The International Journal of Digital Curation
Volume 7, Issue 2 | 2012

The Forensic Curator: Digital Forensics as a Solution to Addressing the Curatorial Challenges Posed by Personal Digital Archives

Gareth Knight,
Project Manager, Research Data Management Support Service
London School of Hygiene and Tropical Medicine

Abstract

The growth of computing technology during the previous three decades has resulted in a large amount of content being created in digital form. As their creators retire or pass away, an increasing number of personal data collections, in the form of digital media and complete computer systems, are being offered to the academic institutional archive. For the digital curator or archivist, the handling and processing of such digital material represents a considerable challenge, requiring development of new processes and procedures. This paper outlines how digital forensic methods, developed by the law enforcement and legal community, may be applied by academic digital archives. It goes on to describe the strategic and practical decisions that should be made to introduce forensic methods within an existing curatorial infrastructure and how different techniques, such as forensic hashing, timeline analysis and data carving, may be used to collect information of a greater breadth and scope than may be gathered through manual activities.

Introduction

The institutional archive is a familiar part of the academic landscape, collecting a broad range of material ranging from corporate business records to the personal collections of notable individuals (e.g. politicians, academics) and making them available for access and use by researchers. Traditionally, these resources have been held in analogue form: paper, cassette tapes, video tapes, and other realia. However, the growth of computing technology during the previous three decades has resulted in a large amount of content being created in digital form. As their creators retire or pass away, an increasing amount of this material is being offered to the academic archive. For the digital curator or archivist, the handling and processing of such personal data collections represents a considerable challenge, requiring development of new processes to address diverse media types, file systems and data structures. This paper discusses work performed by the JISC-funded FIDO project (Forensic Investigation of Digital Objects) at King's College London, describing how digital forensic techniques commonly used in law enforcement and law enforcement may be repurposed to enhance archival processes for acquiring and analysing Personal Digital Archives. It discusses several of the strategic and practical decisions that should be made when applying forensic practices and goes on to highlights software tools that may potentially be adopted by digital archives to partially automate an accession workflow and simplify the decision-making process.

Challenges of Handling Personal Digital Archives

The Personal Digital Archive represents one of several types of digital collection that an institutional archive (or other organisation) may collect and manage. Broadly, a Personal Digital Archive refers to any digital item "within an individual's control that have been stored and maintained by the individual" (Cushing, 2010). It may encapsulate digital information created by one or more individuals for personal and/or work purposes, held on one or more types of media. The type of material found within a Personal Digital Archive and its use by the creator has similarities to personal collections of physical material, potentially containing a mixture of author-created material, correspondence and work created by others. In some cases, the information may be unique, representing the only copy that exists.

The information contained within an Personal Digital Archive may have considerable value to an investigator: private email correspondence and previously unpublished drafts may provide insight into a person's private thoughts and research process, while third party content held on a machine, such as web pages and word processor documents, may provide an understanding of the resources that were consulted and the intellectual context in which work was produced. Preservation of the hardware and software platform in use may provide a future user with an understanding of the environmental conditions in which a creator worked, equivalent to the experience provided through recreation of an individual's working space.[1] The challenge for a digital curator, archivist or researcher is to acquire a Personal Digital Archive in its entirety, analyse the data it contains, and identify information of

[1] For example, the study and library of Sigmund Freud, preserved following his death, has been used to provide visitor's with a better understanding of the environment in which he worked.

 doi:10.2218/ijdc.v7i2.228

relevance to the investigation in a manner that is efficient in the application of methods, accurate in its output and non-invasive in regards to the digital source.

The Personal Digital Archive share similarities with other categories of digital collection handled by a digital archive, utilising storage media and containing information encoded in file formats similar to those used by research data collections and business records. However, differences in the manner that they are used by their owner, the scale of information held on these devices and the method in which they are provided to a digital archive, present challenges that may require the adoption of different approach to their acquisition and analysis.

The first set of challenges to be addressed relate to the process of acquiring the Personal Digital Archive for analysis by a collecting institution. The PDA may be provided by a creator, their family or estate as-is, whom may have little or no knowledge of the information content it contains or the method of obtaining access. Specific issues that must be addressed during the archival deposit process include:

1. **Establishing equipment to be the target for deposit and acquisition:** A Personal Digital Archive may be held on one or more of several types of storage media. Portable media (e.g. 3.5-inch floppy disk, CD-ROM, external hard disk) may be easily accessed using different computing environments. However, other media may be tied to the creation device, e.g. solid-state media installed in a digital camera, phone or tablet, or on hard disk installed within an desktop or portable computer). The challenge will be to determine the electronic equipment to be deposited, whether this be the computer in its entirety[2] or individual items of digital media.
2. **Establishing the method of acquiring the PDA for the archive:** The method adopted to transmit a Personal Digital Archive to the archive must take into account several factors, including the artefact's fragility, size and any depositor-imposed conditions (e.g. the depositor may not wish to provide the storage media or device itself). The challenge will be to establish the most effective method of capturing the PDA in its entirety for subsequent analysis. In these circumstances, it may be impractical to physically transfer the storage media to the archive, requiring the application of alternative methods to acquire the PDA in situ. Possible options may include visiting the workplace or home of the depositor and transferring data using available resources.

A second set of challenges relate to the method of obtaining access to the digital information. Electronic equipment used by a person to create or obtain digital information may be obtained through several routes, including being bought, found, gifted or inherited (Kirk & Sellen, 2008). A variety of electronic equipment is sold to the consumer market, presenting the possibility that a range of electronic devices may be used to create/obtain and store digital information. The time difference between data creation and deposit into a data archive introduces a temporal component, raising the possibility that legacy electronic equipment was used, some of which may be obsolete. Five challenges may be identified as a result of the hardware/software environment that was used to create and store digital information:

[2] The deposit of the original creation environment may resolve the challenge raised in point one and two of physically connecting and accessing digital media.

1. **Obtaining media access:** The digital media selected by the creator to store digital information will be affected by the options available at the time. Contemporary data may be stored on one or more of several devices, including hard disks (e.g. connected via PATA, SATA, SCSI or USB), solid state media (connected via. Secure Digital (SD), CompactFlash (CF), MultiMediaCard (MMC), or USB interface), optical media (Blu-Ray, DVD, CD-ROM), or device-specific storage (e.g. solid state media embedded within a digital camera). Digital information created several years previous may be stored on obsolete media formats. Examples include floppy disk (e.g. 3.5-inch, 5.25-inch, 3-inch) and cartridge formats (e.g. ZIP100, ZIP250, ZIP750, Jaz 1GB), among others.
2. **Interpreting the file system:** A file system acts as a method of storing data on digital media for later retrieval. An operating system will support one or more file systems, one of which may be chosen by a creator when formatting media. Common file systems used by contemporary operating systems include NTFS and FAT32 for MS Windows, HFS and HFS+ for Apple MacOS, and ext2-4 in the Linux OS. Other commonly used formats include ISO 9660 and UDF for optical media, and the Linear Tape File System (LTFS) for digital tape. Legacy OS software may use other file systems (e.g. AmigaOS OFS and FFS, OS/2 HPFS). The challenge will be to identify the file system in use and determine the most effective method of interpreting its structure.
3. **Understanding organisational structure and labelling:** Data may be organized and labelled according to the user's ad hoc needs and/or in accordance with file system requirements with little or no consideration that they would be examined by others at a later date. These may be intrinsic for establishing the purpose that the information performed and the semantic meaning that must be maintained. The challenge will be to understand the creation context and ensure it is transferred into a managed environment.
4. **Identifying digital information of relevance to an investigation:** Large capacity digital media may contain thousands of files obtained from different sources, including operating system files, software application executables and libraries, log records, internet browser cache, temp files, as well as data created by one or more users. The challenge for an investigator will be to locate digital information of value within the digital 'haystack'.
5. **Establishing the provenance of user created data:** User created data held on digital media may be created by different users (the machine owner, users with accounts on the machine, as well as content owned by third parties) and used for different purposes. The challenge for an investigator will be to determine the provenance of the digital information and its implications for curation, preservation and access.

Digital archives, such as the UK Data Archive and Archaeology Data Service, maintain procedures to assist staff to process research data collections and transfer them into a digital archive for curation and preservation. Although these procedures are broadly applicable to personal data collections, they are written with a

 doi:10.2218/ijdc.v7i2.228

presumption that the depositor will have performed basic steps to prepare their data for deposit and the type of digital material that will be received. Unlike research data collections deposited with a digital archive, it is not possible to request that the submitter provide the information in a form that will be simpler to process, and rejection of the PDA may result in potentially unique digital information being lost. The challenge for a digital archive will be to develop new processes to process personal data collections in a manner that is time efficient and provide accurate results.

Overview of Digital Forensics

Digital forensics[3] is a branch of forensic science that emerged from the law enforcement community in the 1980s as a set of methods applied to gather, retrieve, analyse and report upon information held on digital devices, often in relation to a legal investigation (Reith, Carr & Gunsch, 2002). A key feature of digital forensics, which distinguishes it from other activity types, is the emphasis upon "scientifically derived and proven methods" that are acceptable in a legal context (Palmer, 2001). Assessment criteria for determining the validity and accuracy of forensic methods or tools are built upon the Daubert Standard (Ryan & Shpantzer, 2002), a rule of evidence used by US trial judges to assess the relevance and reliability of an expert's testimony. To be accepted in a legal environment, a method or tool must have undergone testing, been subjected to peer review and publication, and be widely accepted by a community of experts. Evidence should be provided that reveals the known error rate for the tool/method (i.e. where it may be used effectively and circumstances where it will produce erroneous or inaccurate results) and indicate standards that govern its application.

The emphasis upon evidence-based evaluation and availability of ready-made tools for performing a digital forensic investigation makes it appealing for those beyond the confines of law enforcement, both as a research and data management tool. Kirschenbaum et al. note the similarity in objectives and methods applied by archivists and forensic investigators, indicating they represent "evidence of something fundamental about the study of the material past, in whatever medium or form" (Kirschenbaum, Ovenden & Redwine, 2010). In a wide-ranging analysis of the application of digital forensics in the cultural heritage domain, they describe how forensic techniques may serve as a component of a risk management strategy, minimizing the risk of media failure or loss through the creation of disk-level backups. The use of non-invasive techniques to acquire and analyse digital material, accompanied by appropriate metadata to establish data integrity and document the investigation process is also recognised as a key requirement that must be performed (Leighton John et al., 2010, Kirschenbaum, Ovenden & Redwine, 2010). This is supported by Duranti (2009), who positions digital forensics as a digital equivalent to the diplomatics functions performed by archivists to establish the provenance of paper records.

[3] Also referred to as computer forensics and digital forensic science.

Applying Digital Forensics Methods Within a Digital Archive

A forensic investigation is comprised of a set of activities performed to gain an understanding of an area of interest. The representation of the investigation process has been a topic for exploration during the past two decades, resulting in the creation of several frameworks to represent the investigation workflow. Although intended for a law enforcement/legal environment, these models describe concepts similar to archival principles, and outline procedures and processes that may be applied equally well to the archival and academic research domains.

Notable work by Pollitt (1995), later adopted by the National Institute of Standards and Technology (Kent, Chevalier, Grance & Dang, 2006), establishes a conceptual model outlining how data held on (analogue/digital) media is analysed and interpreted as information for use in a legal context, and submitted as evidence for use in a court of law. This has broad similarity to the process by which data is transformed into information within the OAIS Reference Model (CCSDS, 2002). In both instances, information is produced using its encoding specification, differing only in the purpose that it serves. OAIS information utilizes representation information to render information in a form understandable by the user, whereas NIST's information uses "knowledge of data file types" (Kent et al., 2006), equivalent to OAIS RepInfo to interpret data and identify information of relevance within the context of the investigation. These commonalities simplify the process of mapping the forensic investigation process onto an OAIS compliant archive.

Several forensic models have been developed during the past 20 years to frame the investigative process. These models build upon the principles of forensic science, information technology and knowledge management, but differ in the composition of these elements, level of prescriptiveness, degree of detail and terminology in use. The work of the first Digital Forensics Research Workshop (DFRWS) has been particularly influential in the field, providing a framework comprised of eight 'activity classes' around which discussion on forensic processes may be framed (Palmer, 2001). Subsequent work by Reith, Carr and Gunsch (2002), Carrier and Spafford (2003), Pollitt (2004), Agarwall and Gupta (2011), and Reith, Carr and Gunsch (2002) build and expand upon the DFRWS model. By synthesizing these models, it is possible to identify six broad phases of an investigation, each of which incorporates a set of one or more common activities.

1. **Prepare:** The set of activities associated with incident recognition, identifying the environmental conditions where a forensic investigation is required, the strategy that should be applied, tools and techniques that must be developed, as well as permissions that must be obtained (e.g. request a search warrant) to undertake the investigation (Agarwal & Gupta, 2011; Carrier & Spafford, 2003; Reith, Carr & Gunsch, 2002).
2. **Acquire:** Data related to a specific event or topic of interest is identified, labelled, recorded and collected. This stage will cover activities necessary to isolate, secure and preserve the state of physical and digital evidence (e.g. preventing people from using the digital device or allowing other electromagnetic devices to be used within an affected radius) and creating a copy of the digital media for later examination.

 doi:10.2218/ijdc.v7i2.228

3. **Examine:** Techniques are applied to perform an in-depth systematic examination of acquired data to identify and locate information of potential relevance to the investigation.
4. **Analyse:** Information contained within the extracted data is manually analysed by an investigator and evaluated for relevance and value in addressing questions raised during the investigation. New questions may be raised as a result of its performance that requires the examination activity to be repeated several times.
5. **Report:** The results of the investigation activity are documented and presented for consideration, on conclusion of the investigation. The report will include details of actions performed, knowledge gained and future steps that must/should be taken.
6. **Review:** The experience of performing the investigation is reviewed to identify improvements that could be made to existing processes (Agarwal & Gupta, 2011; Carrier & Spafford, 2003) and action performed to store the evidence in an appropriate environment for later consultation and/or return to the owner (Reith, Carr & Gunsch, 2002).

The broad investigation model may be applied to the pre-ingest phase of an OAIS-compliant archive, formalizing the activities necessary to locate digital information and transfer it into a managed environment for curation and preservation. These activities may be formalized into a set of policies and procedures for application within and externally to the digital archive.

Practical work in applying digital forensic methods within the archival domain remains at an early stage, although there have been a number of notable developments. The Andrew W. Mellon Foundation-funded AIMS project[4] and JISC-funded FIDO projects (described in this paper) have developed broad procedures and documentation for applying digital forensic practices within an archival environment, while Stanford University Libraries (2011), the Bodleian Library (Thomas, 2011), and the British Library (Leighton John et al., 2010) provide case studies on the process of preserving born-digital and hybrid collections. More recently, the Mellon Foundation funded the BitCurator[5] project to develop forensic tools to enable broader use by collecting institutions.

The remainder of this paper will examine the strategic and practical decisions that must be considered when developing a pre-ingest workflow to locate digital information of relevance to an investigation using digital forensic methods.

Preparation

The initial preparation phase covers a broad set of activities necessary to identify the scenario when an investigation is required, determine the appropriate strategy to adopt, and prepare necessary resources to undertake the investigation.[6] In a law

[4] AIMS project blog: http://born-digital-archives.blogspot.co.uk/

[5] BitCurator: Tools for digital forensics methods and workflows in real-world collecting institutions: http://www.bitcurator.net/aboutbc/

[6] Although unstated within the various investigation models, there is an intrinsic presumption that the institution will possess the existing infrastructure and expertise necessary to perform a forensic

enforcement environment, this will include initial identification of the event that is alleged to have taken place, followed by a set of actions to prepare for the investigation (obtain search warrant, identify incident location and establish tools need to preserve the crime scene). The archive equivalent is likely to be similar to the process currently applied by digital archives to obtain research data produced by funded researchers. Contact is initiated, either by the archive (e.g. making an enquiry regarding the existence of specific work) or by the depositor (e.g. the retiree, their family or estate) and the status of the digital collection is discussed. A depositor/donor agreement is then negotiated establishing the conditions for deposit and an appropriate transfer method established.

Negotiation must take into account the additional complexities introduced by the forensic process. Unlike law enforcement, it is not possible to mandate that all data is provided. Nor is it feasible to establish specific criteria for deposit media and file formats, as defined for funded research data collections. It is therefore important that the investigation process is conducted transparently, with recognition of ethical and confidentiality requirements. Farr (2010) and Redwine (2010) describe the set of challenges encountered when archiving the content of Salman Rushdie's digital collection, arguing that new processes for handling negotiation are required that take into account the archival objective to acquire digital information of research value from a personal digital archives, while respecting the creator's right to privacy. Key issues that an archive may wish to consider during the initial negotiation stage include:

- Conditions of deposit:
  1. The type/extent of analysis authorised by the depositor,
  2. The type of material that they are willing to make available for access and use.
- Approach strategy:
  1. Location of media and physical transfer method,
  2. Software/hardware tools to use to perform the data transfer.

A key issue to be addressed is the analysis type that the depositor authorises may be performed upon the digital media. Forensic techniques, such as data carving and super timeline analysis, enable an investigator to recover data that the creator may not have realised existed or considered removed. The potential that multiple users will have used the drive at different times adds an extra level of complexity, requiring the adoption of a granular approach to the analysis of user data. To meet ethical obligations and demonstrate transparency of operation, depositors should be provided with a description of activities that are to be performed, with the option to opt in or out as necessary. For the FIDO project, depositors were provided with a high-level description of performed activities and a checklist indicating the activities that they do not wish to authorise (e.g. do not recover deleted data, do not use web browser bookmarks). Use of language is considered to be particularly important at this stage, avoiding terms that may imply criminal activity (the phrase 'analysis methods' is used, rather than 'digital forensics') or unnecessary jargon (e.g. 'data carving', 'text

investigation, which will be added to or tailored to the circumstances of the specific incident.

mining'). Depositors are subsequently provided with a list of files that had been selected for retention and relevant examples provided in the reporting phase, to ensure that they authorise that the digital information may be curated and made available. Consultation with several stakeholders may be required, if data held on the drive has been created by two or more users. The licence agreement established as a result of negotiation becomes, in essence, an archival search warrant, providing the digital curator with permission to perform their investigation.

At a practical level, the approach strategy to acquire digital material must also be considered to establish if it is feasible to transport the Personal Digital Archive to the archive. Conditions assigned to the PDA (e.g. the device continues to be in active use, is too fragile, or too expensive to transport) may prevent its transport and, as a result, the digital curator or archivist may be required to visit the depositor's workplace/home to acquire an in situ copy. In these circumstances, the investigator will need to provide relevant tools (e.g. external USB hard disk, boot disc) or work with the restrictions imposed by the host computer system.

Acquisition of Digital Media

Acquisition refers to a process of obtaining data for analysis and examination. The capture of digital media in its existing state – the digital equivalent of preserving the 'crime scene' – is recognized as a challenge in the digital domain. The act of powering on a computer may initiate software tools that read and write data to digital media without user intervention or knowledge. To enable the state of digital media to be acquired in a manner that maintains its integrity, practice within the law enforcement community has focused upon the creation of a bit copy of the digital material – an exact copy of a disk or computer memory – as an image file (Craiger, n.d; Perumal, 2009). A disk image is a set of one or more files that, in combination, contain the content and structure of a mass storage device, including hidden/deleted data that is invisible to the end user. By utilising a disk image, rather than the original media, the investigator is able to apply analysis methods and tools unavailable in the original environment (e.g. data carving) to extract information, and minimise the risk that disk failure or inadvertent, unrecoverable data changes will occur.

A digital archive wishing to apply digital forensic methods must address several strategic and practice issues related to the acquisition phase:

- Strategic decisions:
  1. Type of digital material to be acquired,
  2. Choice of acquisition format,
  3. Retention policy related to the acquired disk image.
- Practical decisions:
  1. Method of obtaining physical access to digital media,
  2. Hardware and software tools to perform acquisition.

A key strategic decision to be made by a digital archive relates to the choice of acquisition formats. Several disk image formats exist, which have had varying degrees

of uptake by the digital forensic and broader IT market. The proprietary EnCase format[7] developed by Guidance Software is considered to be the de facto standard for forensic disk images, due to the popularity of the EnCase software tool within the law enforcement community (Garfinkel et al., 2006). However, in recent years it has been challenged by the Advanced Forensics Format (AFF), an extensible open format for storage of disk images and related forensic metadata. More generally, the raw format is widely used within the IT industry, as a result of being the default type created by the DD command (Garfinkel et al., 2006), provided with all Unix/Linux distributions.[8] In addition, there are a number of application-specific formats maintained by specific software developers[9]. Each disk imaging format is capable of holding a bit copy of digital media. However, differences in additional features provided and level of support, may prompt an institution to adopt one format in preference to another.

To determine the disk image format suitable for the needs of a digital archive, it is necessary to determine a set of evaluation criteria. Relatively little work has been performed on this topic, the most notable being that produced by Garfinkel et al. (2006), which highlights the importance of extensibility, licence status, compression support and data location as factors that require consideration. To determine the disk image format to adopt for the FIDO project, Knight (2011) drew upon work by Todd (2009) for file format selection criteria, as well as the aforementioned work by Garfinkel et al., to propose eight factors that may be taken into account when selecting a disk image format:

1. **Adoption:** the extent to which the format is in widespread use within the forensic community and elsewhere;
2. **Software independence:** the extent to which the format is independent of specific support from hardware and software;
3. **Disclosure:** the extent to which the file format specification is in the public domain;
4. **Metadata support:** the extent to which descriptive information is supported in extractable form within the format;
5. **Licence status:** the licence associated with the format, which may affect the degree of disclosure and adoption;
6. **Level of fixity analysis supported:** the level at which fixity information may be recorded within the disk image;[10]
7. **Support for split files:** the ability to split a large disk image into smaller sections of an arbitrary size for storage on disc or other media;

[7] EWF is also supported by a number of open source tools, via the LibEWF library.

[8] Raw images contain a bit-by-bit copy of a source device, without any attempt made to identify or interpret the filesystem or files held on the disk. As a result, it is a misnomer to describe raw as a disk imaging format.

[9] Examples include the ILook Investigator IDIF, IRBF, and IEIF Formats, ProDiscover image file format, PyFlag's sgzip Format, Rapid Action Imaging Device (RAID) format and Safeback format.

[10] Forensic literature refers to fixity checks being performed at three levels: a fixity check of the data image as a whole, a check on individual files within the data image, and a check on each segment or chunk of data within the image.

 doi:10.2218/ijdc.v7i2.228

8. **Compression support:** the ability to compress the data image to reduce storage space. Compression support is useful but it is not considered mandatory that the format provide built-in support, since it will take longer to locate data on a compressed file.

On the basis of the selection criteria, the FIDO project designated AFF as the preferred format in which to acquire disk images, due to the degree of openness and extensibility.

A second strategic decision for a curatorial institution to make relates to the role of a disk image within an OAIS compliant system and the retention policy that should be applied. Should the disk image be considered a Submission Information Package (SIP), or as packaging that holds digital information of value? If the former, the archive will wish to retain the disk image for long-term storage and preservation. However, if the subset of data extracted from the image is considered to be the SIP, the digital archive may justify the deletion of the disk image. Arguments may be made for each approach: new analysis techniques may be applied to a disk image if it retained, but the storage of a number of disk images, each of which may be several gigabytes or terabytes in size, will require a considerable amount of disk space. Alternatively, a middle-ground approach might be adopted: store disk images for a subset of digital collections (e.g. those that may be assigned specific attributes, such as belonging to a notable individual, or are acquired from low capacity media, that has small storage requirements), while declaring that selected data represents the SIP for other digital collections.

At a more practical level, the issue of how digital media may be connected and accessed at the physical layer must be considered. The level of challenge posed will vary, dependent upon the media type in use for data storage and depositor conditions (e.g. internally mounted disks cannot be removed). For contemporary or widely adopted media formats (e.g. 3.5-inch floppy disks, CD-ROM, DVD-ROM, PATA or SATA hard disks), it may be a simple task to connect and access digital media using current computer platforms. However, the potential gap between date of creation or use and deposit, as well as the potential that the creator uses a less common media type, introduces the possibility that legacy or non-standard media will be provided that cannot be accessed using readily available hardware. A researcher working in the 1980s and 1990s may, for example, have stored their research on 3-inch or 5.25-inch floppy disk, a Iomega ZIP 100/250/750 cartridge, Iomega 1GB Jazz disk, or an obsolete hard disk or digital tape format. Options for gaining physical access to digital media include:

- Connect the digital media to contemporary hardware;
- Use the legacy system owned by the data creator or a model similar to that used (e.g. an Amstrad PCW equipped with a 3-inch disk drive) capable of accessing the legacy media;
- Use hardware compatible with the legacy system (e.g. a 80486 desktop machine fitted with a 5.25-inch disk reader and running appropriate software to access a 5.25-inch CP/M disk;

- Obtain contemporary third-party hardware that enables legacy media and file system formats to be accessed using contemporary hardware.[11]

The chosen approach will influence the processes applied to acquire the digital media and store it as a disk image. Forensic practice for options 1, 2, and 3 recommend that the machine is booted from third party digital media to minimise the risk of accidental data change, and use disk-imaging software to capture an image of the source media (e.g. written to an external USB disk or transferred over ethernet, parallel, or serial). If this fails, it may be advised that the physical disk is removed from the host machine and imaged using a forensic computer (Craiger, n.d.).

A large number of software tools exist capable of capturing an image of diverse types of digital media. These include dc3dd, dcfldd, Guymager, Automatic Image and Restore, FTK Imager, OSFClone, and others. Each tool offers different functionality in terms of the imaging formats (RAW/DD, AFF, EWF), capture speed, metadata supported and so on. To assist evaluators, the National Institute of Standards and Technology has defined a set of criteria for evaluating acquisition tools, identifying eight mandatory and 13 optional requirements that they must or should fulfil (NIST, 2004), and performed evaluation upon several tools in this area (NIST, n.d.). The mandatory requirements focus upon core functionality, indicating, for example, that software tools should be capable of acquiring all visible and hidden data sectors accurately (DI-RM-04 & 05), and should notify the user of error type and location (DI-RM-07). The FIDO project elected to use OSFClone – a self-booting Linux live disc – on the basis that it was simple to use by archival staff (it did not require use of command line parameters), could be configured without the need for a mouse, and supports several common disk formats.

Curatorial institutions seek to perform acquisition and analysis activities in a manner that avoids or minimises the likelihood that an artefact will be damaged. It is common, for example, to wear gloves when handling physical objects to avoid the risk of inadvertent contamination. Similar activities are performed by law enforcement in the digital realm to mitigate the risk that the acquisition process will itself result in data being removed or overwritten from the source media: an event that may result in questions being raised regarding the validity of data collected in a legal context. This is achieved through use of a write blocker, a hardware plug-through unit that connects between the computer system and the media reader that acts to prevent write operations initiated by the operating system being performed on the disk (Craiger, n.d.).

The acquisition of a disk image offers the potential to capture a large amount of digital information in the original digital environment and apply new techniques to analyse digital data. To ensure that forensic practices may be applied in a sustainable manner, a curatorial institution must make a number of strategic decisions to ensure it is sufficiently equipped to forensically acquire digital media and manage disk images in a form that meets curatorial requirements. The decision-making process must be combined with more practical considerations to ensure that the capture method is fit for purpose and may be performed by staff in practice.

[11] Notable development in this area include the Catweasel floppy disk control by Individual Computers and Kryoflux by the Software Preservation Society, enabling a range of 3.5” and 5.25” disk media to be accessed and read.

 doi:10.2218/ijdc.v7i2.228

Analysing digital media

The information contained within an Personal Digital Archive may have considerable value to an investigator: private email correspondence and previously unpublished drafts may provide insight into a person's private thoughts and research process, while third party content held on a machine, such as web pages and word processor documents, may provide an understanding of the resources that were consulted and the intellectual context in which work was produced. The challenge for a digital curator, archivist, or researcher is to locate information of relevance to their investigation in a manner that is efficient in the application of methods, accurate in its output, and non-invasive in regards to the digital source.

Several forensic techniques may be applied to analyse a disk image and output relevant information, dependent upon the research question that the investigator wishes to address. Specific questions that may be queried include:

1. Does the disk contain digital information created by the owner or obtained from another designated source?
2. Does the disk contain material that provides insight into the development process of the owner, including previous drafts and discarded work?
3. What insight does the disk provide in determining how it was used by the owner?

A disk image may contain several thousand data files, ranging from operating system and software applications, through to internet browser cache, log files and user created data, some of which may be relevant to the investigation. Through the application of automated forensic methods, an investigator may analyse the diverse types of data contained on a disk and identify a subset that is relevant to their investigation.

Identifying relevant material by its origin

A first objective that an investigator may wish to perform is to locate digital information relevant to the investigation. For a digital curator, this may be motivated by a desire to locate user created data that should be curated and preserved, whereas a researcher may be more interested in locating information related to a specific topic.

A common technique used to locate relevant information is to search for files based upon their filename (e.g. *report*, *paper*), file extension (.pdf, .doc, .jpg), creation/access/modification date or containing specific content. Although effective in many circumstances, starting a search using these methods is likely to produce superfluous results (e.g. JPEGs associated with a software product), or perhaps more of a concern, omit material relevant to the investigation (files that possess an unexpected filename, format or creation/access/modification date). To improve the efficiency of discovery activities, and reduce the likelihood that user created data will be overlooked, forensic techniques such as forensic hashing may be adopted to differentiate data files obtained from different sources and identify that which should be the target of investigation.

Forensic hashing, also referred to as exclusion hashing (Perlustro, n.d.), builds upon techniques commonly used by digital archives to monitor bit preservation

activities – a hash sum (e.g. MD5, SHA-1) is generated for one or more files held on disk. However, in the forensic usage, the value of each file is compared to a dataset of previously recorded hash sums that have originated from a third-party source, (e.g. Microsoft Windows, Adobe Photoshop), as opposed to the previously generated hash sum, and classified according to their origin (see Figure 1). This classification may be used to identify a subset of data files that require further investigation.

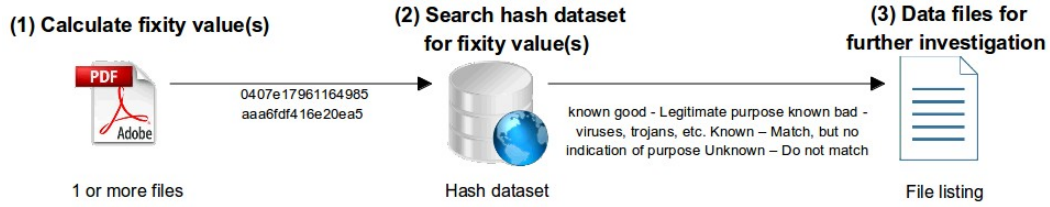

Figure 1. Forensic hashing process.

Several institutions maintain hash sum datasets and provide software tools to assist forensic investigators in determining the provenance of data files. Although intended to serve the investigative needs of law enforcement, they may also be utilised by digital archives and academic researchers to analyse digital media. These include:

- The National Institute of Standards and Technology (NIST)'s National Software Reference Library[12]
- The National Drug Intelligence Center's HashKeeper[13]
- The Online File Signature Database[14]

Each hash sum dataset differs in the number and type of data files catalogued, the extent of information provided and the classification applied. Data files on a hard disk or other digital media may be classified into one of four categories:

1. **Known:** The hash sum for a data file matches one held in the dataset. However, no information is provided on the intended purpose of the data file;
2. **Known Good:** The file(s) originated from a recognized source (e.g. Adobe, Apple, Microsoft, or other known developer) and performs a legitimate purpose on a users' system;
3. **Known Bad:** The file is recognized as belonging to a virus or malware installation;
4. **Unknown:** Hash sums that are not recorded in the dataset.

Forensic hashing is supported by a number of commercial and open source case management tools, including Forensic ToolKit (FTK), OSForensics, Autopsy[15] and

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

[13] Hashkeeper: http://www.justice.gov/ndic/domex/hashkeeper.htm

[14] Online File Signature Database: http://www.filesig.co.uk/ofsdb.html

[15] Autopsy 2 supports hash filtering though integration of a third party plugin. Autopsy has been rewritten in Java for version 3 and, at the time of writing, does not support hash filtering.

 doi:10.2218/ijdc.v7i2.228

PTK, and is utilized by broad forensic toolsets such as The SleuthKit (TSK). However, differences emerge in the level of analysis supported by these tools: open source tools, such as TSK are limited to drive-level analysis, examining each file contained within a disk image and classifying it appropriately. By contrast, commercial case management tools, such as FTK and OSForensics place emphasis upon item-level hash lookup, encouraging the user to establish the origin of one or more files selected through a graphical interface[16]. When implementing forensic hashing with the examination workflow of a digital archive, consideration should be given to the level of analysis that will be performed. Performance of an automated media-level analysis will provide a complete listing of data files held on disk, but may require several days to produce. By contrast, item-level hash lookup will provide more immediate results. However, files must be selected manually, introducing the possibility that relevant files will be overlooked.

The FIDO project developed an automated workflow for performing forensic hashing using The Sleuthkit[17] (TSK) – a compendium of open source forensic tools and scripts developed by Brian Carrier. TSK contains a perl script called 'Sorter' which simplifies the process of characterizing data files (through use of the Unix File command) and classifying data files on digital media as known or unknown using the NSRL dataset. Forensic hashing is a processor-intensive task that can require some time to perform.[18] As an example, it took six days and 12 hours to process 22,672 files held on a real-world 60GB hard disk. However, once finished, four HTML reports were produced, indicating the files that matched entries in the NSRL dataset (exclude.html); files that contain file extension mismatches (mismatch.html), files that contain file extension mismatches and are found in the NSRL dataset (mismatch_exclude.html); and files that are not found in the NSRL dataset (unknown.html).

Program Files/btbb_wcm/html/images/icons/unsecure.gif
GIF image data, version 89a, 36 x 29
Image: /mnt/shared/lo-partition21.img Inode: 94185-128-3

Program Files/btbb_wcm/html/images/icons/unsecure_small.gif
GIF image data, version 89a, 36 x 36
Image: /mnt/shared/lo-partition21.img Inode: 94186-128-4

RECYCLER/S-1-5-21-3943365952-1317941163-395094903-1006/Dc37.asd
CDF V2 Document, Little Endian, Os: Windows, Version 5.1, Code page: 1252, Title: , Author: Lindsay, Template: Normal, Last Saved By: Lindsay, Revision Number: 2, Name of Creating Application: Microsoft Word 10.0, Total Editing Time: 02:00, Create Time/Date: Sat Jan 16 23:50:00 2010, Last Saved Time/Date: Sat Jan 16 23:50:00 2010, Number of Pages: 1, Number of Words: 50, Number of Characters: 285, Security: 0
Image: /mnt/shared/lo-partition21.img Inode: 66445-128-3

RECYCLER/S-1-5-21-3943365952-1317941163-395094903-1006/Dc38.doc
CDF V2 Document, Little Endian, Os: Windows, Version 5.1, Code page: 1252, Title: , Author: Lindsay, Template: Normal, Last Saved By: Lindsay, Revision Number: 3, Name of Creating Application: Microsoft Word 10.0, Total Editing Time: 25:00, Create Time/Date: Sat Jan 16 23:50:00 2010, Last Saved Time/Date: Sun Jan 24 23:11:00 2010, Number of Pages: 1, Number of Words: 88, Number of Characters: 508, Security: 0
Image: /mnt/shared/lo-partition21.img Inode: 1677-128-3

Figure 2. Unknown category.

The adoption of forensic hashing techniques offers the potential to improve the efficiency of discovery activities, providing the investigator with a reduced list of files on which they may focus their investigation. The categorization list may serve to

[16] A hash lookup may be performed upon an entire disk if required through selection of multiple files.

[17] The SleuthKit: http://www.sleuthkit.org

[18] Data files held in a disk image must be extracted for characterisation and classification, which takes additional time and uses disk space on the host machine.

address different questions: a digital curator wishing to identify data unique to the target machine will be interested in the unknown files list; a researcher examining the working environment and data creation practices of a user may be interested in the tools in the known files list; whereas an investigator analysing the impact of malicious software upon a live system would be interested in the known bad list. However, additional filtering is likely to be necessary to exclude data files that are outside the scope of analysis from the list of unknown files, e.g. Windows thumbs.db, log files, cache files, corrupted or patched data files, and uncatalogued software files. To address these issues, it may be helpful to use additional techniques, such as hash de-duplication,[19] fuzzy hashing (as used by SSDeep), or content comparison (as used by XCL), to filter the list using enhanced de-duplication.

Creating a narrative of user activity and identifying material created during a designated time period

A second discovery method available to an investigator wishing to locate specific digital materials is to use temporal information held on digital media. Digital timestamps are commonly used in law enforcement to reconstruct a sequence of events associated with a specific incident (Eiland, 2006). The digital preservation community makes similar use of timestamps embedded within data files – characterisation tools, such as JHOVE are able to extract Creation and Last Modified Date attributes, providing information that may be used in combination with other metadata to establish provenance of content.

Despite being in common use for establishing provenance, it is widely recognized that file timestamps are a potentially untrustworthy information source. Datetime attributes may be applied erroneously by the host system, as a result of mis-configured software, deliberate clock tampering by the user (Rothenberg, 1999), or hardware clock drift (Shatz, Mohay & Clark, 2006). To limit the impact that erroneous file timestamps have upon an investigation, development and discussion in the forensic community has focused upon the creation of super timelines, which take a holistic approach to the extraction of temporal information on a host system (Guðjónsson, 2010). Although these do not resolve problems caused by a misconfigured internal clock or misapplied file timestamps, it provides a large dataset of temporal information, assigned by the host system or third party systems (e.g. mail servers) that may be used by an investigator to establish whether the clock settings have been suddenly altered, or have gradually become out of sync.

The use of super timelines presents the opportunity for new areas of research. A researcher may, for example use a super timeline to create a narrative of how the owner used the machine to perform their work, providing information on the websites that were consulted and the email correspondence that took place during the investigation process. Alternatively, they may wish to establish the date when a keyword associated with a topic was first used and the frequency of its appearance over time (see Figure 3).

[19] If a hard disk contains ten files with the same hash value, only one will be presented to the investigator for consideration and/or duplicate files will be removed.

 doi:10.2218/ijdc.v7i2.228

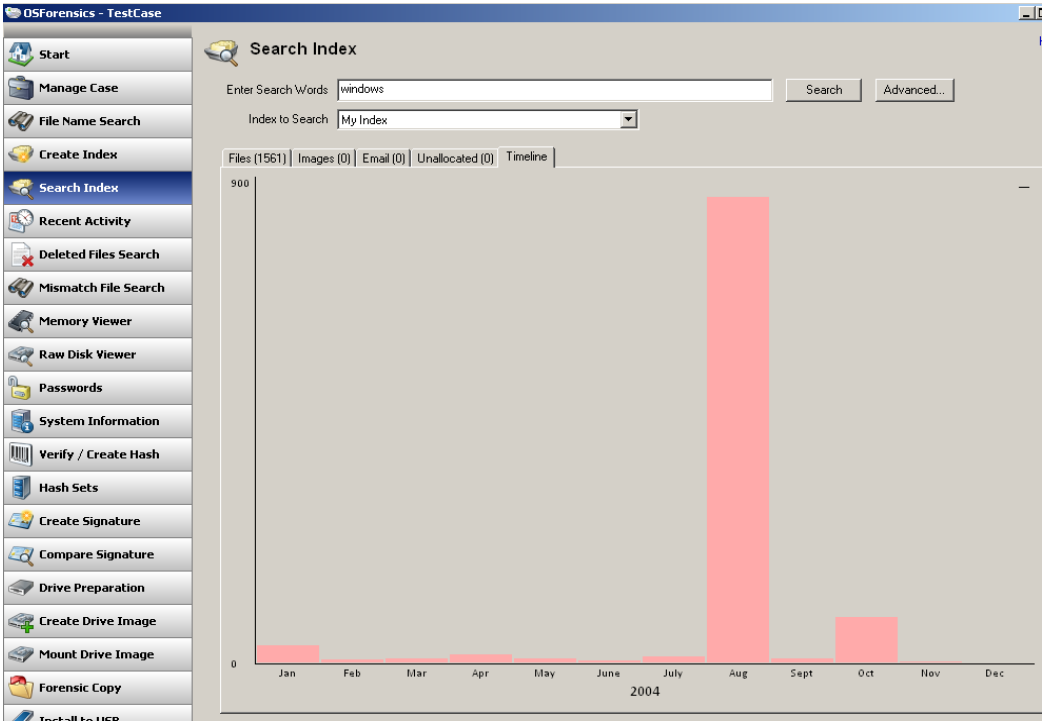

Figure 3. OSForensics keyword search.

The ability to identify, extract and process temporal information held on a host system varies between software tools. The open source, perl-based, TimeScanner and Log2Timeline are, arguably, the most effective tools for generating super timelines, capable of parsing temporal information held in many different locations and formats,[20] normalizing the information,[21] and outputting it to an open, structured format (TSK MacTime, SIMILE XML, BeeDocs, CSV, tab-delimited) for analysis. However, set-up and configuration of these tools require knowledge of perl and familiarity with the command line, which may be onerous for archival staff. By contrast, commercial forensic applications, such as FTK and OSForensics, provide a graphical interface for search and visualization that make them simpler to use. However, at present, these applications support fewer information sources and provide output that is less detailed in comparison to TimeScanner.

Identifying 'lost' material by location

A third discovery method that an investigator may wish to apply is to analyse a disk for 'lost' information that the user has chosen not to retain. For a digital curator or archivist, data recovery may be driven by a desire to locate abandoned or previous versions of works that the creator discarded, or retrieve contextual information that provide an insight into the user's information creation processes. In the physical realm, an archivist might attempt to locate evidence of works by examining a collection for evidence of 'lost works' – paper fragments or imprints upon a piece of

[20] Examples include Internet browser history and cache files, email mail folders, log files, databases, as well as embedded metadata and file attributes.

[21] Each temporal event possesses 17 elements (some of which may be unpopulated). Date and time information extracted from different sources is normalised to a standard format. However, information in other metadata elements remains as-is, resulting in many different terms being used to describe the same event type (e.g. createDate, MediaCreateDate).

paper that show evidence of previously completed work. In the digital realm, an investigator is traditionally reliant upon the existence of .tmp files produced by a software application's autosave function, or the retention of deleted text within a Microsoft Word file to provide an insight into earlier, overwritten work.

Digital storage is classified into two categories: unallocated and allocated space. Unallocated (sometimes referred to as inactive) space refers to disk sectors that are available for use by the operating system. A formatted disk will be comprised of unallocated disk sectors, some of which will be re-designated as 'allocated' (also referred to as active) sectors when data is written to it. The allocated designation indicates that the sectors contain data and should not be overwritten when transferring new data. When a file is deleted, the sectors are reclassified as 'unallocated', allowing them to be reused. However, crucially, the data held in these sectors continue to exist until a point when they are overwritten.

Several forensic techniques exist that may be applied to recover deleted or difficult to access information, with varying success. These include:

1. **File system undelete:** A file system pointer that references the file is used to identify the location of the complete file and reclassify it as 'allocated'.
2. **File Carving:** A raw data object – a disk image, disk, or other file – is analysed and patterns sought in its structure that indicate the presence of specific content types (e.g. a JPEG image). The data is "carved" for further examination.
3. **Text analysis & extraction:** A file is analysed for alphanumeric text contained within and extracted for review by an investigator.

As a data recovery method, file carving represents the most versatile approach, enabling an investigator to analyse unallocated space in a disk/disk image, identify relevant content encoded in different file types, and extract it for examination. It may be used to recover information fragments within a partially deleted file, or recover data from a file system that has been corrupted by mechanical failure or virus attack – the digital equivalent of locating in a scrap of paper in a physical archive.

By contrast, file system undelete is effective only for a short time when the pointer continues to exist on a file system and cannot be used to recover data that has been overwritten in part.[22] Text extraction may be used effectively to output textual information held on disk for analysis (e.g. to establish the type of information held by the owner, or determine if the owner possesses information on a specific topic).[23] However, as the name suggests, it is limited to text content only.

File carving tools use several carving techniques, alone or in combination, to identify embedded data objects within a raw datastream, each of which has benefits and limitations associated with their use. Header-footer (H/F)[24] carving is one of the

[22] See ForensicWiki definition of file carving at: http://www.forensicswiki.org/wiki/File_Carving

[23] Forensic tools, such as Bulk Extractor, may be used to extract specific information, such as email addresses and web sites visited.

[24] Variants of this method include Header/Embedded Length and Header/Maximum Carving. H/E carving uses file size information contained within the header of specific formats (BMP, PDF, AVI) to specify the amount of data to be carved. H/M identifies the header and carves data sequentially until a maximum file size (e.g. 10MB) is reached.

 doi:10.2218/ijdc.v7i2.228

simplest carving methods, and is supported by a large number of forensic tools (Beek, 2011). A datastream is searched for byte sequences that match the header of a known file type (e.g. 'nx47nx49nx46nx38nx37nx61' that denotes a GIF header), followed by the byte sequence for a corresponding footer ('nx00nx3b'). The H/F carving method is developed based upon the assumption that data files are stored contiguously on disk, and is applied to file formats that have a uncommon header and footer byte sequence. These assumptions make it effective when attempting to recover data files that possess a large header and footer, and are held on media that has little or no fragmentation. However, it may produce false positives when searching for file types that have a small or no header/footer (e.g. zip archives, text files) or are stored in non-contiguous locations across a disk (Kloet, 2007). This may result in the creation of invalid files that contain byte sequences that were stored contiguously on disk, or "franken files", in which the header for Object A is matched with the footer from Object B.

To illustrate the capabilities and limitations of open source/free carving tools, a controlled experiment was performed in idealised conditions. A newly purchased 500GB hard disk was formatted to NTFS and 20 files - five 100k text files, five 5Mb JPEGs, five 90MB WMV videos and five 300MB AVI videos[25] - were copied to the disk, and subsequently deleted. Following the deletion, an image was created using dc3dd and a header/footer carving method was applied to the disk image using four file carving tools.[26]

	ASCII text	JPEG	WMV	AVI
PhotoRec	5	5	0	5 (3 complete, 2 incomplete)
Scalpel	0	5	0	3[27]
MagicRescue[28]	0	5	N/A	N/A
Foremost	0	0	0	0
PhotoRec	5	5	0	5 (3 complete, 2 incomplete)

Table 1. Results of a Header/footer carving performed using four file carving tools.

The controlled experiment was performed in ideal conditions: the use of a newly formatted disk reduced the need for the operating system to use non-sequential data storage, and the recentness of the data deletion reduced the likelihood that some of the data had been overwritten. However, it was evident that none of the tools were able to extract every file. PhotoRec was found to be the most effective for the file types

[25] Files are approximated. Text files were a few bytes smaller or large, while other file types were within 500k of the stated file size.

[26] The file carving tools were selected on the basis that they are available under an open source licence or are available for free.

[27] Files were incomplete, but containing several minutes of video that could be played in Media Player Classic.

[28] MagicRescue – Only recovers files it has a 'recipe' for (JPG, AVI, but not txt or WMV) – recovered JPGs, but not AVI. Did not attempt other formats.

selected, extracting 13 complete files and two incomplete files, followed by Scalpel which extracted five complete files and three incomplete files (the incomplete files were a mix of the complete and incomplete files extracted by PhotoRec). Alternative methods have been recognized which may provide greater accuracy in comparison to Header/XXX carving. For example, file structure based approaches use file format documentation to analyse the internal structure of a known file type (in addition to the header and footer) and/or smart carving methods use knowledge of operating system data handling practices to resolve the problem of locating segments of a data file stored in non-sequential order (Pal, Sencar & Memon, 2008). However, tools that support these methods are still under development.

Forensic techniques, such as file carving and text extraction offer new methods that may be applied by digital curators to locate and recover 'lost' digital information, providing insight into the activities of a researcher. However, they should not be considered a perfect data recovery solution. In addition to the need to enhance carving algorithms to improve accuracy (an area which receives considerable attention within the digital forensics domain), there is a need to improve the usability of open source data carving tools to enable them to be applied by less technical users. Development work is required to produce graphical interfaces that may be used to configure and execute disparate forensic tools, and application of appropriate visualisation techniques to render the output in a form that is understandable.

Conclusion

This paper has identified the Personal Digital Archive as a distinct form of data collection that digital archives have begun to collect, highlighting the challenges that must be addressed by archives more familiar with analogue material and research data. It proposes that digital forensic methods and techniques should be reviewed by digital archives and incorporated within pre-ingest and other stages of an OAIS-compliant system to enable these collections to be captured in their entirety and enables digital information of relevance to an investigation to be located in a manner that is efficient, accurate and non-invasive.

The digital forensic model and methods provide a foundation upon which a digital archive may perform pre-ingest activities. The creation of a disk image will provide the investigator with a bit copy of the original environment that was used to create and/or store the digital information, minimising the risk that data of potential value will be lost or damaged. Further examination, through the performance of forensic hashing, data carving and the creation of super timelines will provide information of a greater breadth and scope than could be provided through current manual activities. However, it is evident there is a need for additional tailoring to integrate them within the existing curatorial infrastructure. Strategic decisions must be made on how digital forensic methods are applied, modifying the policy and procedural framework to take into account the additional functionality provided by forensic software. Specific factors, such as the choice of disk image format, documentary metadata format, and its role within the OAIS must also be considered in conjunction with the broader objectives and capabilities of the archival service.

Although the forensic tools examined in this paper provide necessary functionality, further work is required to simplify the process of installing, configuring and applying

them to digital collections. Many of the open source tools available are powerful, but require extensive knowledge of the command line and scripting languages to use. The output provided by these tools can also be difficult to interpret, unless the investigator possesses some understanding of operating system design. Development work performed by the BitCurator project, as well as greater engagement with forensic tools within software development funding projects and at digital preservation 'hack days' may prove helpful in this area. There is also a need to produce case studies that examine how advanced forensic methods, such as those described in this paper can be applied to extract granular information of archival value and address research questions within distinct academic domains.

References

Agarwal, A. & Gupta, M. (2011). Systematic digital forensic investigation model. Proceedings of INDIACom-2011: 5th National Conference on Computing for Nation Development, Delhi, India.

Beek, C. (2011). *Introduction to file carving.* McAfee White Paper. Retrieved from http://www.mcafee.com/us/resources/white-papers/foundstone/wp-intro-to-file-carving.pdf

Carrier, B. & Spafford, E.H. (2003). Getting physical with the digital investigation process. *International Journal of Digital Evidence, 2*(2). Retrieved from https://www.cerias.purdue.edu/assets/pdf/bibtex_archive/2003-29.pdf

Consultative Committee for Space Data Systems. (2002). Recommendation for space data system standards: Reference model for an Open Archival Information System (OAIS). CCSDS 650.0-B-1 Blue Book. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

Craiger, J.P. (n.d.). *Computer forensics procedures and methods.* Retrieved from http://ncfs.org/craiger.forensics.methods.procedures.final.pdf

Cushing, A.L. (2010). Highlighting the archives perspective in the personal digital archiving discussion. *Library Hi Tech*, *28*(2). Retrieved from http://dx.doi.org/10.1108/07378831011047695

Duranti, L. (2009). From digital diplomatics to digital records forensics. *Archivaria 68: Special Section on Queer Archives*. Retrieved from http://journals.sfu.ca/archivar/index.php/archivaria/article/view/13229/

Eiland, E.E. (2006). *Time line analysis in digital forensics.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.3864

Farr, E. (2010). Finding aids and file directories: Researching a 21st century archive. *Digital Humanities 2010 Conference Abstracts*. Retrieved from http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/book-final.pdf

Garfinkel, S., Malan, D., Dubec, K., Stevens, C., & Pham, C. (2006). *Advanced Forensic Format: An open, extensible format for disk imaging*. Retrieved from http://simson.net/ref/2006/ifip119_aff.pdf

Guðjónsson, K. (2010). *Mastering the super timeline with log2timeline.* Retrieved from http://computer-forensics.sans.org/community/papers/gcfa/mastering-super-timeline-log2timeline_5028

Kent, K., Chevalier, S., Grance, T. & Dang, H. (2006). *Guide to integrating forensic techniques into incident response: Recommendations of the National Institute of Standards and Technology*. National Institute of Standards and Technology Special Publication 800-86. Retrieved from: http://csrc.nist.gov/publications/nistpubs/800-86/SP800-86.pdf

Kirk, D. & Sellen, A. (2008). *On human remains: Excavating the home archive.* Microsoft Technical Report. Retrieved from: http://research.microsoft.com/apps/pubs/default.aspx?id=70595

Kirschenbaum, M.G., Ovenden, R. & Redwine, G. (2010). *Digital forensics and born-digital content in cultural heritage collections.* Council on Library and Information Resources. Retrieved from http://mith.umd.edu/wp-content/uploads/whitepaper_borndigital.pdf

Kloet, S.J.J. (2007). Master's thesis: Measuring and improving the quality of file carving methods. Retrieved from http://alexandria.tue.nl/extra2/afstversl/wsk-i/kloet2007.pdf

Knight, G. (2011). *Forensic disk imaging report*. Retrieved from http://researchonline.lshtm.ac.uk/354890/

Leighton John, J. et al. (2010). *Digital lives: Personal digital archives for the 21st century - An initial synthesis.* Retrieved from http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf

National Institute of Standards and Technology. (2004). *Digital data acquisition tool specification.* Draft 1 for Public Review of Version 4.0. Retrieved from http://www.cftt.nist.gov/Pub-Draft-1-DDA-Require.pdf

National Institute of Standards and Technology. (n.d.). *Computer forensic tool testing: Disk imaging*. Retrieved from http://www.cftt.nist.gov/disk_imaging.htm

Pal, A., Sencar, H.T. & Memon, N. (2008). Detecting file fragmentation point using sequential hypothesis testing. *Digital Investigations, S2-S13*. Retrieved from http://digital-assembly.com/technology/research/pubs/dfrws2008.pdf

Palmer, G. (2001). *DFRWS technical report: A road map for digital forensic research*. Report from the first Digital Forensic Research Workshop (DFRWS). Retrieved from http://www.dfrws.org/2001/dfrws-rm-final.pdf

Perlustro. (n.d.). ILook-IX. Retrieved from http://www.perlustro.com/solutions/e-forensics/ilook-ix

Perumal, S. (2009). Digital forensic model based on Malaysian investigation process. *IJCSNS International Journal of Computer Science and Network Security*, *9*(8). Retrieved from http://paper.ijcsns.org/07_book/200908/20090805.pdf

Pollitt, M. (1995). *Computer forensics: An approach to evidence in cyberspace*. Proceedings of the National Information Systems Security Conference, Baltimore, MD. Retrieved from http://www.digitalevidencepro.com/Resources/Approach.pdf

Pollitt, M. (2004). Six blind men from Indostan. In proceedings of the D*i*gital Forensic Research Workshop Baltimore, MD. Retrieved from www.dfrws.org/2004/day1/D1-Pollitt-Keynote.ppt

Redwine, G. (2010). Archives and 'the archive': The computer as archival object. *Digital Humanities 2010 Conference Abstracts*. Retrieved from http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/pdf/book-final.pdf

Reith, M., Carr C. & Gunsch, G. (2002). An examination of digital forensic models. *International Journal of Digital Evidence, 1*(3). Retrieved from http://www.utica.edu/academic/institutes/ecii/ijde/articles.cfm?action=article&id=A04A40DC-A6F6-F2C1-98F94F16AF57232D

Ryan, D.J. & Shpantzer, G. (2002). *Legal aspects of digital forensics.* Retrieved from http://euro.ecom.cmu.edu/program/law/08-732/Evidence/RyanShpantzer.pdf

Rothenberg, J. (1999). *Preserving authentic digital information: Council on library and information resources*. Retrieved from http://www.clir.org/pubs/reports/pub92/rothenberg.html

Shatz, B., Mohay, G., & Clark, A. (2006). A correlation method for establishing provenance of timestamps in digital evidence. *Digital Investigation 3S* (S98 – S107). Retrieved from http://www.dfrws.org/2006/proceedings/13-%20Schatz.pdf

Stanford University Libraries. (2011). *Processing born-digital materials in the STOP AIDS project records: Introduction and preparation for imaging*. Retrieved from http://lib.stanford.edu/digital-forensics-stanford-university-libraries/processing-born-digital-materials-stop-aids-project-

Thomas, S. (2007). *PARADIGM: A practical approach to the preservation of personal digital archives.* Final Report. Retrieved from http://www.paradigm.ac.uk/projectdocs/jiscreports/ParadigmFinalReportv1.pdf

Thomas, S. (2011). Receiving and managing email archives at the Bodleian Libraries. *In proceedings of the Digital Preservation Coalition Conference: Preserving Email: Directions and Perspectives*. Wellcome Collection Conference Centre, London. Retrieved from http://www.dpconline.org/events/previous-events/835-preserving-email-directions-and-perspectives-2011

Todd, M. (2009). *Technology watch report: File formats for preservation*. DPC Technology Watch Series Report. Retrieved from http://www.dpconline.org/component/docman/doc_download/375-file-formats-for-preservation