# Digital Forensics and Preservation

Jeremy Leighton John

DPC Technology Watch Report 12-03 November 2012

**Series editors on behalf of the DPC**
**Charles Beagrie Ltd.**

Charles Beagrie

**Principal Investigator for the Series**
**Neil Beagrie**

DigitalPreservationCoalition

DPC Technology Watch Series

Published in association with Charles Beagrie Ltd.

**ISSN: 2048-7916**
**DOI: http://dx.doi.org/10.7207/twr12-03**

First published in Great Britain in 2012 by the Digital Preservation Coalition.

## Foreword

The Digital Preservation Coalition (DPC) is an advocate and catalyst for digital preservation, ensuring our members can deliver resilient long-term access to digital content and services. It is a not-for-profit membership organization whose primary objective is to raise awareness of the importance of the preservation of digital material and the attendant strategic, cultural and technological issues. It supports its members through knowledge exchange, capacity building, assurance, advocacy and partnership. The DPC's vision is to make our digital memory accessible tomorrow.

The *DPC Technology Watch Reports* identify, delineate, monitor and address topics that have a major bearing on ensuring our collected digital memory will be available tomorrow. They provide an advanced introduction in order to support those charged with ensuring a robust digital memory, and they are of general interest to a wide and international audience with interests in computing, information management, collections management and technology. The reports are commissioned after consultation among DPC members about shared priorities and challenges; they are commissioned from experts; and they are thoroughly scrutinized by peers before being released. The authors are asked to provide reports that are informed, current, concise and balanced; that lower the barriers to participation in digital preservation; and that they are of wide utility. The reports are a distinctive and lasting contribution to the dissemination of good practice in digital preservation.

This report was written by Jeremy Leighton John, a specialist in the theory and practice of digital forensics in the context of personal, cultural and scientific archives. The report is published by the DPC in association with Charles Beagrie Ltd. Neil Beagrie, Director of Consultancy at Charles Beagrie Ltd, was commissioned to act as principal investigator for, and managing editor of this Series in 2011. He has been further supported by an Editorial Board drawn from DPC members and peer reviewers who comment on text prior to release: William Kilbride (Chair), Neil Beagrie (Managing Editor), Janet Delve (University of Portsmouth), Sarah Higgins (University of Aberystwyth), Tim Keefe (Trinity College Dublin), Andrew McHugh (University of Glasgow) and Dave Thompson (Wellcome Library).

## Acknowledgements

This paper has benefited from the input of people occupying a range of professions. The Digital Forensics for Preservation briefing day held by the DPC provided an excellent foundation, and I am indebted to all those who attended and contributed to it, most notably: Cal Lee, Michael Olson and Kam Woods who travelled all the way over the Atlantic. Rachel Beagrie kindly provided her useful notes of the day. I am extremely grateful to William Kilbride who chaired the event and supported the writing of this paper.

Many colleagues and friends have kindly shared their thoughts, links and papers (including in some cases, unpublished drafts): Peter Chan, Aly Conteh, Tom Cramer, Bradley Daigle, Ifor ap Dafydd, Angela Dappert, Rachel Donahue, Erika Farr, Simson Garfinkel, Brad Glisson, Tim Gollins, Matthew Kirschenbaum, Gareth Knight, Andrew Jackson, Adam Jansen, Leslie Johnston, Akiko Kimura, Kari Kraus, Cal Lee, Jerome McDonough, Courtney Mumma, Naomi L. Nelson, Michael Olson, Erin O'Meara, Richard Ovenden, David Pearson, Michael Pearson, Gabriela Redwine, Elfrida Roberts, Corinne Rogers, Seth Shaw, Susan Thomas, Simon Wilson and Kam Woods. Professor Luciana Duranti and the University of British Columbia chapter of the Association of Canadian Archivists provided helpful encouragement.

From the UK forensics and security communities, I should like to acknowledge with gratitude the following researchers and practitioners for their invaluable introductions to the field: Bill Crane, Chris Hargreaves, Geoff Fellows, Brian Jenkinson, Russell May, Tony Sammes, Angela Sasse and Lindy Sheppard. The author is indebted to two anonymous referees for their extensive and careful reviews.

At the British Library, special thanks are directed to Chris Clark and Adam Farquhar who played an instrumental role in the establishment of the eMSS Lab, as well as to colleagues in the newly established Digital Research & Curator Team and Department of Digital Scholarship. Often overlooked is the understated but essential support of very senior colleagues at my institution, who have enabled digital forensics to flourish at the British Library through the provision of the necessary resources for space, equipment and software.

Neil Beagrie has been a truly supportive, patient and friendly editor, and approached me with the idea for this *Technology Watch* topic. He has been encouraging me to write about digital forensics for many years, and at last he may have succeeded.

Jeremy Leighton John
October 2012

The DPC gratefully acknowledges the support of the British Library in presenting this report.

## Contents

## Abstract

In recent years, digital forensics has emerged as an essential source of tools and approaches for facilitating digital preservation and curation, specifically for protecting and investigating evidence from the past. Institutional repositories and professionals with responsibilities for personal archives can benefit from forensics in addressing digital authenticity, accountability and accessibility. Digital personal information must be handled with due sensitivity and security while demonstrably protecting its evidential value. Forensic technology makes it possible to: identify privacy issues; establish a chain of custody for provenance; employ write protection for capture and transfer; and detect forgery or manipulation. It can extract and mine relevant metadata and content; enable efficient indexing and searching by curators; and facilitate audit control and granular access privileges. Advancing capabilities promise increasingly effective automation in the handling of ever higher volumes of personal digital information. With the right policies in place, the judicious use of forensic technologies will continue to offer theoretical models, practical solutions and analytical insights. The purpose of this paper is to provide a broad overview of digital forensics, with some pointers to resources and tools that may benefit cultural heritage and, specifically, the curation of personal digital archives.

## Executive Summary

Digital forensics is associated in many people's minds primarily with the investigation of crime. However, it has also emerged in recent years as a promising source of tools and approaches for facilitating digital preservation and curation, specifically for protecting and investigating evidence from the past.

This report provides a broad overview of digital forensics with pointers to resources and tools that may benefit the preservation of digital cultural heritage. More specifically, the report focuses on the application of digital forensics to the curation of personal digital archives.

Personal digital archives are very complex: the diversity of objects and intricacy of their structural relationships present significant challenges to curation. The nature of personal digital archives reflects both the evolution of technology and its associated social and political impact. Almost anything may appear in a personal digital archive, from emails and poet's drafts, through an astronomer's datasets, to digital workings of the mathematician, and notes of the political reformer. Forensic procedures tested and developed in this context may well be transferable to other areas of digital preservation and scholarship. With their diverse content, organization and ancestry, personal digital archives are the epitome of unstructured information and may serve as a test bed for refining forensic techniques in a curatorial context, as well as being an invaluable primary source of information for analysis.

There are three basic and essential principles in digital forensics: that the evidence is acquired without altering it; that this is demonstrably so; and that analysis is conducted in an accountable and repeatable way. Digital forensic processes, hardware and software have been designed to ensure compliance with these requirements.

Digital forensics is applicable throughout the curatorial and preservation lifecycle. A representative forensic lifecycle for a hard drive would be as follows:

- remove the disk from the originator's computer (it may join the institution's collections even if the computer does not do so);
- attach this collection disk to a curatorial computer via an intervening writeblocker (see Glossary) that prevents the disk from being altered;
- capture a forensic image (see Glossary) of the disk that represents the entire contents of the disk;
- create cryptographic hash values (see Glossary) for each and every digital object and for the disk itself;
- test for malware such as viruses;
- view files;
- view metadata such as the date and time when a file was created;
- extract metadata and pass to metadata marshalling or cataloguing system;
- identify and bookmark privacy concerns, e.g. files with credit card numbers or home addresses;

- export replicates of original objects and of disk image and examine in an emulator;
- convert digital replicates to modern interoperable files known as digital facsimiles that comply with digital preservation guidelines;
- analyse metadata and the content of objects and create exploratory visualizations;
- save log files of the examination process; and
- create a forensic report as documentation of the capture and analysis by the curator.

Information assurance (see Glossary) is critical. Writeblockers ensure that information is captured without altering it, while chains of custody (see Glossary), systems of evidence handling, process control, information audit, digital signatures and watermarking can protect the historical evidence from future alteration and uncertain provenance.

Selective redaction, anonymization and encryption, malware sandbox containment (see Glossary) and other mechanisms for security and fine-tuned control may be required to assure that privacy is fully protected and inadvertent information leakage is prevented. Family computers, portable devices and shareable cloud services all harbour considerable personal information and consequently raise issues of privacy. Digital archivists and forensic practitioners share the need to handle the ensuing personal information responsibly.

The current emphasis on automation in digital forensic research is of particular significance to the curation of cultural heritage, where this capability is increasingly essential in a digital universe that continues to expand exponentially. Current research is directed at handling large volumes efficiently and effectively using a variety of analytical techniques. Parallel processing, for example, through purpose-designed Graphics Processing Units (GPUs), and high performance computing can assist processor-intensive activities such as full search and indexing, filtering and hashing, secure deletion, mining, fusion and visualization.

Digital curation may also be able to derive considerable efficiencies from operating with disk images and using pointers (that reference digital objects and content within a disk image), and not necessarily exporting a multitude of digital objects.

An extensive, complex and detailed forensic study can be very time consuming. In the context of a serious crime, law enforcement may feel duty bound to follow the trail of evidence as far as possible, with less heed to the ultimate cost in resources. In the archival context, while the Digital Capture Imperative requires that the information is safely and accountably secured for the future, the actual forensic analysis may be tailored.

Forensic technologies vary greatly in their capability, cost and complexity. Some equipment is expensive, but some is free. Some techniques are very straightforward to use, others have to be applied with great care and sophistication. There is an increasingly rich set of open source forensic tools that are free to obtain and use. These are a wonderful introduction to the ins-and-outs of digital forensics, and can be used to compare and cross-check the outputs of commercial or other open source tools. They should, for example, produce the same hash values. A healthy forensic ecosystem should include a mix of commercial and open source software, and the strengthening of the forensic open source community is a vital component of scientific assurance which needs to be supported by the preservation community.

Digital archivists and forensic specialists share a common need to monitor and understand how technology is used to create, store, and manage digital information. Additionally, there is a mutual need to manage that information responsibly in conformance with relevant standards and best practice. New forensic techniques are furthering the handling of digital information from mobile devices, networks, live data on remote computers, flash media, virtual machines, cloud services, and encrypted sources. Forensic and archival methodology must retain the ability both to retrospectively interpret events represented on digital devices, and to react quickly to the changing digital landscape by the rapid institution of certifiable and responsible policies, procedures and facilities. The pace of change also has implications for ongoing training of curators and archivists, and there are digital forensics courses endorsed by archival, scholarly and preservation institutions.

In conclusion, there are some deep challenges ahead for cultural heritage and archives, but the forensic perspective is undoubtedly among the most promising sources of insights and solutions. Equally, digital forensics can benefit from the advances being made in the curation and preservation of digital information.

# 1. Introduction

## 1.1. Cultural Data, Personal Data

Digital scholarship and digital humanities are becoming increasingly influential. This is manifested in the emergence of large-scale, computer-intensive humanities projects, Big Humanities, oriented around textual databases and other cultural corpora containing millions of entities, from books to spoken words (Hand, 2011; Leetaru, 2011). Thus these fields of study are beginning to match the handling and analysis of vast volumes of generally well-structured bioinformatic, astronomical and other scientific data.

At the same time, the World Economic Forum has recognized the emergence of 'personal data' as a new and valuable asset class (World Economic Forum, 2011). The International Data Corporation (IDC) has stressed the increasing quantities of information generated by individuals, and concomitant emergence of Big Data Analysis (Gantz and Reinsel, 2011). Meanwhile, a special report on personal technology suggests that '[t]echnology will become even more personal', and portable digital devices converging on ubiquitous computing with personal sensors and mobile connectivity involving millions and ultimately billions of people 'will have a profound impact on the world'(*Economist*, 2011).

These discussions by the IDC and others tend to emphasize the immediate social and commercial analysis of personal information. From the perspective of digital scholarship and science, the raw information will be just as valuable, if not more so, over time - as a burgeoning primary resource for research, subject to ongoing historical, scientific and policy analysis. The field of personal informatics (see Glossary) will in time embrace the distant past as well as the present.

Moreover, this information almost invariably has little or no structure, and its capture, preservation and historical processing and interpretation demand new techniques and adaptable workflows.

## 1.2. Risks to Digital Memory

Memory institutions which are responsible for personal archives already have to care for personal digital objects. Digital preservation is concerned with the sustainability of digital information, notably the resilience and perceptibility of digital objects in the long term. A subtle risk to cultural heritage lies in the erosion of the evidential quality of digital objects, through an uncertain authenticity and completeness of content.

Historical analysis is founded on primary sources, the most valuable of which are archival in nature. Personal archives are extremely diverse in their organization and content, and in the manner and timing of their arrival. Almost any kind of digital object may reside in them. Critically, their historical and scholarly value depends on the evidential properties of these digital objects – reflected in, for example, the nature of embedded metadata.

Another significant risk lies in the loss of trust of potential donors of personal archives, and other people such as third party individuals in email correspondence, due to a failure to protect privacy and other digital rights, and again an understanding of pertinent ancillary and integral information is essential.

## 1.3. Curatorial Forensics

Some repositories have turned to digital forensics, realizing that this approach offers some significant solutions for the effective curation of archives, and for the automated, and quasi-automated, management and analysis of collections (John, 2008; John, 2009; Kirschenbaum *et al*., 2009a; Kirschenbaum *et al*., 2010; Olson 2010; Redwine *et al*., 2010; Carroll *et al*., 2011; Thomas 2011; AIMS, 2012). Forensic tools are potentially useful across a range of digital preservation functions, especially with complex digital objects such as web archives, software applications and systems. They are also highly pertinent in understanding the consequences of file format migration, and in determining the reliability of an emulator: in short, for evaluating digital preservation tools themselves. Conversely, advancing preservation and curation methodologies can contribute significantly to digital forensic practice.

Archival science has also taken up digital forensics as a vital tool and approach (Duranti and Endicott-Popovsky, 2010). A newly formed section of the Archives and Records Association UK & Ireland, concerned with the interface between technology and archives, has incorporated digital forensics within its official remit[1].

## 1.4. Aims and Structure

The aim of this report is to provide an overview of the application of digital forensics in the context of cultural heritage, digital preservation and academic research, oriented towards the curation of personal archives.

Significant work has been done in the emerging field of archival and cultural forensics, but this has mostly concerned the forensics of desktop and laptop computers, entailing the analysis of static file systems situated in self-contained storage media. This report will take a broader view of the digital forensics landscape.

The field of digital forensics is introduced first. Different approaches are then discussed from the perspective of archivists, curators and scholars, noting emerging issues of privacy and legal compliance. Concluding sections look to the future and highlight resources for further reading.

Section 2 provides a brief introduction to digital forensics, while Section 3 introduces and explores the motivations for the use of forensics in the context of digital preservation and curation. Section 4 describes some techniques and tools of potential value to archival practice, and mentions some specialist and emerging topics as well as the continuing coexistence of commercial and open source

[1] Sarah Higgins, personal communication

software. Section 5 turns specifically to digital scholarship, and outlines emerging techniques useful for research in the domain of cultural heritage: use of virtual technologies to provide context, algorithms to study style, authorship, and creative drafting, and methods to advance information extraction, analysis and visualization, especially chronology and social association. Section 6 sets these tools in the wider policy, legal and ethical context. Section 7 gives an overview of future prospects and emerging trends.

## 2. Overview of Digital Forensics

This section introduces the field of digital forensics, examining its definition, scope, core processes and some of its standards.

### 2.1. Definitions

A frequently cited definition for Digital Forensic Science is that of the Digital Forensic Research Workshop (DFRWS) of 2001: ‘The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations’ (DFRWS, 2001).

This definition focuses on criminal and unauthorized actions, but others place less emphasis on this aspect. For example, SY Willassen and SF Mjølsnes (2005), by omitting a reference to criminality, effectively focus an otherwise reminiscent definition on to the reconstruction of events: ‘Digital forensics can be defined as the practice of scientifically derived and proven technical methods and tools toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of after-the-fact digital information derived from digital sources for the purpose of facilitating or furthering the reconstruction of events as forensic evidence’.

The term derives from the Latin word ‘forensis’, referring to the forum, and while dictionary definitions of ‘forensics’ typically specify legal processes, it is also used (to some extent metaphorically) to allude to the notion of exhaustive investigation and argument. A ‘forensic society’ in the USA is equivalent to a debating society in the UK. Other investigatory contexts include the nature of an accident (e.g. an aeroplane crash), and the effectiveness of an individual’s use of equipment or procedure (e.g. during a flight simulation or command and control process) (Dussault and Maciag, 2004).

The use of digital forensic techniques and technologies in the context of archives and cultural heritage inevitably calls for a broadening of the term. For the purposes of this paper, ‘forensics’ essentially refers to the process of in depth analysis of information that exists in the present, in order to reconstruct past events or objects, with the proffered interpretation being subject to scrutiny by others (in some kind of 'forum', general public or specialist public).

## 2.2. Scope and Topics

Forensic computing and computer forensics are seen as essentially synonymous by practitioners, and can be distinguished from computational forensics, which is directed at the use of computing technology in forensics generally (Sammes and Jenkinson, 2007). Digital forensics is an extension of computer forensics, incorporating not only computers but any digital electronic technology, from mobile phones to printers.

Consequently, digital forensics is a very broad subject. Its various activities have been categorized in numerous ways. The 2001 DFRWS road map for digital forensic research distinguished between: (i) media analysis, (ii) code analysis, and (iii) network analysis (DFRWS, 2001).

Forensic sciences may be subdivided according to their 'domain of evidence', the most obvious subdivision being that between digital and analogue (Böhme *et al*., 2009), but 'domain of evidence' might be used to subdivide digital forensics itself, thereby yielding, for didactic purposes, three general areas: (i) desktop and laptop computers, media storage and file system (hard drives, optical discs, and floppy disks); (ii) networks, routers, servers, tapes and computer memory; and (iii) mobile, handheld and embedded systems.

In addition, four basic operational distinctions provide a useful way to specify approaches: (i) live or not (i.e. volatile or static, loosely contrasting information in memory which is lost when power is unavailable with information that persists in media storage such as disk and tape) (Inoue *et al*., 2011); (ii) free and open source software (FOSS) or closed and proprietary (an imperfect but understandable dichotomy); (iii) multimedia or not (i.e. sensor-based system and indeterminate, or finite system and determinate) (Böhme *et al*., 2009) and (iv) file based or not (i.e. oriented towards files or bulk data analysis, regardless of partition structure and file system metadata) (Garfinkel, 2011).

Other classifications and specialities may be apparent for professional practice (e.g. electronic discovery in a corporate context or incident response for security), techniques and procedures (e.g. network analysis, text mining – see Glossary – and regular expressions), specific media and hardware (e.g. optical discs and flash media), specific types of digital objects and applications (e.g. emails).

## 2.3. Core Forensic Process

The forensic process is outlined by JL John (2008) and MG Kirschenbaum *et al*. (2010). A set of three principles lies at the core of computer forensics (e.g. Casey, 2002b; Kruse, II and Heiser, 2002; Sammes and Jenkinson, 2007), and may be paraphrased as follows: (i) acquire the evidence without altering or damaging the original; (ii) establish and demonstrate that the examined evidence is the same as that which was originally obtained; (iii) analyse the evidence in an accountable and repeatable fashion.

These principles have immediate and important practical implications for archivists and curators. For example, faced with a desktop computer a forensic examiner or archivist wanting to undertake a thorough forensic examination would avoid turning the computer on, and would very likely remove the hard drive from the computer. This subject disk would be attached to a specially configured examination computer which would be used to capture and analyse the contents of the disk. An intervening writeblocker has to be installed to ensure that the examination computer does not alter the contents of the subject disk during the process.

At the time of capture it is common for hash values (akin to digital fingerprints) to be created for the entire disk and for each of the digital objects (files) contained within it. If a digital object is subjected to the same hashing algorithm at a later date and the same hash value is obtained, it may be concluded that the object has not changed.

Other functionalities include: file viewing, analysis of file signatures, date-time interpretation, identification of known files (e.g. operating system files) by means of hash libraries, data extraction, file export, searching, indexing, bookmarking, timeline visualization, logging, and reporting, all of which are to be undertaken according to forensically sound principles.

## 2.4. Standards and Testing

### 2.4.1. Standards

Computer forensic practitioners and scientists are routinely expected to meet specific standards in order to satisfy legal authorities. Forensic technologies are regularly justified in court. This stimulates an ongoing and independent and rigorous assessment of tools, and means that practitioners can have significant confidence in tested digital forensic tools. Such confidence is always provisional however, and techniques still have to be applied properly (John *et al*., 2010).

Standards for handling and processing digital evidence have emerged mainly from law enforcement and associated organizations. Notable examples include guidelines produced by the Association of Chief Police Officers (ACPO) in the UK and the National Institute of Justice (NIJ) in the USA. Another source of guidance is the International Organization on Digital Evidence (IOCE: originally called International Organization on Computer Evidence, it retains the acronym). In addition, the Network Working Group of the Internet Engineering Task Force issued a Request for Comments on Guidelines for Evidence Collection and Archiving (RFC 3227, IETF 2002). Similarly, a series of codes of practice have been prepared by the British Standards Institution (BSI) for legal admissibility and evidential weight of digital information, the second edition of which has been recommended by the Lord Chancellor’s Office, a forerunner of the current Ministry of Justice (Shipman, 2004; Shipman and Howes, 2005; BIP, 2008).

The Scientific Working Group for Digital Evidence (SWDGE) has a prominent role in the USA in establishing standards for recovering, preserving, and examining digital evidence. The Digital Evidence Group (DEG) has played a corresponding function in the UK, providing expert opinion and advice regarding the acquisition processing, handling and preserving of evidence, and in various ways the Home Office continues to seek and collate expert advice and guidance concerning digital evidence, security and privacy issues.

The requirement for scientific procedures and methodologies has been emphasized by a number of court cases, notably Frye and Daubert in the USA. In 1923 the Court of Appeal in the District of Columbia ruled that where a scientific procedure has been accepted by the relevant scientific community, the court would defer to this ‘general acceptance’ (Frye vs. United States). This was qualified by the case of Daubert vs. Merrell Dow Pharmaceuticals, 1993, it being established that the judge, acting as a gatekeeper to scientific evidence, would be required to consider not only ‘general acceptance’ but ‘testability, peer review, known error rates, and existence of standards’ (Bell, 2008). Although these laws pertain to the USA, they are of some relevance to other jurisdictions due to the predominance of forensic techniques and technologies emanating from there.

In addition to standards for technologies, the standard ISO[2] 17025:2005 supports and motivates the competence of testing and calibration in scientific laboratories (UNIDO, 2009).

### 2.4.2. Testing

The National Institute of Standards and Technologies (NIST) is one of the principal governing bodies responsible for setting standards in the USA, and is sponsoring a project called Computer Forensics Tool Testing (CFTT) to oversee and coordinate research on computing forensic tools. NIST has created a general approach for testing computer forensic tools, with formal testing criteria (Bryson and Stevens, 2002).

NIST has also developed several mechanisms for evaluating disk drive imaging devices – Forensic Software Testing Support Tools. NIJ works with NIST and other federal partners to develop methods to test commercial forensic software, and has also developed methods and training programs for computer investigations and forensic analysis. A recent conference proposed an extensible and common scheme for evaluating and benchmarking forensic software as well as creating a preliminary development framework (Hildebrandt *et al*., 2011). A wide variety of devices continues to be tested over the years (CFTT, 2012).

Besides testing by official bodies, the most effective digital forensic practitioners and researchers routinely evaluate tools themselves in order to possess the first-hand experience necessary to be able to defend procedures in court.

[2] http://www.iso.org/

## 3. Digital Curation and Forensic Possibilities

This section introduces the benefits to archival practice in adopting digital forensics, and briefly notes the handling of digital evidence, the longstanding applicability of the 'multi-evidential' approach in detecting forgery and in forensics generally. The ongoing relevance of forensic expertise in investigating ancestral computers is noted.

### 3.1. Distinction of Curatorial Forensics

Conventional uses of digital forensics commonly pertain to a unique and specific context, such as a court case. However, digital preservation assumes that an archival resource will be used and reused by many potential users, for diverse purposes, and over an indefinite period. This outlook towards digital preservation and access is the most distinctive aspect of curatorial forensics in comparison with forensic practice generally.

### 3.2. Motivations for Archival Repositories

Some of the initial motivations for curators adopting forensic technology have been:

- prevention of changes to dates and time, location data and other associated data;
- application of searching and indexing capability for locating private content including credit card numbers, postal and email addresses and so on;
- recovery and diligent processing of compound files such as Microsoft Word documents with embedded earlier drafts and content;
- identification and protection of authenticity along with detection of digital forgery;
- adoption of contextual perspective with capture not only of individual digital objects but of whole disks and of an entire collection of personal media; and
- assimilation of the evidential value of layout, ornament and style.

### 3.3. Adoption of Digital Forensics in an Archival Context

Over recent years digital forensics has moved into the archival and curatorial universe. Significant and early contributions have been made in digital record forensics by Luciana Duranti, most especially in diplomatics with a legal context (see Glossary), following on from Elizabeth Diamond's perusal of record keeping in a forensic light (Diamond, 1994;  Duranti, 2009; Duranti and Endicott-Popovsky, 2010). Alastair Irons (2006) observed the consonance between records management and computer forensics. Curators and archivists have put the approach into practice at the British Library, the Bodleian and other libraries in the UK, and at archival repositories worldwide, notably Stanford University Libraries (John, 2008, 2009; Redwine *et al*. 2010; Thomas, 2011). Perhaps most tellingly of all, the scholars themselves are advocating and adopting the approach, most prominently at the University of Maryland (Kirschenbaum, 2008).

Notwithstanding the different perspectives and objectives, the digital preservation community has long employed techniques and concepts that are closely allied to forensics (Dappert *et al*., 2011), as do many computer security practitioners. Recently, interest in digital forensics specifically has intensified in this community too. In a report on the preservation of virtual worlds, it was observed that forensic tools may be uniquely useful in appraising the authenticity and provenance of computer gaming material received from individuals in the player community (McDonough *et al.*, 2010).

Digital forensics researchers have also become interested in the archival context, notably Simson Garfinkel who gave a keynote paper at the 2009 Digital Lives Research Conference, and who is contributing to the BitCurator project, and Phil Turner who attended the 2010 Digital Lives Research Seminar [3]. Tony Sammes and Brian Jenkinson (formerly of the Defence Academy of the United Kingdom and currently affiliated with DeMontfort University, Leicester) are sharing with the British Library their methods in the application of forensics with ancestral computers.

## 3.4. Due Diligence in the Security of Evidence

Considerable effort in the forensic community is directed at guiding the first responders, the people who collect evidence at the scene of investigation; but evidence handling continues throughout the lifecycle, and many of the procedures and stipulations will strike a chord with archivists and curators as well as digital preservation technologists.

It is strongly recommended by authorities that the primary evidence is stored in a dedicated safe or strong room to which access is controlled and documented by ‘evidence custodians’ (Mandia *et al.*, 2003).

The Digital Evidence Bags (DEBs) proposed by Philip Turner are aimed at ‘bundling digital evidence, associated metadata, and audit logs into a single structure’ (Richard III and Roussev, 2006). Under this scheme, the audit log is updated whenever evidence is transferred to and from, and processed within, the universal container, the DEB (P Turner, 2005).

The management and storage of digital information is a key research requirement in forensic science generally. A number of digital systems have been designed to assist in the creation of documentation and data, evidence tracking, chain of custody and case management (including HOLMES2, LOCARD and Digital Investigation Manager)[4].

Archival science, and digital curation and preservation fields of research have much to offer to this topic (e.g. Duranti, 2009).

[3] http://britishlibrary.typepad.co.uk/files/digital-lives-seminar-5july2010-v8-1.pdf

[4] HOLMES2, http://www.holmes2.com/holmes2/whatish2; LOCARD: Evidence Tracking System, http://www.locard.co.uk; Anite, http://www.anite.com/secure-information-solutions-products-anite.html?Itemid1⁄4129; http://www.incman.dflabs.com/digitalinvestigationmanager.html

## 3.5. The Multi-Evidential Perspective

The role of established forensic principles in the digital arena is not straightforward since digital objects can be replicated exactly. Nonetheless, longstanding principles remain relevant, due to the scale, complexity and intertwining levels of abstraction that exist in modern computing systems. Notwithstanding the preeminence of the Exchange Principle ('every contact leaves a trace') commonly attributed to Edmond Locard (1877–1966), the fundamental cornerstone of forensic, and indeed any retrospective, analysis lies in the multi-evidential approach: the manner in which small, seemingly independent, extant traces serve to corroborate each other making it possible to build up a picture of past events or objects. Some understanding of this notion has existed since classical times (Nickell, 2005; Bell, 2008), as has been manifest in scholarly and scientific practice, and continues to be applicable in the digital era.

In the digital arena, attempts have been made to devise methods for estimating and categorizing uncertainty of digital data such as network logs, and for assessing reliability of evidence (Casey, 2000; Casey, 2002a). Eoghan Casey (2000) produced a chart for scaling certainty ranging from C0 (evidence contradicts facts) through to C6 (the evidence is tamper proof and unquestionable, a theoretical goal for the future). Appropriately, levels C4 and C5 both incorporate a requirement of 'multiple, independent sources of evidence' that agree[5].

It may be useful to emphasize, therefore, that there are two core purposes to digital forensics, for in addition to the appraisal and protection of evidential value such as embedded metadata and other latent information, forensic science seeks to reconstruct events and objects from information that is no longer otherwise wholly available or coherent, where overt contextual information has been lost or become obscure.

## 3.6. Date and Time

One consideration which highlights more than any other the need for digital forensics in cultural and historical contexts is the challenge of establishing the timing of events in sequence and in relation to a reference time such as Universal Coordinated Time (UTC, the acronym being a compromise between English and French). The operating system, the file system, the file formats, the condition and nature of networks, and the assiduousness, preferences and mobility of the user, and the type of media on which a file is located (e.g. external, removable or internal), all influence the forensic interpretation of date and time (Willassen, 2008, Casey, 2010).

Thus in some cases the determination of time can be a labyrinthine task, and forensic specialists may need to understand not only what software is doing in the real world, but also what the forensic tool is doing in the laboratory. Nearly every digital object has some kind of date and time, and using the multi-evidential approach of forensics it is possible to establish dates and times with some reliability, and with measures of accuracy. Corroboration can be sought not only internally but externally, in an approach that is strongly reminiscent of scholarly methods of textual analysis.

[5] The author thanks an anonymous referee for pointing out the certainty scale.

### 3.7. Forensics of Ancestral Computers and Code

The rapid pace at which technology advances means that forensic researchers are continually exploring and developing new techniques and tools. Yet the forensics of early computers remains directly relevant and applicable to criminal cases (personal communication, T Sammes).

With time, a modern forensic community will tend to become less focused on earlier computing technologies. Even so, these will remain crucial to scholars such as historians and archivists, and not just historians of computer science. This requirement will motivate continuing research into the forensics of ancestral computers and other digital devices, as well as code[6].

A specific issue is the question of changing forensic standards as time passes.

## 4. Archival Forensics in Practice: Procedures and Tools

This section outlines the practice and lifecycle of archival forensics, introduces a series of tools, both general and specialist, and discusses the open source approach to forensics. It also draws a distinction between digital conservation and digital archaeology.

### 4.1. Lifecycle: Archival and Curatorial

A few words about archival and curatorial lifecycles may be useful for the digital forensics scientist and practitioner.

The almost universally adopted archival model is the OAIS Reference Model which originated within the space science community. It has become an ISO standard (ISO 14721:2003) and continues to evolve with input from a variety of interested parties, notably specialists concerned with digital preservation and digital library systems, as well as archivists and curators of scientific data, institutional records and personal archives. One change in emphasis has been a greater recognition of emulation alongside file format migration as a valuable approach to digital preservation (Farquhar and Hockx-Yu, 2007).

The Digital Curation Centre (DCC) produced the DCC Lifecycle Model, which provides an excellent introduction to archival thinking and concepts, and is being used for training purposes (Higgins, 2008). The InterPARES projects (International Research on Permanent Authentic Records in Electronic Systems) have produced an exceedingly rich resource for archival and records management lifecycles and practices in a digital context[7].

[6] A joint paper by Corinne Rogers and Jeremy Leighton John elaborates on the theme, to be published under the auspices of the UNESCO Memory of the World conference 2012, Vancouver

[7] http://www.interpares.org

Keywords and a paraphrasing of the DCC model may provide an illustration[8]:

- **Conceptualize** (conceive data, capture and storage)
- **Create or Receive** (create or receive data including metadata – administrative, descriptive, structural, technical)
- **Appraise and Select** (evaluate and select data for long term curation)
- **Ingest** (transfer information to repository)
- **Preservation Action** (undertake actions for long-term preservation and retention of authenticity including data cleaning, validation, and provision of suitable data structures and file formats)
- **Store** (store data securely)
- **Access, Use and Reuse** (ensure the accessibility of the information on a day-to-day basis, with necessary robust access controls)
- **Transform** (create newly modified data or subsets of data for specific purposes, e.g. publication)

At each stage, processes are conducted in accordance with documented guidelines, policies and legal requirements.

## 4.2. Lifecycle: Digital Forensics

A quick introduction with a model forensic lifecycle may be helpful before discussing tools. There are a number of models due to the diverse situations that digital forensics must address. For simplicity, incident response methodology is overlooked, and an amalgamation of models derived from the Association of Chief Police Officers (ACPO), US Department of Justice and US Air Force (see ACPO, undated; Janssen and Ayers, 2007) and a reading of T Sammes and B Jenkinson (2007) may serve as an illustrative sequence for this paper:

- **Identification** (recognize incident, requirement for action, intelligence for investigation)

[8] http://www.dcc.ac.uk

- **Authorization** (approval)

- **Preparation** (intelligence for search, adequate toolkits, operational briefing, task allocation)

- **Securing and Evaluating the Scene** (ensure safety, confirm computer equipment present and recognize further possibilities, secure equipment, identify and protect evidence, conduct interviews)

- **Documenting the Scene** (create a permanent record of the scene by means of photography and note taking, document condition and location of computers and related components whether these are to be removed or not, mark and label artefacts, use seals and sealable containers, evidence bags)

- **Evidence Collection** (cater for computer devices found to be switched on or off, attending to order of volatility (see Glossary), collect computer hardware and media while preserving evidential value, obtain analogue evidence such as passwords, handwritten notes, computer manuals, printouts)

- **Packaging, Transportation and Storage** (protect equipment and media during transfer avoiding extreme temperatures, physical impact and vibration, static electricity and magnetic sources, establish procedures for reception and storage of machines and media, maintain chain of custody, inventory for storage in secure area free of contaminants)

- **Initial Inspection** (identification of devices, external and internal physical examination of computers, tool selection and expectations)

- **Forensic Imaging and Copying** (e.g. for hard drive – removal of physical disk from computer, digital preview and capture using physical or logical disk acquisition, with writeblockers, followed by return of original media to evidence custodian)

- **Forensic Examination and Analysis** (use forensic techniques and tools for analysis and processing including: creation of cryptographic hash values and filtering with hash libraries, file viewing, file exporting and expansion of compound files (e.g. email), extraction of metadata, searching and indexing)

- **Presentation and Report** (document procedures, analysis and findings, use log files, bookmarks and notes made during the examination, make conclusions, prepare exhibits suitable for court)

The topic of digital investigation process models is covered in some detail by E Casey and B Schatz (2011).

## 4.3. Mutual Incorporation of Procedures

Although not identical, the parallels between the two lifecycles are striking, and speak for some natural integration of forensic and archival workflows (as is being actively pursued by the Digital Records Forensics Project and others). The forensic workflow places more emphasis and detail in the preparation for collection and its actual conduct, while the archival workflow is oriented more towards long-term preservation and reuse. Appraisal and selection of evidence, data and records, and the maintenance of provenance, a chain of custody, are prominent in both fields.

Initial incorporation of forensic procedures in workflows for personal archives has been outlined previously (e.g. John, 2008; John *et al.*, 2010; AIMS, 2012) but more detailed work remains to be done.

Every step of the archival lifecycle may be influenced by the forensic approach. Even the integration of digital with analogue is embraced by the forensic workflow, with digitized objects being imported into the digital forensic case – accordingly, Forensic Toolkit (FTK) has an OCR (Optical Character Recognition) capability.

Although the curatorial site visit can be seen quite naturally as part of the forensic life cycle, seeking as it does the context of the digital life, much of this activity is outside the scope of this paper. The obvious exception is the need to document the computer, hardware and network system that the originator of the archive has used during their life.

For the purposes of this report, the following simplified but focused workflow may suffice (the workflow contemplates a hard drive for illustration).

**Digital Capture**

- **Arrival and Transfer to Holding Storage** (securely held within holding system, e.g. Digital Objects Curatorial System at the British Library, registered with evidence custodian, for examination and processing prior to transfer to digital repository system)
- **Inventory on Reception** (naming and registration of all digital media and hardware)
- **Digital Intake** (inspection, boxing and labelling of equipment and media)
- **Digital Acquisition: Physical or Logical** (e.g. forensic imaging of entire hard drive (physical acquisition) or set of files on hard drive (logical acquisition) with hash values created for every digital object, repeated to ensure that same hash values are obtained using independent forensic software and hardware)

### Processing

- **Inspection and Appraisal** (initial inspection of disk image and digital objects within it, prioritizing and filtering of digital objects such as software using hash libraries, malware checking)
- **Metadata Extraction** (metadata are examined and extracted, e.g. for selective passing to a catalogue system)
- **Digital Replicates: Export** (export of exact replicates of digital objects, e.g. for use in an emulator, for preservation characterization, and as exact copies of the original objects)
- **Digital Elements: Formation** (expansion of compound files, or complex digital objects, e.g. word-processed documents attached to email messages must be separated out for preservation)
- **Digital Facsimiles: Conversion, Migration** (conversion of digital replicates and elements to a modern interoperable type for long-term preservation)

### Curatorial Examination and Analysis

- **Emulation with Digital Replicates and Disk Images** (explore original look and feel using emulators and virtual machines)
- **Metadata Creation and Content Elucidation** (curatorial examination of content, creation of metadata and archival description, with special attention directed at privacy and digital rights annotation, using searching and indexing functionality)
- **Visualization and Content Analysis** (including chronological mapping, visual analytics, multimedia and text mining)
- **Restriction, Redaction and Selective Encryption: Access Versions** (where necessary access versions of digital objects and of disk images are created with information redacted or in some way restricted)

### Digital Archival Storage System

- **Consolidation and Package Preparation** (preparation of disk images and digital objects for transfer into a digital repository system, including metadata pertaining to curatorial forensic processing, e.g. Digital Library System of the British Library)
- **Transfer and Ingest to Digital Repository System** (actual ingest into the digital repository system)

### Access and Resource Discovery

- **Concluding Preparations for Discovery** (including digital policy compliance and release authorization)
- **Enabling Access** (uploading or linking for reading room, institutional or online access)

Ongoing measures include audit and enhancement of preservation and curation with repeatable evaluation of procedures and outcomes.

A potentially valuable procedure that has not been much commented on in an archival context is the possibility of incorporating a series of floppy disks, optical discs and other media within a single disk image (or series of segments of a disk image) facilitating their joint analysis and investigation (see John, 2008). Written and photographic records by the curator during a site visit to the originator's home similarly may be introduced.

## 4.4. Toolkits for the Lab

Digital forensics is a fast moving area with the key tool producers releasing new features in their products frequently. There is endless debate in the forensic community regarding their virtues. One of the professional roles of a digital curator is to monitor changing technology.

Smaller institutions will be able to do much with a combination of free and inexpensive tools (writeblockers, open source software, FTK Imager and others); larger institutions may be able to justify greater expenditure in part so that a wide range of tools can be tried and tested for the benefit of the wider community, and as a means of cross checking analyses. Examination of the capability of existing forensic software provides useful insights, ideas that might be transferred to other contexts, enabling ongoing advancement.

Functionally, there are five categories of hardware: (i) computing machines for activities across the lifecycle from capture through analysis and presentation; (ii) media drives, interfaces and writeblockers for digital capture; (iii) reception storage for digital objects during and following capture; (iv) holding storage for high performance analysis, fast duplication and local network testing; (v) toolkits and consumables for dismantling and reconstructing devices, for cleaning, for processing and labelling, and for protection.

In essence, there are two kinds of forensic software: (i) comprehensive and integrated software more or less directed at the entire lifecycle; and (ii) specialist software for particular forensic purposes.

The two most established software packages remain EnCase by Guidance Software and FTK by AccessData. A third contender is Paraben's P2 Commander which essentially integrates a number of independent tools, with particular strengths in the forensics of emails and handheld devices. X-Ways Forensic Toolkit is an excellent fourth tool. Although less polished and advanced in its presentation, it is highly respected by many forensics experts. It is built upon the reputable WinHex tool (which is still available separately).

Two emerging sets of tools are provided by Digital Detective Group and PassMark Software. The FIDO project (Forensic Investigation of Digital Objects) funded by JISC has explored the use of

PassMark Software[9]. There are many specialist tools; one particularly useful one is Infinadyne's CD/DVD Inspector.

Several useful lists of tools and outlines of archival lifecycles that incorporate forensic activities have been published (PARADIGM, 2007; John, 2008; Garfinkel and Cox, 2009; John *et al*., 2010; Kirschenbaum *et al*., 2010; Thomas, 2011, AIMS, 2012;).

## 4.5. Emerging Forensics

### 4.5.1. Handheld Devices

Most archival forensics has been oriented towards the desktop and laptop and associated media, notably the hard drive, floppy disk and optical disc accessed and analysed through the file system by means of an operating system, commonly Microsoft Windows. Increasingly other forms of digital forensics will become essential for digital curation and preservation.

A pragmatic distinction can be made between computer forensics that does and does not require specific knowledge of the hardware (Van Der Knijff, 2010). Of most relevance to personal archives are those embedded systems that belong to the subcategory of small-scale digital devices, specifically handheld and tablet computers; personal digital assistants (PDAs); mobile phones; digital cameras and digital video (DV) cameras; GPS (Global Positioning System) receivers and transmitters, and digital audio recorders.

Although smartphones are relatively novel, handheld devices such as electronic organizers have been forensically investigated for many years (Sammes and Jenkinson, 2000; Sammes and Jenkinson, 2007). With the iPad and other tablets taking on aspects of the role of the PC in digital life, the more prevalent these devices become, the more archivally valuable will be information derived from them.

Two of the leaders in mobile forensics dedicated to this field are CelleBrite with its UFED Physical Analyzer and Micro Systemation which produces .XRY Complete. Within the broader forensic field, Paraben's Device Seizure has established a reputation in handheld forensics.

Aspects of flash media are reminiscent of disk forensics in the adoption of the FAT32 file system, but the autonomous processes of wear levelling (aimed at distributing the use of each area of the flash store equally), frequent updating and error checking makes for a dynamic and complex organization. It also means that the potential for recovery of earlier and duplicated versions of information is magnified, through physical extraction processes that are not normally available to the user. This may be countered somewhat by the nature of TRIM, a special command that indicates to the operating system which areas of data in solid state disks are no longer necessary (King and Vidas, 2011).

[9] http://fido.cerch.kcl.ac.uk/

The extensive diversity of handheld devices and interfaces means that there is an imperative for repositories to collect both data cables and power cables in readiness for future arrival of devices. Most providers of mobile forensic equipment will supply sets of cables, at least for those devices that are current or only recently obsolete. The British Library has for a number of years been gathering such cables for future use.

### 4.5.2. Mac Forensics

Tools for use with Microsoft Windows such as EnCase, FTK, Paraben and X-Ways have long been able – to some extent – to capture and interpret the file systems used by Apple Macintosh without being able to handle and fully process the files themselves. There have been limitations, and, in some instances, even partition structures are not identified or displayed correctly. It has been suggested that it is most effective to conduct forensics of modern Apple computers using a native Macintosh environment (Kokocinski, 2010).

Following the groundbreaking Expert Witness for the Macintosh (the forerunner of EnCase), the principal software specific to the Macintosh computer, and running on OS X systems, are BlackLight (replacing the Forensic Suite), MacQuisition of BlackBag Technologies and MacForensicsLab of SubRosa Software, using many of the operating system's own utilities. Another tool is Mac Marshall of ATC-NY Corp.

A distinct functionality that can be used to acquire Apple Mac systems is the Target Disk Mode. However, care is needed, because there are times when it is unavailable due to a firmware password, resulting in the operating system being engaged, and consequent write protection failure.

With the proliferation of Apple laptops and mobile devices, the iPod, iPhone and iPad, OS X and iOS have attracted much more attention (Joyce *et al*., 2008). It is increasingly expected that a forensic lab should be equipped with a Mac computer system for forensic investigation.

### 4.5.3. Unix Forensics, Linux Forensics

For about a decade the operating system of the Macintosh computer, OS X, has been founded on a Unix system codenamed Darwin. This alone points to the need for some familiarity with Unix (McElhearn, 2005).

Curators who work with scientific archives can certainly expect Unix systems and need to use the corresponding forensic tools (Pogue *et al*., 2008; Seglem *et al*., 2002; Altheide and Casey, 2010). At the British Library, two of the first computers to arrive in a personal archive were Silicon Graphics workstations with their own flavour of Unix. Linux systems are also popular with scientists.

Linux systems are favoured by many computer forensics practitioners, and there are some commercial tools. Forward Discovery has produced Raptor, a Linux-based tool for previewing and acquiring disks. A longstanding Linux tool for analysis as well as acquisition is SMART from ASR Data. BJ Grundy (2008) provides an introduction to the use of Linux and some of its own utilities.

### 4.5.4. Live Forensics

There is an increasing interest in the acquisition of live data, of the physical memory. In part this is in order to bypass encryption. While it may not be an immediate priority for digital curators, it may become useful in scenarios such as cloud computing. It is commonly used in seeking to understand and target malicious processes in Windows (Pittman and Shaver, 2010).

### 4.5.5. Remote Acquisition

Currently curators travel to the house of a writer or scientist in order to preview and capture the contents of a donor's personal digital objects. Potentially, time and expense could be saved if some of these activities could be conducted over the Internet as an aspect of online curation. Given the support and knowledge of the donor, standard means might include secure email (akin to services such as Voltage), uploading by the donor through a secure form of ftp, or the use of a remote access service such as LogMeIn.

Forensic technologies offer several key advantages: (i) forensically sound inspection and acquisition, (ii) security and control of access, and (iii) detailed auditing and logging of the activities of the examiner. For example, the Field Intelligence Model (FIM) of EnCase illustrates the use of a forensically sound and accountable authentication administration server, linked via secure connections over a network using a public key AES encryption system (Bunting and Wei, 2006). FIM is available to bona fide law enforcement and security professionals but the concept could be readily adopted for preservation and archival purposes with overt functionality.

## 4.6. Digital Conservation, Digital Archaeology

Most digital forensics is concerned with relatively healthy media, but it does encompass the investigation of damaged media and objects. Even when media are not degraded it is sometimes necessary to examine them physically at the microscopic scale as well as digitally. A classic digital preservation paper on the topic of digital archaeology describes the use of advanced microscopic techniques (Ross and Gow, 1999). It may be helpful to adopt the term 'digital conservation' (see Glossary) for those situations where the storage media and other hardware are significantly degraded or damaged, or where the media and hardware are being investigated at fundamental levels with a view to enhancing or expanding the recovery and preservation of digital information.

It seems sensible to reserve the word 'archaeology' (digital, media or otherwise) for the situation where fragments of information are not only recovered but used to investigate and interpret social circumstances, such as online communities or early computer game players, using the phrase 'digital conservation' for the recovery and care of the information itself.

Techniques for recovery from malfunctioning hard drives operate at several levels of intervention. For example, Disk Labs (UK) offers a data recovery service that extends to the rebuilding of a hard drive in a dust-free clean room.[10] The international standard for airborne particulate cleanliness is ISO 14644.[11] Some of the concepts of data recovery and the repair of hard disks are introduced by K Kaspersky (2006).

Besides the more extreme options of atomic microscopy for scanning the media surface and the replacement of components of a drive's electronics in a clean room, there are two less drastic measures worth mentioning. Firstly, the DeepSpar Disk Imager 3 has an ability to cope with bad sectors, while moderating any strain on the drive. Secondly, SignalTrace technology in combination with PRMLpro is said to be able to bypass key elements of drive electronics to read and decode data on the media surface. The SignalTrace technology appears to have been absorbed by Seagate where it resides as a service[12].

## 4.7. Open Source and Commercial Software

### 4.7.1. Open Source Digital Forensics Frameworks and Toolsets

An appealing introduction to the setting up of an open source examination platform is centred around Linux as a host, specifically Ubuntu; and incorporates the high-level programming languages Perl, Python and Ruby (Carvey and Altheide, 2011). It outlines the use of FUSE (File Systems in User Space) and associated modules that together provide great flexibility in interpreting file systems as well as volumes and containers (digital structures for holding digital objects and associated data in an orderly way) providing ready access to their contents; forensic programs in this context include MountEWF, AFFuse and XMount, all of which enable access to specific types of forensic images.

Prominent frameworks and toolsets (Carvey and Altheide, 2011; Huebner and Zanero, 2010) include:

- The Sleuth Kit (TSK) with Autopsy, an open source GUI (Graphical User Interface) as browser;
- AFFLib incorporating Advanced Forensic Format (AFF) (for introduction see John et al., 2010);
- Fiwalk or 'file&inode walk' directed at automated and rapid processing of disk images;
- PyFlag, aimed at unifying the forensic examination of diverse sources of data through a Python GUI;
- Digital Forensics Framework (DFF), otherwise known as Open Source Digital Investigation Framework from ArxSys;
- Open Computer Forensics Architecture (OCFA) a modular design by Dutch police for automating the analysis of large volumes of digital evidence; and

[10] http://www.disklabs.com/

[11] http://www.iest.org/StandardsRPs/ISOStandards/ISO14644Standards/tabid/10135/Default.aspx

[12] http://services.seagate.com/signal_trace.aspx

- Computer Aided Investigative Environment (CAINE).

The advancement of bulk data analysis as an efficient approach to digital forensics has yielded bulk_extractor, which is written in C++, along with associated tools for processing output (Garfinkel, 2011). It focuses on the extraction of salient features such as email addresses, GPS coordinates and credit card numbers, with filtering that attempts to take into account the local context of the features. An interesting challenge that this study tackles is the necessity of filtering the many extraneous email addresses that reside in system and application files.

These activities are most effective when they adopt components and modules from each other or enable such borrowing. Although C and C++ are faster, Python is a popular way to provide extensibility. For example, the Python module fiwalk.py makes it possible to design forensic tools that leverage fiwalk's capabilities; similarly, another project has developed pytsk in order to provide Python access to the SleuthKit libraries.

### 4.7.2. The Forensic Ecosystem

The relative merits of open source software and commercial, typically proprietary and closed, software are a matter of continuing discussion. Two landmark papers emphasize the crucial role of scrutiny of open source software and invoke, among other things, the Daubert requirements (Carrier, 2002; Kenneally, 2001). In the face of ever-increasing volumes of evidence and data and the necessity of rapid processing, many forensic researchers favour the greater adaptability and potential efficiencies of open source software. From the archival perspective, moreover, commercial software is generally expensive, tedious to license, and employs archivally discordant terminology. (Although this last disadvantage commonly applies to open source software too it may be easier to modify.)

Open source supports learning through open knowledge, flexibility of use, versatility at low cost, portability across systems, tool and error checking for diligence and adaptability (Huebner and Zanero, 2010; Carvey and Altheide, 2011), and a vibrant research community.

Despite the current weaknesses of commercial software products, there are clearly some strengths. Some of the products have existed for a long time, are familiar, and have established a reputation supported by regular scrutiny in legal courts. Notwithstanding the cost, training by commercial vendors does at least offer some introduction to the field of digital forensics for archivists and curators. A degree of healthy competition among vendors exists, which results in a manifest desire to support and assist the customer, and most tellingly the major tools strive to be comprehensive and integrated and have succeeded to some extent. Two examples of where a vendor has quickly adopted new techniques from academic research are the incorporation of AFF and fuzzy hashing (see Glossary) in FTK by AccessData, and the move by Technology Pathways to make it possible to output report material from ProDiscover to an ODBC (Object Database Connection) data source.

For the foreseeable future, the optimum scenario is the coexistence of open source with commercial software. On balance, the crucial necessity is to have a multiplicity of tools for meeting various functional requirements, for cross verifying outcomes, and for continuous evaluation within a healthy ecosystem of open source and commercial tools involving academic researchers and vendor developers.

A very important step is the BitCurator project led by the University of North Carolina at Chapel Hill and the University of Maryland.[13] As with any open source venture, it will depend greatly on the ongoing contributions of the archival and forensic community.

There are major challenges ahead for digital forensics generally, and these issues, together with a strong and growing research community, will likely change the business model and relationship with the commercial vendors. A survey conducted by PLANETS in another context, namely digital preservation, concluded: 'Open-source and proprietary software are used equally by respondents, and often combined in the same solution. In the future, respondents expect to continue using this pick and mix approach, with three fifths predicting that they will use a mixture of open-source and proprietary software. However, the proportion that will rely on purely proprietary solutions will decline seven-fold from 14% to 2%' (PLANETS, 2010).

## 5. Digital Scholarship, Digital Creativity

Having considered the potential of digital forensics in archival practice, this section examines the ways that digital forensics may serve digital scholarship beyond helping to ensure authenticity. Forensic tools also provide analytical capacity and introduce and stimulate concepts of physically and virtually extended context, stylometry, phylogenetic relatedness, mining of less structured information, visual analytics and chronological mapping, digital materiality and haptic emulation. These techniques may be useful for curators and scholars alike, but for different reasons.

### 5.1. Digital History in a Digital Milieu

Archival theory has long given fundamental prominence to fonds, the original arrangement of a personal archive's contents. Preserving fonds helps to maintain context and structure, enabling access and understanding. Circumstance and setting remain crucially important in the digital arena too. For a digital archive, context can be perceived in three layers (each with physical and digital (logical or virtual) manifestations):

- the microscopic scale of physical magnetic flux transitions, and logical binary and hexadecimal code subject to file system and bulk data analyses;
- the mesoscopic scale of the original computer environment, the graphical user interface, complete with the desktop layout, folder directories, application toolbars, and network volumes and resources, and in the selection of menu items through physical mouse, trackpad or touch screen; and

[13] http://www.bitcurator.net

- the macroscopic scale, beyond the immediate computer environment to the physical environment of local landscapes of home and study, of lab and studio, and (increasingly) to the remote physical environment experienced through sensors, and the sound and vision environment of virtual worlds.

The scholar or scientist in the pursuit of meaningful textual and graphical encodings needs to be able to move to and from one level to another, integrating the analogue and digital environments, and dynamically to infer and scrutinize possible reconstructions and sequences of events. It is also desirable to have a record of the exploration and examination for later reflection and for referencing. Emulation, virtualization and virtual reality will play pivotal roles in bringing these three layers of context together.

## 5.2. Virtual Archives

The conventional approach by digital archives and libraries has been to focus on the individual digital objects, making them available independently. Context may be catered for through records of the original arrangement of these objects. The careful investigation of individual digital (and analogue) objects in this way will remain a key focus of research; but, with modern technology, scholars should be able to experience the original arrangement directly, and not conceive it only from an arrangement record.

In the digital era, there is the possibility of making available an entire disk, even the entire personal digital archive, so as to allow the researcher to encounter, and search across, the original creative and operating environment of the originator. Virtual computing offers a route towards this functionality.

Forensic products such as EnCase have for some years provided for the possibility of mounting disk images through the use of integrated Physical Disk Emulator (PDE) and Virtual File System (VFS) modules. The possibility of not only mounting a virtual disk but also booting it using a virtual machine was discussed by JL John (2008).

The product VFC (Virtual Forensic Computing) from the company MD5 in the UK, based on research at Cranfield University (Penhallurick, 2005b) has been designed specifically for the purpose of 'experiencing' the original environment. Essentially the software examines the disk image and determines a suitable VMware virtual machine on which the original disk image can be booted (Penhallurick, 2005a; Penhallurick, 2005b). LiveView is a free product[14] with a similar objective (Bem and Huebner, 2007b). At the British Library, Mount Image Pro has been used to mount a bootable disk image derived from a personal computer of evolutionary biologist John Maynard Smith, with VFC configuring a VMware virtual machine on which it could be booted (using the resident operating system, in this case Microsoft Windows 98).

[14] http://liveview.sourceforge.net

It is possible to configure the virtual machine manually, not only with VMware but other virtualization products too. Common virtual desktops include: VMware Fusion, Microsoft Virtual PC, Windows Virtual PC, and Parallels, and open source VirtualBox and Xen (Bem and Huebner, 2007a; Shavers, 2008; Barrett and Kipper, 2010).These products are aimed at contemporary machines, specifically entailing Windows, OS X, and Linux.

For earlier machines, there are a number of emulators, some of which are deliberately designed to be preservation friendly. Important progress has been made by the Keeping Emulation Environments Portable (KEEP) project[15] with its Emulation Framework, emerging Universal Machine and database for Trustworthy Online Technical Environment Metadata (TOTEM).

The open source emulator SheepShaver is derived from the classic computer enthusiast community, and it can be configured to mount and boot a forensically sound disk image derived from an Apple Macintosh system that predates OS X. As reported at the Digital Lives Research Seminar 2010, SheepShaver has been adopted at the British Library in order to boot a disk image of one of the hard drives (G3 PowerMac with Macintosh System 8) from the evolutionary biologist WD Hamilton. Each time the computer disk is booted, one of several potential desktop pictures is revealed; for example, personally taken aerial photos of the river system in Amazonia. After a while, a screensaver appears. In addition to the usual applications such as Microsoft Word, Acrobat Reader and Photoshop 4.0, it has been possible to open CodeWarrior and run C++ programs residing on the original disk, displaying dynamic graphics.

The archival and technical team at Emory University has pioneered a similar arrangement for the digital archive of author Salman Rushdie, and as a consequence it is possible to browse the directory structure and open applications such as MacWrite Pro and ClarisWorks (Loftus, 2010; Carroll *et al*., 2011).

There is an interesting difference in emphasis between the approach to emulation currently practised in the preservation community and the forensic and curatorial approach outlined. In the preservation context, a typical scenario is to have an object of interest (e.g. a computer game application or a word processing document) and seek to play it or view it. To do this, a suitable emulator is found for the game or document. With virtual archival computing, it may well be unnecessary to find application emulators since the necessary application software already exists within the forensically captured disk. Nonetheless, both scenarios will ultimately depend on a preservation compliant lower level virtual machine.

## 5.3. Forensic Animation, Forensic Virtual Reality

Virtualization extends into virtual reality and 3D imagery, and offers the opportunity of integrating the presentation of analogue and digital environments through virtual and graphic technologies. Three-dimensional virtual reconstructions, augmented reality, simulations, animations and digital data displays are being conducted as a means of presenting forensic evidence. Critically, in a

[15] http://www.keep-project.eu

forensic context, there is significant effort to validate virtual environments for their evidential reliability through reference to the original data, while ensuring that the 'inherently persuasive nature' of visual presentation does not lead to 'undue reliance' on it. Conversely, the same techniques can be used to generate and explore competing hypotheses, 'exposing any inconsistencies' (Schofield, 2009). To ensure that the computer graphics presentations are appropriate, fair and authentic, the accuracy of data, methods and final image generation are subject to testing and verification by forensic researchers (Ma *et al*., 2010). If this functionality is achievable with some realistic measures of reliability it would have potential use in the scholarly context.

## 5.4. Drafts, Fuzzy Hashing and Phylogenetic Relatedness

Scholars often seek to identify versions of a creative production. A significant source of interim versions or drafts is to be found in the virtual snapshots that are made by modern operating systems through Volume Shadow Copy (Microsoft Windows) and Time Machine (Apple Macintosh).

Cryptographic hashes provide a digital fingerprint but are ill suited for identifying similar digital objects. Fuzzy hashes, or, more formally, context-triggered piecewise hashes (CTPH) produced with the program ssdeep (Kornblum, 2006) and similarity digests produced by the program sdhash (Roussev, 2010;  Roussev, 2011) supply measures of similarity for a pair of objects. An outline of CTPH is provided in H Baier and F Breitinger's examination of security implications (2011).

Beyond similarity there is the question of relatedness by descent, since this is a major means by which similarity arises; and the primary way to determine such relationships is through phylogenetics (see Glossary). Various methods of phylogenetics and associated techniques such as the application of Kolmogorov Complexity as manifested by Normalized Compression Distance (NCD) have been applied to digital objects. (NCD, for example, presumes that two objects are close if one can be compressed using information from the other.) Both scholarly and scientific in outlook, the techniques have shed light on chain letters, computer viruses and other malware, software evolution, digital images, music, and plagiarism (Bennett *et al*., 2003; Cilibrasi *et al*., 2004; Cilibrasi and Vitányi, 2005; Aquilina *et al*., 2008; Ji *et al*., 2008; Cebrián *et al*., 2009; Kraus, 2009; Dias *et al*., 2010). Kari Kraus draws out some of the intriguing parallels and influences between textual scholarship, historical linguistics and bioinformatics.

## 5.5. Stylometrics and Individuality

The identification of individuals through the style of their handwriting (palaeography) is an ancient skill; critical scholarship has long fostered an ability to interpret the style of content, a style in the choice of words and ornament, both literal and pictorial – based on a multi-evidential prospect. At times it might be possible to conclude that a set of writings is derived from a single individual, who nevertheless remains unidentified. Furthermore, even without revealing identity, it may be possible to surmise social and educational status, gender, age and even emotional state of the writer.

With the adoption of careful measurement and application of advanced statistics, the field of (scientific) stylometry has begun to become more effective. Patrick Juola provides a useful review of the field and definitions (Juola, 2006). Multivariate statistics, multiple metrics, natural language processing and data mining techniques have been applied to authorship attribution and characterization with electronic documents, emails and internet chat (Hadjidj *et al*., 2009).

Two other areas of forensic research and practice are of importance to digital scholarship, including, but not only, authorship attribution and the recovery of drafts. Firstly, there is electronic document analysis, or more generally the use of embedded data. The critical aspect of this well-established area of digital forensics is the careful use of multiple, mutually corroborative sources of evidential artefacts (not least date-time stamps), internal and external, as well as extraction of earlier drafts from within compound word processing files (Bryson and Stevens, 2002). Secondly, file carving (see Glossary) and fragment identification are pertinent in several ways including the automated reassembly of draft versions, and the authorship of fragments (Garfinkel, 2007; Pal *et al*., 2008; Axelsson, 2010). A recent study examined the problem of linking carved information to the individual when more than one person has used the computer storage media such as a disk drive (Garfinkel *et al*., 2010). It was argued that the automated solution developed 'is superior to the manual approach' because it (i) considered *all* of the available data on the computer's hard drive, and (ii) supplied an error rate, which means that it complied with the Daubert requirement for potential or known rate of error with empirical testing.

Besides digital scholarship per se, these areas of digital forensics are of specific relevance to long-term digital preservation and the technical understanding of digital objects.

## 5.6. Text and Multimedia Mining

While data mining is designed for formal databases, text mining and its multimedia counterparts (image mining, audio mining and video mining) are directed at extracting meaning from unstructured information through, among other things, analysis of multiple associations. Text mining alone is a huge field of research, which has attracted the participation of digital forensic researchers as a means of promoting evidence discovery and relation extraction. One study concluded that forensic investigation is frequently hampered by incomplete analysis of available digital information, and by the amount of time it takes to garner and annotate the necessary background information that allows the evidential base to be queried effectively through conventional search (Louis, 2009). Mining concepts and relations in combination with interactive exploration through a visual interface was found to be effective in identifying entities, events and associations of interest in unstructured textual content, and in stimulating systematic hypotheses.

Commercial software includes modules of SPSS and SAS; open source tools include RapidMiner and an extension of Weka and GATE which is written in Java. Text mining functionality is available from R by means of the tm package which interfaces with Weka and openNLP (a machine learning toolkit from the Apache Software Foundation) (Feinerer *et al*., 2008). In this context, it is worth highlighting NLTK (Natural Language Toolkit) from the Python community. In the arena of multimedia, facilities and prototypes for emotion recognition, human age estimation based on facial information, motion intensity, energy measures, colour histograms, dominant colour, colour content and orientation have been employed. The approach is centred around the extraction of pivotal features, identification of unusual patterns, and extraction of implicit information.

A report into mining multimedia observed: ‘An interesting research direction on web content mining is the integration of heterogeneous information sources’ (Kamde and Algur, 2011). This concurs with the emphasis placed in forensics on multiple sources of evidence. The strength of heterogeneous material including personal archives is that it offers a stronger multi-evidential foundation. Similarly, research into the automatic generation of captions or annotations for still images has embraced multifaceted associations: ‘The motivation for using multi-relational association rule mining for multimedia data mining is to exhibit the potential accorded by multiple descriptions for the same image (such as multiple people labeling the same image differently). Moreover, multi-relational association rule mining can also benefit the auto-annotation process by pruning the number of trivial associations that are generated if text and image features were combined in a single table through a join’ (Teredesai *et al*., 2005).

The discovery of objects in images has been explored through the use of a visual analogue of a word (Sivic *et al*., 2005). Object categories are treated as topics, and an image with several categories is modelled as a mixture of topics in the same way that text analysis discovers topics in a corpus using a ‘bag-of-words’ representation of a document.

Text mining algorithms have been adapted and optimized for the massive parallel data processing capacity of GPUs, specifically for use with the software CUDA (Compute Unified Device Architecture) (Zhang *et al*., 2009).

Also important in this context is the extensive research into the forensics of audio, video and multimedia including detection of forgery (Farid, 2008; Wang and Farid, 2006; Farid, 2009; Wang, 2009).

## 5.7. Chronological Mapping and Visual Analytics

Visual presentations of data and analysis such as diagrams, graphs and plots are invaluable in conveying and interpreting complex information. A continuum can be seen between visual data analytics, where very careful consideration of statistical principles is required, and visualization, where the emphasis may be more on exploration, on discovering the unexpected and on informal presentation. Well-known statistical packages with graphical components include SPSS and the GPL (Graphics Production Language). The open source R statistical package offers an inexpensive route to sophisticated statistical analysis with many additional modules created worldwide by scientists and others. Graphiz provides open source graphics capability.

The upsurge in interest in visual analysis and presentation arises from the vast amount of information that has become available and the desire to analyse it as quickly as possible, if not dynamically in realtime, because the information is needed promptly or because it is continuously changing. Commercial software adopted by financial institutions for fast analysis include

Panopticon and Tableau Software, providing a plethora of presentations from heatmaps and time series to candlestick graphs and theta analysis.[16] While these are fairly straightforward to use, due to the degree to which the software seeks to reduce effort, considerable care is required when setting up analyses for interpretation.

A similar requirement for dynamic fast visualization exists in cyber security and command and control. Indeed, system administrators were early adopters of visualization for monitoring their networks, and software tools such as the NVision-PA continue to be developed for this purpose.

Two longstanding tools for mapping and visualizing networks in the intelligence and forensic fields are i2 Analyst Notebook and Sentinel Visualizer. While these were originally designed for intelligence analysis, the software can be used for digital forensics as demonstrated with the i2 product in an exploratory analysis and visualization of malfeasant activity collected through a Voice over Internet Protocol (VoIP) honeypot (see Glossary) (Valli, 2010). Trace analysis combined with graphical visualization revealed that the emulated VoIP system was successfully compromised by attackers.

An attractive aspect of these tools is that they allow customization of icons, and in a limited way could incorporate metadata icons in a manner similar to that outlined by the Digital Lives research (John *et al*., 2010). The proprietary products are expensive and flexibility is modest; although Sentinel Visualizer is cheaper and more open, i2 has established itself firmly within the intelligence and investigation community (recently becoming a member of the IBM group).

Acknowledging the large volumes of information, some forensic scientists have proposed a need not only for a visualizing of the ultimate findings, but an explicit recording and graphical formalization of the process of detection and examination. It is suggested that a methodology for 'archiving, retrieving, and reasoning about forensic knowledge' would help to improve forensic skills, aid team succession, and allow the reuse of forensic knowledge (Bruschi *et al*., 2004). A graphic representation can help to structure the argumentation from evidence to hypothesis (Reed and Rowe, 2001; Van Gelder, 2002).

With the advent of social networking visualization there are many open source tools available from NodeXL (based on Excel spreadsheets) through to Gephi. Peter Chan and Michael Olson at Stanford have conducted some effective visualizations with data extracted forensically from personal archives with MUSE prototype software by Sudheendra Hangal (AIMS, 2012). At the British Library some graphs and network visualizations have been compiled using dates and times of email messages and of files generally from personal archives of the poet Wendy Cope and biologist Bill Hamilton for instance; for example daily routine has been plotted by using the time without the date of date-time stamps. The fascinating possibilities of personal analytics have been demonstrated by a blog entry by Stephen Wolfram (2012).

[16] http://www.panopticon.com/; http://www.tableausoftware.com/

A central concern is that of reliability of data extraction, analysis and interpretation. Will presentations stand the test of examination in a public forum, a legal environment? Although there are many visualization software packages available, few may be said to have been validated for forensic purposes. As might be expected, one of the most common requirements in the forensic literature is timeline visualization (Buchholz and Falk, 2005; Olsson and Boldt, 2009; Guđjónsson, 2010; Hallman, 2011; Marrington *et al*., 2011). Other facilities aim to interpret the organization of evidence (Vlastos and Patel, 2007).

## 5.8. Digital Materiality and Haptic Emulation

The concept of digital materiality (see Glossary) is still in flux, but evidently it embraces the architecture of the digital device as an environment where creativity manifests itself. It incorporates both the physical and the virtual, from keyboard and mouse to graphical user interface, toolbars, dropdown menus, and the individual's manner of using the virtual desktop, nesting folders and naming files. A number of textual and literary scholars have joined together in emphasizing the necessity of attending to these aspects just as other scholars undertake codicology and iconography in the investigation of the creation of meaning and of trails of provenance: bindings, illustrative layout and paper quality (Manoff, 2006; Kirschenbaum *et al*., 2009b; Redwine *et al*., 2010; Trace, 2011). Physically material aspects of computing are obviously a potential source of evidential traces useful to forensic procedures, highlighting the blending of real analogue and digital worlds. In time, haptic technology (pertaining to touch and tangibility) may come to emulate the physical computing experiences such as the feel of a unique plastic casing, the response of a keyboard, and the behaviour of the mouse and joystick.

# 6. Legal, Ethical, Historical: Imperatives and Dilemmas

So far this report has simply presented forensic approaches and their potential in technological terms. It should be obvious, however, that the application of these tools reflects new kinds of tensions for archivists in data protection and management. This section highlights issues of privacy and the potential role of digital forensics in helping to protect it. Information assurance and intellectual property are also briefly discussed. The combination of temporary anonymization with mediation by a trusted repository is outlined as a possible way to couple privacy protection with the legitimate requirements of scientific research in the near term, while securing the information for later biographical and historical research when the period of anonymization comes to an end. The related concept of the 'digital shadow' is also outlined.

## 6.1. Issues of Privacy and Intellectual Property

The use of forensic technology at a time when surveillance, mobile phone hacking and the use of personal profiles by social networking sites are of widespread concern, clearly calls for comment. There is the inherent power of forensic computing in its ability to reveal private aspects of a life, and there is simply the use of the term 'forensics' rather than, say, 'in depth analysis'. The present report concurs with the view of Kirschenbaum, Ovenden, Redwine and Donahue that originators and others may well prefer an openly principled approach to privacy issues and digital rights

(Kirschenbaum *et al*., 2010). All digital technology poses a threat to privacy wherever and whenever technical mechanisms and social policies are not in place to counter it.

A primary purpose of the archival application of forensics is in fact to help protect the privacy of originators and third parties. This has long been a role for archivists. Forensic search functionality makes it possible to identify quickly and efficiently credit card numbers, telephone numbers, email addresses, postal addresses and the like. The day may come when many aspects of privacy and intellectual property (legal, ethical and cultural compliance) can be identified automatically without the direct agency of a curator, but until this time some other approach is necessary. The tried and tested method is for a professional archivist or curator to examine the material prior to granting access. While originators can be encouraged to attend to their own privacy wishes prior to their archive's transfer to a repository, it is still necessary to address the interests of the third party. Quasi-automated analysis and extraction supervised by an archivist may be productive in this context.

Correspondingly, forensic techniques can ascertain whether privacy has already been breached, with information existing on the Internet for example. Similarly, it is not merely a matter of forensic revelation; simply being able to read obsolete media using technology no longer available to most people may impact privacy.

Forensic analysis may also benefit intellectual property and other digital rights.

## 6.2. Procedures, Protocols, Policies for Privacy and Other Rights

### 6.2.1. Meeting the Expectations of the Individual

Central to reassuring originators and others is to match procedures as far as permissible to the varying attitudes and beliefs of individuals. Conventional approaches such as the possibility of periods of reservation (embargo) and selective redaction can be applied to personal digital archives. Whatever the technical reliability, it ultimately depends on appropriate, effective and open policies and protocols, and astute curatorial decision-making to ensure a high degree of trust in institutional repositories (as has existed for some institutions over decades and centuries) to keep information private for some considerable time.

Broadly a series of steps may be recommended: (i) establish open policies and procedures; (ii) inform and seek consent of donors and families; (iii) preview content of a personal archive; (iv) discern as far as feasible the interests of third parties; and (v) take actions to comply with policies and expressed wishes. Some of the onus may be placed on professional researchers but the policy boundaries must be clear.

Approaches towards establishing informed consent and understanding warrant further study and clear guidelines. Some of these questions are being explored by an international Born Digital

Acquisitions Group[17] consisting of digital curators with complementary expertise in subject and technical aspects.

Technological options include logical acquisition of active files instead of physical acquisition that includes unallocated space where deleted files reside, or the editing of disk images with forensically facilitated redaction and selective encryption, and secure deletion where agreed (perhaps with the family accepting a duplicate so that the institution does not find itself deleting the last copy in existence). Tools for manipulating disk images and virtual machines are useful in several ways: VirtualBox is versatile and can be used for conversion from one VM file format to another; Winimage can be used to delete folders and the like; Microsoft's Disk2vhd enables migration of disks from physical to virtual (Barrett and Kipper, 2010). Of particular interest is the ability to securely but selectively delete within a virtual structure.

Virtualization is also useful in forensics for creating a sandbox environment in which to examine systems that have, or may have, been compromised by malware.

A significant provision within computer forensics is a system for controlling and auditing the capture, handling and analysis of digital evidence. Some software incorporates a degree of automated audit, recording the actions of examiners and controlling access according to their status or identity. The documentary information is held with the evidence, and is available as a report at the end of an investigation.

The extent to which these functions are incorporated in an archival context will depend on the institution concerned; however, it does suggest a way to help protect curators and the institutions which they represent from any misapprehensions and to help assure depositors of the seriousness with which the handling of personal information is taken.

Portable forensics, as represented by the eponymous product of Guidance Software, has two components: triage and data collection. 'Triage' is aimed at reviewing quickly in the field, in real time, information residing on a computer without altering the information. Searches may be precisely configured in the lab beforehand or may be quickly customized by advanced users while in the field. 'Collect' makes it possible for the actual collection of information on media or in memory to be preconfigured and targeted by the specialist according to agreed criteria, allowing others to undertake the capture in a forensically sound way; collection might be restricted to emails for example.

Other approaches towards controlling access to private information beyond initial metadata extraction might be helpful: (i) the precise use by curators of the most intense forensic and data fusion techniques might be sensitively controlled; (ii) careful mediation of permitted access, with potential restrictions on *dynamic* analysis, perhaps in some cases with no or limited further analysis *across* archives being permitted.

[17] The group is coordinated by Gabriela Redwine, University of Texas at Austin

On the other hand, security can be enhanced through forensically enabled facilities such as watermarking, digital signatures, selectively encrypted metadata and redacted content, and forensically tested procedures and technologies such as monitoring and evaluating the security of virtual networks.

### 6.2.2. Information Assurance and the Internet

The establishment of a secure repository system is fundamental, and in its own right can reassure originators, family and others concerned about information assurance and security of personal information. The ISO 27000 series provides for the analysis of risk (Higgins, 2010; Ball and Billenness, 2012). Two specific concerns are cyber security and the Internet (Sommer and Brown, 2011), and the origins of hardware as well as software (Pfleeger, 1997). In the UK guidance for information security is provided by Communications-Electronics Security Group. Such advice in information assurance includes assessments of technical risk and of technology suppliers concerning the securing of information especially personal data (NTAIA, 2009; NTAIA, 2011).

The juxtaposition of personal archives and national security has led to governments, their agencies and other organizations raising the profile of security with National Cyber Security Awareness Month (NCSAM) and Data Privacy Day (DPD) and providing guidelines for cleaning your machine at home.[18] A specific concern has been the proliferation of botnets (see Glossary) (ENISA, 2011).

New approaches to security include schemes for anonymous attestation of hardware devices with enhanced privacy identification (Brickell and Li, 2009).

## 6.3. Increasing Historical Information (and Social Research)

The other risk is that digital objects and content are not captured in the first place. History is first and foremost about people and events in their lives. The numbers of people are increasing, and people are living longer. Vastly more personal and family information is being created than ever before. The capture of this information will depend on the policies and remits of cultural organizations, which in turn depend on the practicalities of what is feasible.

Catering for private information and personal identity in a nuanced way is among the most costly aspects of holding and making available personal archives. It may be more cost effective to capture and hold significant quantities of information but only release it for access conservatively. At the same time, publicly funded repositories are understandably reluctant to store large volumes of information without making it available for research. Is there some other way that forensics could facilitate academic research with personal data?

One option might be to temporarily anonymize information. Such information may be of immediate interest in social science and cultural studies, even if it is not yet available in a form

[18] eg http://stopthinkconnect.org/campaigns/keep-a-clean-machine/

suitable for biography. This immediate access for social analysis would justify the investment by funding and governmental bodies on the acquisition of high volumes of life information, while securing quantities of personally identifiable information useful for future release to historians and biographers. This high volume throughput does not preclude the possibility of continuing with a bespoke curation of the smaller volume of prominent and identifiable personal archives for quick access.

There are notable limitations to the actual processes of de-identification and anonymization in practice (e.g. Neamatullah *et al*., 2008; Ohm, 2010). A key point is that digital objects in an archival repository are not necessarily released into the public domain. A subset of objects may be made available online in some way, but many others will not be, and will remain within the confines or control of the repository. This means that the nature of any analysis may be mediated by the repository. Anonymization (temporary or otherwise) may not have to be completely unbreakable, but be secure using the techniques made available to the researcher.

It is even conceivable that analysis is conducted by curators on behalf of researchers. Forensically extracting data in a carefully systematic way would help to ensure that the information that is *extracted* for release is anonymized reliably – at least as far as current technology is concerned. Practical grades of security for classes of anonymization according to the context of current technology might be useful.

Digital forensic analysis itself will help archivists and computer forensic practitioners alike to better understand and research the limitations of anonymization of less structured information. Concerns about the feasibility of long-term anonymization apply to personal information outside any repository, of course, and belong to a broader social issue for digital life (e.g. Mayer-Schönberger, 2009).

The forensic community has already recognized the challenges ahead in terms of capturing and processing large volumes of personal digital information. A combination of forensics, early intervention and a focus on scale could make a substantial difference, imparting some urgency to the exploration and testing of the practicalities of automated forensics.

## 6.4. Digital Detritus, and the Forensics of the Digital Shadow

Individuals are not only creating personal information directly, but are stimulating its production indirectly, as a personal digital shadow (see Glossary).

A high proportion of personal information is currently generated and held by organizations as people go about their daily lives, and is reflected in bank account statements, eBay reputation metrics, Amazon book buying habits, CCTV, RFID, London Transport Oyster Card records, and so on. As the digital shadow and the sharing and mining of personal information expands, there is a concern that security is not keeping up (Gantz and Reinsel, 2011).

What are the implications for authenticity, for provenance and for historical integrity? If remnants of a person's digital shadow emerge in the future, how will it be interpreted? It may well be historically invaluable, but how will it be authenticated and corroborated? How will identity be established (Walden, 2004)? An inkling of how it might be done is by examining an individual's own computers and mobile devices, their immediate personal belongings and comparing this with information in the cloud, which in turn raises privacy issues (Garfinkel and Cox, 2009). An open source tool has been developed for reconstructing a user's online activity by extracting data from their hard drive: Offline Windows Analysis and Data Extraction (OWADE) (Bursztein *et al*., 2011). Its principal author is currently employed by Google to research internet security and privacy.[19]

In short, through an understanding of the contents of a person's desktop and laptop and their immediate digital belongings, it may be possible to characterize their online life, their participation in gaming communities, social networks and the like. An interesting challenge is how one undertakes this process within the wishes of a depositor when there is so much uncertainty about exactly what is retrievable from the cloud. Creators may seek to obtain and practically manage their online profile as a routine aspect of digital life.

## 7. Archival and Forensic Perspectives: Future Vision

This section considers future prospects. Although significant steps have already been taken by the curatorial community in adopting and adapting digital forensics for archival purposes, more progress remains to be done; scholarly referencing is an area of special interest. Other prospects for change are shared with the forensic community at large: scale and the necessity of automation; virtual entities to be investigated and to be used forensically; challenges of cloud computing and networks; and the advancement of digital forensic science.

### 7.1. Advancing Digital Forensics within an Archival Context

Despite the overlapping and parallel perspectives that typify forensics and other approaches to the study of the past, the first decades of digital forensics have been mostly directed towards the investigation of wrongdoing. Until recently, it has not been directed at serving the archival community.

It is necessary to further adapt: (i) the forensic technologies and methodologies for archival purposes and perspectives, as well as (ii) the practices of the archival community to absorb digital forensics.

Suitable terminologies and scholarly referencing systems are required for effective integration of archival and investigatory practices. Bates numbering (see Glossary) is common in forensics and the

[19] http://elie.im/

legal context generally, and occurs in some digital forensic tools.[20] T Larson (2002) outlines the use of Bates numbers in the context of electronic discovery or eDiscovery.

Scholarly referencing of complex digital objects calls, however, for more sophisticated solutions. A version or extension of Ted Nelson's strong hypertext may be applicable (Nelson, 1995; Lowe and Hall, 1999). The concept of 'transclusion', for example, is characterized as 'reuse with original context available', with documents having windows to other documents, and with intellectual property rights being managed on a microscale (see Lowe and Hall, 1999).

Education has been a longstanding concern of the computer forensics and law enforcement community (e.g. Yasinac *et al*., 2003). It would be beneficial for training to be tailored for the purposes of archivists and information scientists. Pioneering work in this area has been done at the University of North Carolina at Chapel Hill and at the University of Glasgow. The Rare Book School held at the University of Virginia has incorporated forensics in their classes. Stanford University has created a video outlining the workflow with FTK. Gareth Knight organized a forensic imaging session for archivists during the FIDO project. The Digital Preservation Coalition, the Archives and Records Association UK & Ireland, and others have a range of training activities that build capacity in the workforce. Readers may wish to consult the DPC's *What's New* bulletin to find out about forthcoming opportunities.

## 7.2. Scale and Automation

As already intimated, scalability of volume and processing is a shared concern for digital forensics and digital curation and preservation. The need for fast processing to cope with the rapidly expanding volumes of personal digital belongings and information generally has been discussed repeatedly (Brueckner *et al*., 2008; Grillo *et al*., 2009; Sommer, 2009; Garfinkel, 2010;) , and is being actively addressed by the BitCurator project.

Efficient, prioritized triage would allow a curator to quickly appraise a personal archive for possible acquisition; the ability to resolve identity and to profile an individual's social networks are obviously useful to the archivist as is the ability to process, transfer and ingest large volumes of information using automation and parallel computing. One study addresses bulk data analysis in the automation of forensic analysis (Garfinkel, 2011). File fragment type identification was originally aimed at file carving and analysis of memory; however, it can also serve to determine efficiently a statistical analysis of the content (in terms of file type) of a hard drive through random sampling (Cohen and Schatz, 2010; Garfinkel *et al*., 2010).

Many of these pioneering efforts are directed at efficient partial extraction and rigorous randomized statistical sampling, without necessarily relying on high processing power. However, other forensic activities continue to benefit from high performance computing. Not least among these procedures – along with decryption and password recovery – are text and multimedia mining and network visualization, all remaining extremely computer-intensive. Like other scientists,

[20] eg ProDiscover DFT, http://www.techpathways.com

forensic and text mining researchers have begun to adopt GPU processing using open source software such as CUDA and OpenCL (Open Computing Language)[21] (Marziale *et al*., 2007). A balance has to be found between speed and comprehensiveness.

Equally, with automation and increased usability it is even more imperative that the procedures are demonstrably rigorous, forensically reliable, and certifiably accurate (even if not sensitive in detail). Careful documentation of the actions and capabilities of automated tools will be essential.

## 7.3. Personal Virtualization, Cloud Computing

Noting the proliferation of virtual environments in the data centre and on desktop, laptop and removable media, a review of virtualization anticipates (i) not only its use as a forensic tool, but (ii) an increasing need to investigate virtual environments (Barrett and Kipper, 2010).

### 7.3.1. Personal Use of Virtual Machines, Disks and Appliances

Of immediate interest to curatorial forensics is the existence on personal computers of diverse virtual machines, virtual disks and virtual appliances that may need to be identified, captured and analysed. Similarly, the use of emulators such as Bochs and DOSBox by individuals may need to be accounted for.

An initial question is the quality of the forensic capture and examination of virtual machines, virtual disks and emulators. Analysis of virtual entities is an emerging area of digital forensics, and there remains considerable potential for developing accepted tools and approaches for verifying and handling virtual machines in forensically sound ways (Bates, 2009).

### 7.3.2. Forensic and Preservation Use of Virtualization

One consideration is the use of emulators in digital preservation. When objects are viewed, rendered, using an emulator, how accurate and sensitive is the emulation of the hardware and software being emulated?

A virtual machine may be configured to behave like the original machine from which a collection hard drive has been removed (as outlined in § 5.2). As the user explores the digital objects within this emulating or virtual environment, changes take place, and in many cases can be seen. Usually with a modern virtual machine the original disk image remains unchanged, with the changes (such as when a user moves or views a file) represented in temporary files associated with the virtual machine. (This is akin to the way a *dynamic* hardware writeblocker such as Voom's Shadow Drive can be used to prevent changes to a physical hard disk that is subject to examination, and write protect a physical disk while allowing dynamic interaction: the changes are cached in temporary

[21] http://www.khronos.org/opencl/

files leaving the hard drive unchanged.) There is the very useful option of occasionally restoring an earlier or even original state using 'snapshots' of the virtual machine.

Even with the source disk image protected from change when booted and actively explored, the curator (or scholar) would still *perceive* changes to files. It would be of further benefit, therefore, if this were not so, or (more realistically) if perceived changes are managed and those that do occur are identified and indicated to the curator. Thus the curator should be able to browse an opened word-processed document in an environment matching the original environment but without being potentially misled by ostensible changes to it. Paraben's Email Examiner, for example, allows the researcher to browse email messages without modifying them; e.g. if an email message is 'unread' (and this state is shown with 'bold' style) it remains 'unread' even after the researcher has in fact read it. Email Examiner, however, does not show the emails in the original environment (e.g. Microsoft Outlook).

The use of virtual machines and bootable disk images in forensic presentation might be advanced further if the perception of change could be minimized beyond essential navigation, and the opening and closing of digital objects could be conducted in 'read only' fashion regardless of the original software; ideally, such functionality would be of forensic standard, tested and certified according to its limitations and capabilities.

### 7.3.3. Cloud Computing and Virtual Worlds

As with virtualization, a cloud computing environment can be both (i) a subject of forensic investigation, and (ii) a tool for conducting one (Lillard *et al*., 2010, Grispos *et al*., 2011; Birk, 2011).

There are limitations in gaining access through service providers to the information of a specific individual even with the permission of the originator, since data from multiple customers may be co-located and dynamically distributed over changing hosts and data centres (Barrett and Kipper, 2010). Although it is still early days, these challenges speak for a proactive approach tailored for individual writers, scientists and other creators, with regular downloading of personal information to a local site.

Specifically, with regard to virtual worlds and social networking sites, how can information from a virtual world be captured in a forensically sound manner? Techniques are being researched for gathering 'online evidence' from social networking services in ways that are more efficient and effective than web crawling, even when the service provider is uncooperative. Such techniques will require a strong legal remit (Huber *et al*., 2011).

A further approach is for a memory institution to function to some extent as a cloud computing provider itself, offering specialist curatorial services. Some of the activities such as preview and acquisition that curators and archivists undertake could be conducted remotely.

## 7.4. Forensic Science: Corpora and Hash Libraries

An important resource that is required for research and testing as well as training is the establishment of forensically appropriate corpora. Corpora may be of computing devices (e.g. mobiles), media items (e.g. hard drives and floppy disks), digital objects (of diverse file formats), and digital content (e.g. text for linguistic analysis, sound for speech recognition). The value of corpora for forensic and preservation testing and analysis is attested by a number of studies (Garfinkel *et al*., 2009; Axelsson, 2010; Aitken *et al*., 2008).

In some situations natural corpora may be unavailable or unfeasible. Sets of digital content may be generated artificially, using a genetic engine, as in a study by Cebrián *et al*., (2009) of plagiarism detection tools that adopted 'grammar evolution' for producing originals and artificial plagiarisms that mimic genuinely plagiaristic methods.

Similarly, cryptographic and fuzzy hash libraries need to be collected for some specific archival contexts. For example, a hash library that caters more fully for Korean software has been developed locally (Kim *et al*., 2009). A degree of overlap with existing registries is acceptable given the value of redundancy, but nonetheless the aim would be to focus on special aspects. The personal archives themselves will contribute to the hash libraries, and there will be a call for collaboration among repositories regionally, nationally and internationally.

In addition to the public hash libraries compiled as the National Software Reference Library (NSRL) and HashKeeper, there is a comprehensive commercial service known as Bit9.

It would be useful to research the integration of the forensic use of reference hash values with the preservation use of file characterization techniques.

The curation and preservation of media images, specifically disk images, and the design of repository architectures that support their sustainable use warrants much more attention: the Advanced Forensic Format and Digital Forensics XML provide a potentially effective means of organizing and managing disk images (Woods *et al*., 2011).

An example of the kind of potential efficiencies that could be explored is the transfer of a forensic disk image over a network while omitting common files such as those found in the operating system, with the disk image being reconstituted at the other end from a corpus that includes those same common files identified through their hash values (Watkins *et al.*, 2009; see Cohen and Schatz, 2010).

## 7.5. Anti-forensics

Measures for promoting deception and even for pre-empting or countering forensic analysis have long existed and continue to emerge in the digital era. In archival forensics it is to be hoped that such activities are less likely than in the context of law enforcement and cyber warfare.

Nevertheless, forgery and obfuscation, misinformation, and plain mischief remain distinct possibilities.

P. Juola (2006) stresses the necessity of sustaining advanced forensic research, by alluding to a never-ending arms race, as an expression of the ongoing relationship between forensics and anti-forensics. Thus as with digital preservation generally, it is necessary for digital challenges to be met by dynamic and proactive action and policy.

## 8. Conclusions

As should be now obvious, digital forensics is a broad field which intersects with preservation in a number of ways. It is hard to provide a complete summary, and conclusions are at best provisional given the rapid development in both fields. Instead of a simple conclusion, this report closes with a series of points as a way to help archivists and other readers understand and exploit the potential of forensics.

**Emerging Opportunities:** The landscape of personal and cultural information is changing rapidly, and new approaches to archives are necessary in order to remain effective and to exploit new opportunities.

**Personal Digital Archives as Very Complex Digital Objects:** The multiplicity of personal digital objects and the intricacy of their structural relationships pose major challenges to curation. By the same token, this collective complexity means that archives provide an invaluable window to the entire digital universe. Almost anything may appear in the personal digital archive of the poet, astronomer, mathematician, or political reformer: from emails and draft text, through datasets and technical workings, to jotted notes and travel films. Procedures explored and elaborated in this context may be transferable to other areas of digital scholarship and preservation.

**Inertia:** There has been a degree of inertia, if not stasis, in the handling and acquisition of personal archives among cultural repositories and archival institutions. For more than two decades concerns have been expressed about the curation of personal digital archives. In some cases repositories have simply let computer media sit on shelves quietly degrading; in others there has been an unspoken, if not official, policy of not accepting digital media. This appears to be due to an understandable trepidation about technological obscurity and transience, data protection, confidentiality and digital rights compliance.

**Evidential Value:** Along with the general concerns of digital preservation, notably the fragility of file formats and the absence of sustained interoperability, the erosion of the evidential integrity of historical digital objects poses another serious threat to long term cultural heritage.

Beyond essential provenance, the authenticity of embedded and associated metadata, notably date and time, is absolutely critical to historical scholarship and scientific analysis of personal digital objects and other informal, less structured resources, such as websites.

The Digital Records Forensics Project at the University of British Columbia is playing a key role in further strengthening the quality and rigour of forensic chains of custody and standards of evidence.

**Privacy, Security and Information Assurance:** It is crucial that repositories maintain the ability to protect privacy and digital rights and ethics with rigorous security. Curators have always been in a privileged position due to the necessity for institutions to appraise material that is potentially being accepted for long-term preservation and access; and this continues with the essential and judicious use of forensic technologies. The fields of information assurance and security usability are closely allied to digital forensics and serve as another source of tools and perspectives.

**Context and Integration:** Forensics and the curatorship and scholarship of personal archives both hinge on an understanding of the value of context. In forensics this is known as the multi-evidential perspective, where a number of diverse extant traces are examined and interpreted in order to retrospectively infer an ancestral state or event. Context may be captured for study at one of three general levels: (i) microscopic (hexadecimal code, magnetic flux transitions); (ii) mesoscopic (human-computer interface with mouse, trackpad, menu, toolbar); and (iii) macroscopic (physical landscape, studio, garden, virtual island in virtual world).

Historical evidence may reside in any one of these general levels, and may be digital or analogue (which may be captured digitally). The original personal digital archive may be enhanced through complementary information such as the virtual panoramas of the creative environment of the writer, video capture of a sculptor at work, audio interviews of political figures, and live data capture of a computer game or scientific experiment in play along with real world video of participants. All of this information may be fed into the forensic process, and subject to information fusion (see Glossary), text and multimedia mining, phylogenetic analysis of similarity and relatedness, and network and multinodal visualization.

**Virtual Archives through Haptics:** Virtual archival computing – forensically valid – makes it possible to experience the original look and feel of digital objects within original digital environments. It may in time be extended beyond the emulation of sound and vision to the emulation of digital materiality through haptic technology. Correspondingly, a field of haptic, sensor and 3D (digital material) forensics can be anticipated.

**Recovery:** The ability to recover passwords and deleted or encrypted information with the cooperation of the depositor is potentially very useful when done with due process, as is the ability to recover the files on obsolete media which the originator and family are no longer able to read. Increasingly, digital objects are encrypted, and the scientist might be anxious to make a dataset file available, but the password has been lost. A forensic approach combined with suitable institutional policies, e.g. controlling analysis and access, offers a way out of these quandaries.

**Forensics of Ancestral Computing:** Even after much of the digital forensic community has moved on with the investigation of modern technologies, there will be an abiding scholarly interest in earlier computers and code, inspiring a subfield of ancestral digital forensics founded on the retention of existing forensic science and scholarship.

**Open Source and the Ecosystem of Forensic Tools:** Although memory institutions will need to cut their cloth according to their resources, it is possible for even the smallest institutions to undertake the core requirements with a modest set of forensic tools, and certainly to do a great deal more than at present.

With the blossoming of open source forensic tools and an increasingly vibrant research community in digital forensics, including the new field of curatorial and archival digital forensics, it can be expected that the benefit-to-cost relationship will rise much further. As well as being transparent, open source practices serve to instigate innovation and encourage commercial entities to cater for the customer through early adoption of emerging capabilities.

The BitCurator project based at the University of North Carolina at Chapel Hill and the University of Maryland is playing a prominent role in fostering an archival forensics community.

**Personal Information as an Incomparable Resource:** With more and more people leading digital lives, the size and extent of personal digital holdings are growing. Away from the archival community, much of the emphasis is on personal information being a contemporary resource for marketing, business and services; but its value will potentially increase much further in future, not only in the realms of history and biography but also in social, medical and natural sciences. The archival community has an essential role to play in this process ensuring the evidential quality, preservation and utility of this unstructured information for scientific as well as scholarly purposes.

**Personalization, Identity and the Digital Shadow:** Increasingly personal information is being used to tailor the functions and services provided to individuals. As this practice evolves issues of ownership and use will intensify. Personal information, its curation, its integrity and its personal reuse may become crucial to the daily life of the individual.

Some personal information does not exist as conventional digital holdings but is created indirectly by individuals as a digital shadow, while personal identity likewise may be unclear without careful forensic analysis. Stylometric characterization, if not authorship attribution, of digital content is a potentially significant though still emerging subfield of forensics.

**Anonymization, with Quasi-Automation and Supervision:** Personal archives could simply be held for future in-depth processing but public institutions are reluctant to hold substantial quantities of objects without making them available promptly. Temporary anonymization of personal information could provide a solution. An understanding of forensics emphasizes the nuanced caution required in quasi-automated anonymization processes. Forensic research and testing will help to develop and appraise supervised redaction and anonymization techniques.

**Curatorial Extraction and Interpretation:** Where reliable automation of data anonymization is not feasible, selective analysis for, say, social science may be conducted by specialist curators who subsequently release only high level findings and statistical data that are inherently anonymous. As trustworthy institutional mediators, curators may research the content of personal archives for social and policy purposes, and provide safely and securely anonymized findings and interpretation.

**Digital Forensics is Integral to Digital Preservation:** In jointly advancing the curation and long-term preservation of digital objects, there are potential benefits to be enjoyed by both digital communities. Digital forensics can play a major role in sustaining the value of digital objects and content, and needs to be seen as an integral part of digital preservation and curation. Conversely, digital preservation and archival practice can further the aims of digital forensics through effective archival curation of evidence and contextual information. An area of special interest is the long-term preservation and curation of disk images and other media images.

**Digital Forensics is for Life:** Besides offering a set of essential tools, digital forensics provides an invaluable strategic paradigm, and a productive source of knowledge from a research community with overlapping interests. Ongoing and highly active research into digital forensics is inevitable, due to the need to cater for the appearance of new digital technologies, and to counter continuing anti-forensic activities.

## 9. Recommended Actions

### 9.1. Strategic Activities for Information Schools and Professional Bodies

1. Increase and maintain awareness of the existence and evolution of forensic tools through examples of usage by small and large archival institutions.

2. Put in place mechanisms for initial and ongoing training in forensic theory and practice.

### 9.2. Customary Practice for Memory Institutions

3. Establish clear open policies and guidelines for originators and family and others specifically regarding privacy and evidence protection.

4. Follow recognized forensic principles such as (i) not relying on any single digital technology, and (ii) corroborating and consolidating findings through a multi-evidential approach.

## 9.3. Ongoing Research and Development: Personal Informatics

5. Scope the feasibility of maintaining distinct lines between the processing of overt, active information and the recovery of ostensibly deleted, hidden information.

6. Ascertain means and ways for gauging the value of forensically authenticated digital objects and the reliability of investigatory tools in digital scholarship.

7. Explore and test the use of automation including digital acquisition and hashing, bulk extraction, indexing, text and multimedia mining, visual analysis and information fusion, and anonymization.

## 10.Glossary

| | |
|---|---|
| botnet | Currently one of the most potent means of breaching cybersecurity measures, a botnet is a set of computers which are connected over the Internet and which have been usurped by malicious software. The botnet may consist of many thousands of compromised and compliant devices, and may be exploited for a variety of inappropriate, criminal or counterstate activities. |
| chain of custody | A key concept in forensics whereby the custody and provenance of digital hardware, media and files are safeguarded through, for example, the appointment of evidence custodians. The purpose of the Digital Evidence Bag (DEB) is to hold digitally, along with the evidential digital objects, provenance metadata that can be updated as required: a concept that is familiar to digital preservation practitioners. |
| cryptographic hash value | The outcome of a cryptographic algorithm (such as MD5 and SHA1). The hash value may serve as a ‘digital fingerprint’. In principle, a change to one bit in the object being cryptographically hashed (e.g. a file or disk) will yield a different hash value. |
| digital conservation | The retrieval of information that resides on degraded media, the recovery of damaged digital objects, and the care of this fragmented and disrupted information. Preventive digital conservation may entail low-level examination at the interface where digital information is represented by analogue phenomena such as magnetic flux |

| | |
|---|---|
| | transitions. Digital archaeology looks beyond digital recovery to social interpretation. |

| | |
|---|---|
| digital materiality | Highlights the importance of context in digital information; specifically, it directs attention towards the way digital technology influences and constrains the creative process. Examples are writing technologies such as the dropdown menu and toolbar of word processing and the mouse and keyboard. It is anticipated that in time haptic emulation may help to make experiences of physical components available to future scholars. Similarly, haptic forensics can be expected to emerge as a means of authentication and reconstruction in the context of tangibility. |

| | |
|---|---|
| digital shadow | Refers to the personal information that is created indirectly and retained (typically by organizations) as individuals go about their daily digital lives. |

| | |
|---|---|
| diplomatics | Originally concerned with the production and transfer of genuine documents by official institutions – charters, diplomas and so on. (Correspondingly, the role of a diplomat has long been to ensure the secure communication of authentic messages from one government to another.) While palaeography focuses on handwriting, diplomatics addresses broader aspects of documents and records including textual and historical criticism. In the UK, the field may be referred to simply as ‘diplomatic’. |

| | |
|---|---|
| file carving | Refers to the process of extracting files and remnants of the files on a disk independently of the file system. Bulk extraction may be seen as a more general concept that is not restricted to the complete or partial restoration of files per se, but aims to obtain quickly and efficiently useful content and features. |

| | |
|---|---|
| forensic image or media image | The bitstream representation of a digital media object such as a hard drive, sector-by-sector. In principle it is a single file that represents the entire disk, although for convenience this complete file may be split into a series of segment files. |

| | |
|---|---|
| honeypot | A digital contrivance such as an ostensible computer that appears to be on a network and contains enticing information of apparent value, but is – in fact – a trap which is both isolated and able to monitor and record unauthorized attempts to gain access to the system. The honeypot is one of a number of active and passive techniques for measuring, detecting and countering security threats. |

| | |
|---|---|
| information assurance | An aspect of digital security, specifically directed at ensuring that the quality of the information is demonstrably safeguarded, that it has not been tampered with or accessed inappropriately. |

| | |
|---|---|
| information fusion | Refers to the process of bringing together information from disparate sources in novel ways to yield unanticipated insights. Whereas digital forensics is primarily concerned with the reconstruction of objects that existed and events that happened in the past, information fusion may bring elements of information together in entirely new ways. |

| | |
|---|---|
| order of volatility | Reflects the necessity of capturing the more volatile information, such as memory, prior to capturing less volatile information that is stored on media. |

| | |
|---|---|
| personal informatics | Concerned with the study of all aspects of personal information including such topics as privacy protection, the reuse of personal digital archives, personalized usability and personal information management. |

| | |
|---|---|
| phylogenetics | A biological discipline that aims to discern and understand patterns of descent and origin. The quintessential instance is the ‘tree of life’, the map of evolutionary relationships among living organisms, some being more closely related than others. Originally based on morphological characteristics, the field is increasingly founded on genetic and genomic data. The phenomenon of ‘descent with modification’ (or, loosely, imperfect replication with error or change) is not confined to life, and consequently phylogenetic techniques have been applied in other contexts, notably with documents such as literary, classical or ecclesiastical manuscripts (stemmatics), artefacts and tools (comparative anthropology), and language (historical linguistics). |

| | |
|---|---|
| sandbox containment | A secure computing environment for running novel, unattested or experimental code or changes in code, including potentially malicious code. The environment is self-contained with tightly controlled resources and is characteristically virtual. |

| | |
|---|---|
| scholarly referencing, Bates numbering | An essential aspect of digital archiving is the establishment of reliable systems for referencing digital objects and elements of their content. Hypertext may play a significant role. Bates numbering and stamping originated in the 19th century with Edward G. Bates, as an automatic system for numbering many documents consecutively, often for legal purposes. The approach has been adopted in the form of 'digital stamping' in Electronic Discovery, a branch of forensics most concerned with the legal disclosure of corporate information. |

| | |
|---|---|
| similarity digests, fuzzy hash values | Algorithmic representations that provide an indication of similarity among digital objects. By this means, files that share much content can be identified. |

| | |
|---|---|
| text and multimedia mining | Seeks to extract meaningful information from unstructured freestyle digital sources through automated or supervised procedures. |

| | |
|---|---|
| writeblockers | Tools that prevent an examination computer system from writing or altering a collection or subject hard drive or other digital media object. Hardware writeblockers are generally regarded as more reliable than software writeblockers. |

There are additional and very extensive glossaries on the websites of the Digital Records Forensics Project, the InterPARES projects (especially InterPares2), and the Digital Curation Centre.

## 11.Further Reading

### 11.1. Texts and Guidelines

All hypertext links checked on 5 October 2012.

ACPO, undated. *Good practice guide for computer-based electronic evidence*. Official release version, supported by 7safe Information Security, http://www.acpo.police.uk/policies.asp.

AIMS Work Group, 2012. AIMS Born-– ‐ digital collections: an inter-– ‐ institutional model for stewardship, http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_A4.pdf. (Contributors: Nicole Bouché, Judy Burg, Peter Chan, Bradley Daigle, Glynn Edwards, Michael Forstrom, Kevin Glick, Gretchen Gueguen, Tom Laudeman, Mark Matienzo, Michael Olson, Simon Wilson.)

Aquilina, J.M., Casey, E. and Malin, C.H., 2008. *Malware Forensics. Investigating and analyzing malicious code*. Burlington, MA, Syngress Publishing.

Beagrie, N., 2005. Plenty of room at the bottom? Personal digital libraries and collection. *D-Lib Magazine* 11(6).

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

Carroll, L., Farr, E., Hornsby, P. and Ranker, B., 2011. A comprehensive approach to born-digital archives. *Archivaria* 72, 61–92.

Casey, E. (ed.), 2010. *Handbook of Digital Forensics and Investigation.* London: Elsevier Academic Press.

Duranti, L., 2009. From digital diplomatics to digital records forensics. *Archivaria* 68: 39–66.

Farmer, D. and Venema, W., 2005. *Forensic Discovery.* Upper Saddle River, NJ, Addison-Wesley.

Fraser, J. and Williams, R. (eds), 2009. *The Handbook of Forensic Science*. Cullompton: Willan.

NIJ, 2004. *Forensic examination of digital evidence: a guide for law enforcement. NIJ Special Report*. Washington DC: National Institute of Justice, US Department of Justice.

Forstram, M., 2009. Managing electronic records in manuscript collections: a case study from the Beinecke Rare Book and Manuscript Library, *American Archivist* 72: 460–477.

John, J.L., 2008. Adapting existing technologies for digitally archiving personal lives. Digital forensics, ancestral computing, and evolutionary perspectives and tools. *The Fifth International Conference on Preservation of Digital Objects (iPRES 2008)*, British Library, London, http://www.bl.uk/ipres2008/presentations_day1/09_John.pdf

John, J.L., Rowlands, I., Williams, P. and Dean, K., 2010. *Digital Lives. Personal digital archives for the 21st century. An initial synthesis*. A project funded by the Arts and Humanities Research Council, http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf

Kirschenbaum, M.G., Ovenden, R., Redwine, G. and Donahue, R., 2010. *Digital forensics and born-digital content in cultural heritage collections*, Washington, DC, Council on Library and Information Resources, http://www.clir.org/pubs/abstract/reports/pub149.

Kirschenbaum, M.G., 2008. *Mechanisms. New media and the forensic imagination*. Cambridge, MA, The MIT Press.

Lee, C.A. (ed.), 2011. *I, Digital. Personal collections in the digital era*, Chicago, Society of American Archivists.

Olson, M 2010. Computer forensics in the archive: an analysis of software tools for born digital collections. Digital Humanities 2010 (DH 2010). King's College, London.

PARADIGM, 2007. *Personal archives accessible in digital media project. Workbook on digital private papers*, http://www. paradigm.ac.uk/workbook/. Principal authors: Susan Thomas, Renhart Gittens, Janette Martin and Fran Baker.

Politt, M. and Shenoi, S. (eds), 2005. *Advances in digital forensics: IFIP International Conference on Digital Forensics.* Orlando, Florida: National Center for Forensic Science.

Sammes, T. and Jenkinson, B., 2007. *Forensic Computing. A practitioner's guide.* London, Springer-Verlag.

Sommer, P., 1998. Digital footprints: assessing computer evidence. *Criminal Law Review Special Edition*, 61–78.

Thomas, S., 2011. Curating the I, Digital: experiences at the Bodleian Library. In C.A. Lee (ed.) *I, Digital, personal collections in the digital era*. Chicago: Society of American Archivists.

## 11.2. Academic Journals

*Digital Investigation*

*IEEE Transactions on Information Forensics and Security*

*International Journal of Digital Crime and Forensics*

*International Journal of Digital Evidence*

*International Journal of Electronic Security and Digital Forensics*

*Journal of Digital Forensic Practice*

*Journal of Digital Forensics, Security and Law*

*Small Scale Digital Device Forensics Journal*

## 11.3. Electronic Resources

BitCurator Project: http://www.bitcurator.net

Digital Forensics / Magazine. The Quarterly Magazine for Digital Forensics Practitioners: http://www.digitalforensicsmagazine.com/

Digital Forensic Research Workshop: http://www.dfrws.org/

Digital Lives Research Project: for report see http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf; for the original website (http://www.bl.uk/digital-lives) refer to the UK Web Archive

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

E-evidence info. The electronic evidence information center: http://www.e-evidence.info/

FIDO. Forensic Investigation of Digital Objects: http://fido.cerch.kcl.ac.uk/

Forensic Focus. For digital forensics and ediscovery professionals: http://www.forensicfocus.com

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

futureArch: http://www.bodleian.ox.ac.uk/beam/projects/futurearch and http://futurearch.blogspot.co.uk

Maresware. Software links for forensics investigative tasks: http://www.maresware.com/maresware/SITES/tasks.htm

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

## 12.References

ACPO, undated. Good practice guide for computer-based electronic evidence. Official release version.

AIMS, 2012. AIMS Born-digital collections an inter-institutional model for stewardship. Available: http://www2.lib.virginia.edu/aims/whitepaper/AIMS_final_A4.pdf.

Aitken, B, Helwig, P, Jackson, A, Lindley, A, Nicchiarelli, E and Ross, S 2008. The Planets testbed: science for digital preservation. *The code{4}lib Journal* [Online], 2008-06-23. Available: http://journal.code4lib.org/articles/83.

Altheide, C and Casey, E 2010. UNIX forensic analysis. In*:* Casey, E (ed.) *Handbook of Digital Forensics and Investigation.* London: Elsevier Academic Press.

Aquilina, JM, Casey, E and Malin, CH 2008. *Malware Forensics. Investigating and analyzing malicious code.* Burlington, MA, Syngress Publishing.

Axelsson, S 2010. The normalized compression distance as a file fragment classifier. *Digital Investigation,* 7**,** S24–S31.

Baier, H and Breitinger, F 2011. Security aspects of piecewise hashing in computer forensics. *Sixth International Conference on IT Security Incident Management and IT Forensics.* IEEE Computer Society. 21-36

Ball, A and Billenness, C 2012. Issues of information security applicable to the preservation of digital objects. In*:* Delve, J, Anderson, D, Dobreva, M, Baker, D, Billeness, C and Konstantelos, L (eds) *The preservation of complex objects. Volume 1. Visualisations and simulations.* POCOS, Preservation of Complex Objects Symposia, Portsmouth, University of Portsmouth. Available: http://eprints.port.ac.uk/7745/.

Barrett, D and Kipper, G 2010. *Virtualization and Forensics. A digital forensic investigator's guide to virtual environments.* London, Syngress.

Bates, P 2009. The rising impact of virtual machine hypervisor technology on digital forensics investigations. *ISACA Journal,* 6, 1–4. (ISACA, currently acronym only, formerly known as Information Systems Audit and Control Association.)

Bell, S 2008. *Crime and Circumstance: investigating the history of forensic science*. Greenwood Press.

Bem, D and Huebner, E 2007a. Analysis of USB flash drives in a virtual environment. *Small Scale Digital Device Forensics Journal,* 1, 1–6.

Bem, D and Huebner, E 2007b. Computer forensic analysis in a virtual environment. *International Journal of Digital Evidence,* 6**,** 1-13.

Bennett, CH, Li, M and Ma, B 2003. Chain letters and evolutionary histories. *Scientific American***,** June 2003, 76–81.

BIP, 2008. *Evidential Weight and Legal Admissibility of Information Stored Electronically*. Available http://www.thecabinetoffice.co.uk/page28.html

Birk, D 2011. Technical challenges of forensic investigations in cloud computing environments. Workshop on Cryptography and Security in Clouds, 15-16 March 2011, Zurich. Available: http://www.zurich.ibm.com/~cca/csc2011/abstracts.html

Böhme, R, Freiling, FC, Gloe, T and Kirchner, M 2009. Multimedia forensics is not computer forensics. In: Geradts, ZJMH, Franke, KY and Veenman, CJ (eds) *3rd International Workshop on Computational Forensics (IWCF 2009),* pp 90–103*.*

Brickell, E and Li, J 2009. Enhanced privacy ID: a remote anonymous attestation scheme for hardware devices. *Intel Technology Journal,* 13, 96–111.

Brueckner, S, Guaspari, D, Adelstein, F and Weeks, J 2008. Automated computer forensics training in a virtualized environment. *Digital Investigation,* 5, S105–S111.

Bruschi, D, Monga, M and Martignoni, L 2004. How to reuse knowledge about forensic investigations. *Digital Forensic Research Workshop (DFRWS 2004).* Linticum, MD.

Bryson, C and Stevens, S 2002. Tool testing and analytical methodology. In: Casey, E (ed.) *Handbook of Computer Crime Investigation. Forensic tools and technology.* London: Academic Press.

Buchholz, F and Falk, C 2005. Design and implementation of Zeitline: a forensic timeline editor. *Digital Forensic Research Workshop (DFRWS 2005).* New Orleans, LA.

Bunting, S and Wei, W 2006. *EnCase Computer Forensics. The Official EnCE EnCase Certified Examiner Study Guide.* Indianapolis, IN, Wiley Publishing.

Bursztein, E, Fontarensky, I, Martin, M and Picod, J-M 2011. Doing forensics in the cloud age. OWADE: beyond files recovery forensic. *Black Hat USA 2011.* Available: http://elie.im/talks/beyond-files-recovery-OWADE-cloud-based-forensic & http://www.blackhat.com/html/bh-us-11/bh-us-11-archives.html.

Carrier, B 2002. Open source digital forensics tools. The legal argument. *@stake Research Report.*

Carroll, L, Farr, E, Hornsby, P and Ranker, B 2011. A comprehensive approach to born-digital archives. *Archivaria* 72, 61–92.

Carvey, H and Altheide, C 2011. *Digital Forensics With Open Source Tools.* Waltham, MA, Syngress.

Casey, E 2000. *Digital Evidence And Computer Crime: Forensic science, computers and the Internet.* London, Academic Press.

Casey, E 2002a. Error, uncertainty, and loss in digital evidence. *International Journal of Digital Evidence,* 1.

Casey, E (ed.) 2002b. *Handbook of Computer Crime Investigation. Forensic tools and technology.* London: Academic Press.

Casey, E and Schatz, B 2011. Conducting digital investigations. In: Casey, E (ed.) *Digital Evidence and Computer Crime. Forensic science, computers and the Internet,* 3rd edn*.* London: Academic Press, Elsevier.

Cebrián, M, Alfonseca, M and Ortega, A 2009. Towards the validation of plagiarism detection tools by means of grammar evolution. *IEEE Transactions on Evolutionary Computation,* 13, 477–485.

CFTT, 2012. *Computer Forensics Tool Testing Handbook*. Computer Forensics Tool Testing Program, National Institute of Standards and Technology, US Department of Commerce. 73pp

Cilibrasi, R and Vitányi, PMB 2005. Clustering by compression. *IEEE Transactions on Information Theory,* 51, 1523–1545.

Cilibrasi, R, Vitányi, PMB and De Wolf, R 2004. Algorithmic clustering of music based on string compression. *Computer Music Journal,* 28, 49–67.

Cohen, ML and Schatz, B 2010. Hash based disk imaging using AFF4. *Digital Investigation,* 7, S121–S128.

Dappert, A, Jackson, A and Kimura, A 2011. Developing a robust migration workflow for preserving and curating hand-held media. *iPRES 2011 Conference. The Eighth International Conference on Preservation of Digital Objects.* Singapore.

DFRWS, 2001. A road map for digital forensic research. Report from the First Digital Forensic Research Workshop (DFRWS), 7–8 August 2001. *DFRWS Technical Report.* Utica, New York: Digital Forensic Research Workshop (DFRWS 2001).

Diamond, E 1994. The archivist as a forensic scientist: seeing ourselves in a different way. *Archivaria,* 38, 139–154.

Dias, Z, Rocha, A and Goldenstein, S 2010. First steps towards image phylogeny. *Workshop on Information Forensics and Security (WIFS 2010).* Seattle, WA: IEEE.

Duranti, L 2009. From digital diplomatics to digital record forensics. *Archivaria,* 68, 39–66.

Duranti, L and Endicott-Popovsky, B 2010. Digital record forensics: a new science and academic program for forensic readiness. *The Journal of Digital Forensics, Security and Law,* 5.

Dussault, HMB and Maciag, CJ 2004. Forensics, fighter pilots and the OODA loop: the role of digital forensics in cyber command and control. *Digital Forensic Research Workshop (DFRWS 2004).*

*Economist, The* 2011. Special report. *Personal technology. Beyond the PC*. 8 October 2011.

ENISA, 2011. Botnets: detection, measurement, disinfection and defence. European Network and Information Security Agency (ENISA) 153 pp.

Farid, H 2008. Digital image forensics. *Scientific American,* June 2008, 66–71

Farid, H 2009. Image forgery detection: a survey. *IEEE Signal Processing Magazine March 2009*, 16–25. Available http://www.emc.co.uk/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

Farquhar, A and Hockx-Yu, H 2007. Planets: integrated services for digital preservation. *The International Journal of Digital Curation,* 2, 88–99.

Feinerer, I, Hornik, K and Meyer, D 2008. Text mining infrastructure in *R. Journal of Statistical Software,* March 2008, 25(5), 1–54 .

Gantz, J F and Reinsel, D 2011. Extracting value from chaos. *IDC iView, June 2011.* Framingham, MA: IDC.

Garfinkel, S 2007. Carving contiguous and fragmented files with fast object validation. *Digital Investigation,* 4, S2–S12.

Garfinkel, S, Nelson, A, White, D and Roussev, V 2010. Using purpose-built functions and block hashes to enable small block and sub-file forensics. *Digital Investigation,* 7, S13–S23.

Garfinkel, S L 2010. Digital forensics research: the next 10 years. *Digital Investigation,* 7, S64–S73.

Garfinkel, S L 2011. Digital media triage with bulk data analysis and *bulk_extractor*. *Preprint submitted to Elsevier*.

Garfinkel, S L and Cox, D 2009. Finding and archiving the internet footprint. *Digital Lives Research Conference: Personal Digital Archives for the 21st Century.* British Library, London.

Garfinkel, S L, Farrell, P, Roussev, V and Dinolt, G 2009. Bringing science to digital forensics with standardized forensic corpora. *Digital Investigation,* 6**,** S2–S11.

Garfinkel, S L, Parker-Wood, A, Huynh, D and Migletz, J 2010. An automated solution to the mulituser carved data ascription problem. *IEEE Transactions on Information Forensics and Security,* 5**,** 868–882.

Grillo, A, Lentini, A, Me, G and Ottoni, M 2009. Fast user classifying to establish forensic analysis priorities. *Fifth International Conference on IT Security Incident Management and IT Forensics.* IEEE Computer Society 69-77.

Grispos, G, Glisson, W B and Storer, T 2011. Calm before the storm: the emerging challenges of cloud computing in digital forensics. *Preprint*, draft published for comment. Available: http://www.dcs.gla.ac.uk/~tws/papers/grispos11calm-rev2425.pdf.

Grundy, B J. 2008. The law enforcement and forensic examiner's introduction to Linux. A practitioner's guide to Linux as a computer forensic platform. Available: http://www.linuxleo.com.

Guđjónsson, K. 2010. Mastering the super timeline with log2timeline. *SANS Institute InfoSec Reading Room*. Available*: http://www.sans.org/reading_room/whitepapers/logging/mastering-super-timeline-log2timeline_33438*.

Hadjidj, R, Debbabi, M, Lounis, H, Iqbal, F, Szporer, A and Benredjem, D 2009. Towards an integrated e-mail forensic analysis framework. *Digital Investigation,* 5**,** 124–137.

Hallman, M 2011. log2timeline. Timeline creation and analysis. McLean, Virginia: SleuthKit and Open Source Digital Forensics Conference 2011, 14 June 2011 .

Hand, E 2011. Word play. *Nature,* 474**,** 436–440.

Higgins, S 2008. The DCC Curation Lifecycle Model. *The International Journal of Digital Curation,* 3**,** 134–140.

Higgins, S. 2010. Information security management: the ISO 27000 (ISO 27K) SERIES. *Briefing Papers* [Online]. Available: http://www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/information-security-management-iso-27000-iso-27k-s.

Hildebrandt, M, Kiltz, S and Dittmann, J 2011. A common scheme for evaluation of forensic software. *Sixth International Conference on IT Security Incident Management and IT Forensics (IMF 2011).* Stuttgart, Germany.

Huber, M, Mulazzani, M, Leithner, M, Schrittwieser, S, Wondracek, G and Weippi, E 2011. Social snapshots: digital forensics for online social networks. *Annual Computer Security Applications Conference.* ACM December 2011.

Huebner, E and Zanero, S (eds) 2010. *Open Source Software For Digital Forensics.* Springer Heidelberg.

Inoue, H, Adelstein, F and Joyce, R A 2011. Visualization in testing a volatile memory forensic tool. *Digital Investigation,* 8**,** S42–S51.

Irons, A 2006. Computer forensics and records management: compatible disciplines. *Records Management Journal,* 16**,** 102–112.

Janssen, W and Ayers, R 2007. Guidelines on cell phone forensics. Recommendations of the National Institute of Standards and Technology. Gaithersburg, Maryland: National Institute of Standards and Technology.

Ji, J-H, Park, S-H, Woo, G and Cho, H-G 2008. Generating p[h]ylogenetic tree of homogenous source code in plagiarism detection system. *International Journal of Control, Automation, and Systems,* 6**,** 809–817.

John, J L 2008. Adapting existing technologies for digitally archiving personal lives. Digital forensics, ancestral computing, and evolutionary perspectives and tools. *iPRES 2008 Conference. The Fifth International Conference on Preservation of Digital Objects.* The British Library, London.

John, J L 2009. The future of saving our past. *Nature,* 459**,** 775–776.

John, J L, Rowlands, I, Williams, P and Dean, K 2010. Digital Lives. Personal digital archives for the 21st century. An initial synthesis. *Digital Lives Research Paper. A project funded by the Arts and Humanities Research Council, UK.*

Joyce, R A, Powers, J and Adelstein, F 2008. MEGA: a tool for Mac OS X operating system and application forensics. *Digital Investigation,* 5**,** S83–S90.

Juola, P 2006. Authorship attribution. *Foundations and Trends in Information Retrieval,* 1**,** 233–334.

Kamde, P M and Algur, S P 2011. A survey of web multimedia mining. *The International Journal of Multimedia and Its Applications,* 3**,** 72–84.

Kaspersky, K 2006. *Data recovery. Tips and solutions: Windows, Linux and BSD.* Wayne, PA, A-LIST Publishing.

Kenneally, E 2001. Gatekeeping out of the box: open source software as a mechanism to assess reliability for digital evidence. *Virginia Journal of Law and Technology,* 13, Fall 2001. Available: http://www.vjolt.net/vol6/issue3/v6i3-a13-Kenneally.html.

Kim, K, Park, S, Chang, T, Lee, C and Baek, S 2009. Lessons learned from the construction of a Korean software reference data set for digital forensics. *Digital Investigation,* 6**,** S108–S113.

King, C and Vidas, T 2011. Empirical analysis of solid state disk data retention when used with contemporary operating systems. *Digital Investigation,* 8**,** S111–S117.

Kirschenbaum, M G 2008. *Mechanisms. New media and the forensic imagination.* Cambridge, MA, The MIT Press.

Kirschenbaum, M G, Farr, E, Kraus, K M, Nelson, N L, Stollar Peters, C, Redwine, G and Reside, D 2009a. Approaches to managing and collecting born-digital literary materials for scholarly use. *White Paper to the NEH Office of Digital Humanities.*

Kirschenbaum, M G, Farr, E L, Kraus, K M, Nelson, N, Peters, C S, Redwine, G and Reside, D 2009b. Digital materiality: preserving access to computers as complete environments. *iPRES 2009 Conference. The Sixth International Conference on Preservation of Digital Objects.* Stanford University, California.

Kirschenbaum, M G, Ovenden, R, Redwine, G and Donahue, R 2010. Digital forensics and born-digital content in cultural heritage collections. Washington DC.

Kokocinski, A 2010. Macintosh forensic analysis. In: Casey, E (ed.) *Handbook of Digital Forensics and Investigation.* London: Academic Press.

Kornblum, J 2006. Identifying almost identical files using context triggered piecewise hashing. *Digital Investigation,* 3S**,** S91–S97.

Kraus, K 2009. Conjectural criticism: computing past and future texts. *Digital Humanities Quarterly,* 3.

Kruse, W G, II and Heiser, J G 2002. *Computer Forensics. Incident response essentials.* Boston, Addison-Wesley.

Larson, T 2002. The other side of civil discovery: disclosure and production of electronic records. In: Casey, E (ed.) *Handbook of Computer Crime Investigation. Forensic tools and technology.* London: Academic Press.

Leetaru, K H. 2011. Culturomics 2.0: forecasting large scale human behavior using global news media tone in time and space. *First Monday. Peer-reviewed Journal on the Internet* [Online], 16. Available: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/3663/3040.

Lillard, T V, Garrison, C P, Schiller, C A, Steele, J and Murray, J 2010. *Digital Forensics for Network, Internet, and Cloud Computing. A forensic evidence guide for moving targets and data.* Burlington MA, Syngress, Elsevier.

Loftus, M J 2010. The author's desktop. How famous writers' computers – like Salman Rushdie's Macs in Emory's Manuscript, Archives, and Rare Book Library – and born-digital content are creating a revolution. *Emory Magazine*. Available: http://www.emory.edu/EMORY_MAGAZINE/2010/winter/authors.html.

Louis, A L. 2009. *Unsupervised discovery of relations for analysis of textual data in digital forensics.* Master of Science, University of Pretoria.

Lowe, D and Hall, W 1999. *Hypermedia and the Web. An engineering approach.* Chichester, John Wiley & Sons.

Ma, M, Zheng, H and Lallie, H 2010. Virtual reality and 3D animation in forensic visualisation. *Journal of Forensic Sciences,* 55, 1227–1231.

Mandia, K, Prossise, C and Pepe, M 2003. *Incident Response and Computer Forensics.* Emeryville, CA, McGraw-Hill/Osborne.

Manoff, M 2006. The materiality of digital collections: theoretical and historical perspectives. *portal: Libraries and the Academy,* 6**,** 311–325.

Marrington, A, Baggili, I, Mohay, G and Clark, A 2011. CAT Detect (Computer Activity Timeline Detection): a tool for detecting inconsistency in computer activity timelines. *Digital Investigation,* 8**,** S52–S61.

Marziale, L, Richard, I, Golden G. and Roussev, V 2007. Massive threading: using GPUs to increase the performance of digital forensic tools. *Digital Investigation,* 45**,** S73–S81.

Mayer-Schönberger, V 2009. *Delete. The virtue of forgetting in the digital age.* Princeton, Princeton University Press.

McDonough, J, Olendorf, R, Kirschenbaum, M, Kraus, K, Reside, D, Donahue, R, Phelps, A, Egert, C, Lowood, H and Rojo, S 2010. Preserving virtual worlds. Final report. Available: http://hdl.handle.net/2142/17097.

McElhearn, K 2005. *The Mac OS X Command Line. Unix under the hood.* Alameda, California, Sybex.

Neamatullah, I, Douglass, M M, Lehman, L-W H, Reisner, A, Villarroel, M, Long, W J, Szolovits, P, Moody, G B, Mark, R G and Clifford, G D. 2008. Automated de-identification of free-text medical records. *BMC Medical Informatics and Decision Making* [Online], 8. Available: http://www.biomedcentral.com/1472-6947/8/32.

Nelson, T H 1995. The heart of connection: hypermedia unified by transclusion. *Communications of the ACM,* 38**,** 31–33.

Nickell, J 2005. *Detecting Forgery. Forensic investigation of documents.* Lexington, University Press of Kentucky.

NTAIA 2009. Technical risk assessment, HMG IA Standard Number 1. Cheltenham: Communications-Electronics Security Group (National Technical Authority for Information Assurance).

NTAIA 2011. Supplier information assurance assessment framework and guidance, Issue Number: 1.0. Cheltenham: Communications-Electronics Security Group (National Technical Authority for Information Assurance).

Ohm, P 2010. Broken promises of privacy: responding to the surprising failure of anonymization. *UCLA Law Review,* 57**,** p. 1701-77.

Olsson, J and Boldt, M 2009. Computer forensic timeline visualization tool. *Digital Investigation,* 6**,** S78–S87.

Olson,M 2010.Computer forensics in the archive: an analysis of software tools for born digital collections. Digital Humanities 2010 (DH 2010). King's College, London.

Pal, A, Sencar, H T and Memon, N 2008. Detecting file fragmentation point using sequential hypothesis testing. *Digital Investigation,* 5, S2–S13.

PARADIGM, 2007. Workbook on Digital Private Papers. Paradigm Project. Available: http://www.paradigm.ac.uk/workbook/index.html.

Penhallurick, M A 2005a. Methodologies for the use of VMware to boot cloned/mounted subject hard disk images. *Digital Investigation,* 2**,** 209–222.

Penhallurick, M A 2005b. Methodologies for the use of VMware to boot cloned/mounted subject hard disk images. Cranfield University.

Pfleeger, C H 1997. *Security in Computing.* London, Prentice-Hall International.

Pittman, R D and Shaver, D 2010. Windows forensic analysis. In: Casey, E (ed.) *Handbook of Digital Forensics and Investigation.* London: Elsevier Academic Press.

PLANETS. 2010. *The digital divide. Assessing organisations' preparations for digital preservation* [Online]. Available: http://www.planets-project.eu/docs/reports/planets-market-survey-white-paper.pdf.

Pogue, C, Altheide, C and Haverkos, T 2008. *UNIX and Linux forensic analysis DVD toolkit.* Burlington MA, Syngress Publishing, Elsevier.

Redwine, G, Kirschenbaum, M, Olson, M and Farr, E 2010. Born digital: the 21st century archive in practice and theory. *Digital Humanities 2010 (DH2010).* King's College, London.

Reed, C A and Rowe, G W A 2001. Araucaria: software for puzzles in argument diagramming and XML. *Technical Report.* Department of Applied Computing, University of Dundee, see also http://araucaria.computing.dundee.ac.uk/doku.php.

Richard III, G G and Roussev, V 2006. Secure, audited processing of digital evidence: filesystem support for digital evidence bags. *Second Annual IFIP Working Group 11.9 International Conference on Digital Forensics.* Orlando, FL: International Federation for Information Processing Working Group 11.9 on Digital Forensics.

Ross, S and Gow, A 1999. Digital archaeology: rescuing neglected and damaged data resources. Technical report. *British Library Research and Innovation Report.* London: British Library.

Roussev, V 2010. Data fingerprinting with similarity digests. In: Chow, K and Shenoi, S (eds) *Research Advances in Digital Forensics.* Springer.

Roussev, V 2011. An evaluation of forensic similarity hashes. *Digital Investigation,* 8**,** S34–S41.

Sammes, T and Jenkinson, B 2000. *Forensic Computing. A practitioner's guide,* London, Springer-Verlag.

Sammes, T and Jenkinson, B 2007. *Forensic computing. A practitioner's guide,* London, Springer-Verlag. 2nd edn.

Schofield, D 2009. Graphical evidence: forensic examinations and virtual reconstructions. *Australian Journal of Forensic Sciences,* 41, 131–145.

Seglem, K K, Luque, M and Murphy, S 2002. Unix system analysis. In: Casey, E (ed.) *Handbook of Computer Crime Investigation. Forensic tools and technology.* London: Academic Press.

Shavers, B. 2008. *Virtual machine forensics, a discussion of virtual machines related to forensic analysis* [Online]. Available: http://www.forensicfocus.com/downloads/virtual-machines-forensics-analysis.pdf.

Shipman, A 2004. Code of practice for legal admissibility and evidential weight of information stored electronically. London: British Standards Institution pp141

Shipman, A and Howes, P 2005. *Code of practice for legal admissibility and evidential weight of linking electronic identity to documents*. London: British Standards Institution.

Sivic, J, Russell, B C, Efros, A A, Zisserman, A and Freeman, W T 2005. Discovering objects and their location in images. *Tenth IEEE International Conference on Computer Vision (ICCV 2005).* IEEE Computer Society.

Sommer, P 2009. Forensic science standards in fast-changing environments (based on a presentation, Keeping up: testing methodologies in digital forensics). *In:* European Academy of Forensic Science (EAFS 2009). *(Prepublication: final version in Science and Justice* 50(11): 12–17, March 2010*.)*

Sommer, P and Brown, I 2011. Reducing systemic cybersecurity risk. *OECD/IFP Project on 'Future Global Shocks'*, 14 January 2011, OECD: IFP/WKP/FGS (2011) 3. Available: http://www.oecd.org/dataoecd/3/42/46894657.pdf.

Teredesai, A M, Ahmad, M A, Kanodia, J and Gaborski, R S 2005. CoMMA: a framework for integrated multimedia mining using multi-relational associations. *Knowledge and Information Systems*. 10, 135–162. Available: http://faculty.washington.edu/ankurt/Publications_files/COMMAs10115-005-0221-x.pdf.

Thomas, S 2011. Curating the I, Digital: experiences at the Bodleian Library. In: Lee, CA (ed.) *I, Digital: personal collections in the digital era*. Chicago: Society of American Archivists.

Trace, C B 2011. Beyond the magic to the mechanism: computers, materiality, and what it means for records to be 'born digital'. *Archivaria,* 72, 5–27.

Turner, P 2005. Unification of digital evidence from disparate sources (digital evidence bags). *Digital Forensic Research Workshop* (DFRWS 2005). New Orleans, LA.

UNIDO 2009. Complying with ISO 17025. A practical guidebook for meeting the requirements of laboratory accreditation schemes based on ISO 17025:2005 or equivalent national standards. Vienna: United Nations Industrial Development Organization.

Valli, C 2010. An analysis of malfeasant activity directed at a VoIP honeypot. *In:* Woodward, A (ed.) *8th Australian Digital Forensics Conference.* Perth, Western Australia: Security Research Centre (secau), School of Computer and Security Science, Edith Cowan University, pp 169–174.

Van Der Knijff, R 2010. Embedded systems analysis. In: Casey, E (ed.) *Handbook of Digital Forensics and Investigation.* London: Elsevier Academic Press.

Van Gelder, T 2002. Enhancing deliberation through computer-supported argument visualization. *In*: Kirschner, P A, Buckingham Shum, S J, and Carr, C S. *Visualizing Argumentation: Software Tools for Collaborative and Educational Sense-Making*. London: Springer.

Vlastos, E and Patel, A 2007. An open source forensic tool to visualize digital evidence. *Computer Standards and Interfaces,* 29, 614–625.

Walden, I 2004. Forensic investigations in cyberspace for civil proceedings. *International Review of Law, Computers and Technology,* 18, 275–287.

Wang, W. 2009. *Digital Video Forensics.* PhD, Dartmouth College.

Wang, W and Farid, H 2006. Exposing digital forgeries in video by detecting double MPEG compression. Proceedings of the 8th Workshop on Multimedia and Security (*MM&Sec 2006),* 26-27 September 2006, Geneva, Switzerland.

Willassen, S Y and Mjølsnes, S F 2005. Digital forensics research. *Telektronikk, 1.2005, 92-97.*

Willassen, S Y 2008 Methods for enhancement of timestamp evidence in digital investigations, PhD, Norwegian University of Science and Technology.

Wolfram, S. 2012. The personal analytics of my life. *Stephen Wolfram Blog* [Online]. Available from: http://blog.stephenwolfram.com/2012/03/the-personal-analytics-of-my-life/.

Woods, K, Lee, C A and Garfinkel, S 2011. Extending digital repository architectures to support disk image preservation and access. *Joint Conference on Digital Libraries (JCDL 2011).* Ottawa, Ontario, Canada.

World Economic Forum, 2011. Personal data: the emergence of a new asset class. Cologny/Geneva: World Economic Forum (with Bain & Company, Inc.).

Yasinac, A, Erbacher, R F, Marks, D G, Pollitt, M and Sommer, P, M. 2003. Computer forensics education. *IEEE Security and Privacy***,** 15–23.

Zhang, Y, Mueller, F, Cui, X and Potok, T 2009. GPU-accelerated text mining. Exploiting Parallelism Using GPUs and Other Hardware-Assisted Methods *(EPHAM 2009)*, 2009 International Symposium on Code Generation and Optimization*,* 22–25 March 2009, Seattle, Washington: ACM.