Adapting Existing Technologies for Digitally Archiving Personal Lives

Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools

Jeremy Leighton John

Department of Western Manuscripts, Directorate of Scholarship and Collections, The British Library
96 Euston Road, LONDON NW1 2DB, United Kingdom
jeremy.john@bl.uk

Abstract

The adoption of existing technologies for digital curation, most especially digital capture, is outlined in the context of personal digital archives and the Digital Manuscripts Project at the British Library. Technologies derived from computer forensics, data conversion and classic computing, and evolutionary computing are considered. The practical imperative of moving information to modern and fresh media as soon as possible is highlighted, as is the need to retain the potential for researchers of the future to experience the original look and feel of personal digital objects. The importance of not relying on any single technology is also emphasised.

Introduction

Archives of 'personal papers' contain letters, notebooks, diaries, draft essays, family photographs and travel cine films; and in 2000 the British Library adopted the term eMANUSCRIPTS (eMSS) for the digital equivalent of these 'personal papers', having begun accepting diverse computer media as part of its manuscript holdings (Summers and John 2001, John 2005).

These media include punched cards, paper tapes, magnetic tapes, program cards, floppy disks of several sizes (8", 5.25", 3.5" and 3"), zip disks, optical disks (eg CDRs and DVDRs) and various hard drives, both internal and external. All three major contemporary operating system families are represented: Microsoft Windows, Apple Macintosh and Unix/Linux as well as earlier systems.

Beyond the library's own collections, the Digital Manuscripts Project has enabled digital capture for the Bodleian Library, the Royal Society (with the National Cataloguing Unit for the Archives of Contemporary Scientists), and the Wellcome Library.

Digital Manuscripts at the British Library

The primary aim of the project is to develop and put into place the means with which to secure the personal archives of individuals in the digital era in order to enable sustained access. This entails the capture of the digital component of the archive alongside its corresponding analogue component.

The project is also addressing in tandem the digitisation of the conventional papers in personal archives (and in that sense is also concerned with digital manuscripts beyond eMSS). Among other benefits, this will make it easier for researchers to work with an entire personal archive in an integrated way; but this work along with cataloguing and resource discovery is beyond the scope of the present paper, which aims to focus on the curatorial role in digital acquisition, examination and metadata extraction.

Theoretical and Practical Considerations

The challenges of technological obsolescence, media degradation and the behaviour of the computer user (eg failure to secure and backup information including passwords) are long familiar to the digital preservation community. Personal collections raise issues, however, that are different from those arising with publications, which have received far more attention.

Of special relevance is the means of acquiring personal archives. Central to the process is the relationship between the curator and the originator or depositor, and in particular the need to deal with personal matters in a sensitive way, ensuring robust confidentiality where necessary.

Three key requirements have been identified and promoted: (i) to capture as far as possible the whole contextual space of the personal computer (the entire hard drive or set of hard drives for example) and not just independent individual files, thereby strengthening authentication; (ii) to replicate and retain exact copies of the original files, recognising their historical and informational value (and not just rely on digital facsimiles, even if these match modern standards for interoperability); and (iii) to meet the special requirements for a confidentiality that is sensitive and reassuring to potential depositors as well as being technically convincing.

A pragmatic philosophy is to provide for immediate access to basic text, images and sounds (eg raw alphanumeric content which will suffice for many scholarly purposes); but to retain (by capturing and keeping exact digital replicates of disks and files) the potential to make available high fidelity versions that respect original styles, layout and behaviour.

The Digital Capture Imperative

Future work with personal archives can be expected to be increasingly proactive and entail a close understanding with and involvement of originators and their families and friends. The single most important consequence of the increasingly digital nature of personal archives is the need to preempt inadvertent loss of information by providing advice and assistance.

The key threshold is the initial digital capture: the movement of the eMANUSCRIPT information to modern, fresh and secure media.

Adapting Existing Tools

An effective and potentially efficient route for successful digital capture, preservation and access is to adopt and modify existing technologies for new purposes rather than necessarily designing from scratch.

In this spirit, three key technologies are being examined: (i) computer forensic software and hardware; (ii) ancestral computers, disk and tape drives with associated controllers and software emerging from communities of enthusiasts; and (iii) evolutionary computing techniques and perspectives.

Computer Forensics

In computer forensics there are three text book stipulations (eg Kruse and Heiser 2001, Casey 2001, Carrier 2005, Sammes and Jenkinson 2007): (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. There are, moreover, certifiable standards with which computer forensic scientists must comply in order to satisfy legal authorities. Guides to good practice include ACPO (2003) and NIJ (2004). These requirements match in a number of ways the concerns of the digital curator of personal archives.

A wide range of forensic software and hardware has been explored at the British Library for its applicability in capturing, examining and authenticating eMSS. Software that has been and is currently being surveyed and tested includes: Forensic Toolkit (FTK) of AccessData; Macintosh Forensic Suite (MFS) of BlackBag Technologies; Image, PDBlock, DriveSpy and others of Digital Intelligence; Helix of e-fense; Encase of Guidance Software; CD/DVD Inspector of InfinaDyne; Device Seizure, Email Examiner and others of Paraben. Products which have not been examined include ILook Investigator and X-Ways Forensics.

Open source forensic software tools include: Back Track; Coroner's Toolkit; Foremost; Foundstone Forensic Tools; Open Computer Forensics Architecture; Scalpel; and Sleuth Kit with Autopsy.

Forensic hardware includes high specification workstations with forensically compatible BIOS (eg Digital Intelligence), diverse write-blockers (eg Tableau) and robotic floppy disk and optical disk imagers (eg WiebeTech) as well as numerous connectors and adaptors.

Overview of Available Functionality

This equipment provides a plethora of capabilities including the write-protection of original collection source disks, certified wiping of target receiving disk (even brand new drives can contain digital artefacts), the forensically-sound bitstream 'imaging' of the original disk, the creation of unique hash values (MD5, SHA1 and related algorithms) for the entire disk and for individual files, and the recovery of fragments of lost files.

Other functionality includes the ready export of replicate files, the bookmarking and annotating of files of interest for summary reports, timeline viewers for investigating times and dates of file creation, modification and access, while taking into account different time zones, provisional identification of file types based on file signatures and extensions, maintenance of an examination audit trail, filtering of files that are not of immediate interest to an examining curator (eg software files), sophisticated searching (with GREP), file viewing, and reading of emails with carving out of attachments.

Available forensic products are subject to ongoing and rapid development and any attempt to identify the best of them risks being anachronistic. There is no single product that will meet all requirements of the forensic examiner or for that matter the digital curator or preservation expert, which explains why there is a flourishing diversity of specialist products.

Two of the most well established are Encase and FTK, both of which seek to be comprehensive, encompassing in one package much of the functionality just outlined. Both work with a wide range of file systems, and are convenient and comparatively straightforward to use, while still providing capabilities for hexadecimal viewing and analysis of disk and file system geometry. Encase has recently incorporated "Outside In" technology from Stellent for the viewing of files from over 400 file formats. Following its recent major upgrade, FTK now works natively with Oracle's database technology. Other companies such as Paraben provide numerous software modules that are dedicated to specific capabilities and are able to work either separately or together as a more integrated whole, as with P2 Commander.

On the other hand, CD/DVD Inspector specialises in the analysis of optical discs, which show some profound differences from hard disks in the forensic context (Crowley 2007). A standard ISO 'image' does not capture all of a CD's potential contents, but CD/DVD Inspector is able to do so, producing a file that can be imported into Encase for example. It is also able to work with the sometimes awkward Universal Disk Format.

Helix is another specialist: essentially a forensically customised adaptation of Knoppix. In this Linux mode it serves as a bootable CD with a self-contained operating system that will not write to the attached hard drives, and which can create nonproprietary forensic 'images'. (It also

operates in a Windows mode, mainly concerned with the forensics of live machines.) Moreover, it is accompanied by an assortment of other largely standalone tools (including some of the openly available ones mentioned), making it a kind of forensic Swiss Army knife.

The essential workflow adapted for the curation of information from contemporary computer media in personal digital archives can be considered in two phases: before and after capture.

Phase 1. The Core of the Capture Workflow

There are three initial key requirements: (i) audit trail; (ii) write-protection; and (iii) forensic 'imaging', with hash values created for disk and files. (The term 'disk' is being used here loosely to refer to floppy, zip, optical and hard disks, flash media and others.)

The first recommended practice is for there to be a chain of custody from the moment that the original materials become available continuing throughout the lifecyle of the entire capture process, recording procedures undertaken by the curator. (At the end of the workflow, the audit culminates with a detailed report.) It is possible to use specialist tools such as Adepto (from e-fense) which will provide an audit log and chain of custody form on acquiring a forensic 'image'. An advantage of the more comprehensive packages is that the audit control, record making and documentation, is seamlessly integrated and automatic, and in some cases embedded along with the 'forensic' image. Digital photos taken by the curator at the time of collection, can be imported into the integrated systems, as can photos of all of the computer media (along with labels) in the personal archive.

The initial motivation for adopting computer forensics arose from the simple concern that even turning on the computer of an originator risked modifying important dates and times of historic interest. It is one of the rules (sometimes needing to be broken) of forensic science not to switch on the originator's computer (even lifting the lid of some laptops may turn them on); but instead to remove all of the hard drives and connect them to the examiner's computer using write-blockers.

The main and sometimes necessary alternative to the use of a hardware write-blocker is to again connect the original hard drive to the examining workstation but to boot this computer from a forensically prepared floppy disk or CD, being very careful (typically by modifying BIOS) not to allow the computer to boot from the original hard drive (eg Helix, Encase for DOS, or LinEn).

The long established workhorse of bitstream 'imaging' is the 'dd' command under UNIX. In principle, this produces a single file encapsulating the entire digital contents of the disk (in practice, it is often a series of conveniently smaller files). An open source forensic version has been developed (dcfldd) with hash values produced on the fly, and additional features (originally developed by the Department of Defence Computer Forensics Laboratory, and available at sourceforge.net). One drawback of Encase's compressed 'image' file from the perspective of digital curation is its proprietary nature. FTK Imager (which is part of FTK but obtainable separately and free of charge) can create both proprietary and nonproprietary 'images' including "dd', as well as computing hash values for disk and files.

The strategy adopted by the Digital Manuscripts Project has been to use both facilities, checking that the same hash values are achieved, as a means of corroborating successful capture, while retaining the nonproprietary 'image' file and independently obtained hash values for future reference. It is strongly recommended that digital curators do not rely on any single tool or technology.

Phase 2. Consolidation of the Capture Workflow

The workflow continues with four remaining functional activities: (iv) examination and consideration by curators (and originators), with filtering and searching; (v) export and replication of files; (vi) file conversion for interoperability; and (vii) indexing and metadata extraction and compilation.

With the successful capture of the disk and checks for viruses and other malware completed, examination of its contents can proceed. Sometimes this will be the first time that curator and originator are able to look extensively at the eMSS.

The hash values of the files can be compared with a known hash library for application and operating system software files, allowing these to be identified and filtered out from immediate consideration. Scripts are available and can be customised for refined searching and filtering, based on file signatures, keywords and other criteria. Digital content entailing specific digital rights issues such as intellectual property, data protection or requested confidentiality can be identified by the curator and bookmarked. Any files with credit card numbers, telephone numbers, post codes or email addresses for example, can be automatically located and listed.

Files can be exported in their original form as exact digital replicates providing future scholars with the potential for use with, for example, an authenticated emulator of application software. For more immediate and practical access, the files can be converted into an interoperable form (with low fidelity if not high fidelity) such as a member of the XML or PDF families, where deemed appropriate by the digital preservation community.

Moreover, a digital replicate of the original drive can be restored to a similar or larger hard drive to be inserted into an appropriate computer if desired; however, this presents the same potential problem as before, interacting with this computer will alter the system. The Digital Manuscripts Project is currently examining the use of special and general hardware write-blockers for interacting with a dynamic system.

Encase provides two other options: Physical Disk Emulator (PDE) and Virtual File System (VFS). These modules allow the 'image' bitstream to be mounted in read-only mode in a Windows environment. A key difference between them is that PDE, in contrast to VFS, will behave as a normal volume, and not provide access to

unallocated space or deleted files. One useful aspect is that these read-only systems can be scanned for viruses and other malware before exporting any files to the examiner's computer. PDE can be used with the virtualisation software VMware, which will mount the PDE disk as a virtual machine that can be booted virtually. PDE and VFS are both proprietary; but the nonproprietary 'dd' file can also be mounted in VMware (which is itself proprietary though very widely available; open source software such as Xen and QEMU may be useful, however).

InfinaDyne offers a sister tool that can be used to produce a replicate optical disk from the CD/DVD Inspector 'image' file.

The principal forensic tools can conduct deep indexing (incorporating text within files) and extraction of metadata relating to files including file extension, file type, file signature, dates and times, permissions, hash values, logical size, physical location, file extents (fragmentation). In addition, metadata associated with emails (and webmail), photos and instant messages for instance can be extracted. The open source Sleuth Kit (with or without the GUI of Autopsy) is a useful alternative. Metadata for the disks and tapes themselves can be compiled.

Originators, Other Depositors and Third Parties

The essential need to involve potential depositors in the capture process cannot be overemphasised. In addition to assisting in the identification of eMSS where there are data protection and confidentiality requirements (including for third parties), originators can provide contextual and corroborating information that increases the scholarly and historical value of the entire digital archive.

Recovery, even if only in the form of fragments, of partially overwritten, inadvertently or regretfully deleted, earlier drafts of creative works could be of great scholarly interest but it must involve the originators and accord with their wishes. On the other hand, establishing the provenance of fragments of deleted files can sometimes be forensically demanding (Sammes and Jenkinson 2006), and again the creator's confirmation of authenticity might be invaluable. Much better in the long run, of course, would be if creators would know how to manage and care for their personal archive, assisted perhaps by advice from curators and digital preservation specialists.

Passwords are sometimes forgotten or records are accidentally lost, and with the permission of family and originators, decryption and password recovery tools can be used with varying levels of success.

An initial examination of a digital archive can be facilitated at the home of the creator using a forensic laptop and a preview facility that does not entail actual acquisition, helping curators and creators decide whether an archive fits into the collection development policy of the repository before being transferred there. It may be the intention of the originator to simply donate some specific folders or files rather than a disk. A 'logical' acquisition of files can be conducted forensically in much the same way as a 'physical' acquisition of an entire disk.

Ancestral Computing

At present there are, for archival purposes, two limitations to computer forensic technologies: (i) a limited ability to cater for legacy computers, storage media and software even with regard to the initial capture of the information that exists on obsolete media; and (ii) a limited ability to present the files and computer working environment identical or close to the way it was perceived by the creator (even in the case of many contemporary files) with styles, layout and behaviour accurately demonstrated and certified.

It is also necessary to understand the way users interacted with their computers, how these worked technically, the applications that were available to users and the nature of the files produced — just as curators of conventional manuscripts are required to know about the ways in which writing media (wax, parchment, vellum, paper) and associated technologies (pen, ink, pencil, stylus) were designed and used.

This section looks at the initial capture of the information on ancestral computer media. As with deleted files, it is essential to involve originators and their families, as they may not have seen the files residing on the obsolete media for many years.

There is an important and frequently misunderstood distinction between digital capture and digital preservation. Guides to digital preservation have been anxious to dispel any notion of technology preservation as a tenable solution. However, the use of ancestral computer technology for digital capture is unavoidable at present.

Files existing on 3.5" and 5.25" floppy disks and derived from Microsoft DOS and early Windows systems can often be replicated within Windows 98, in DOS mode where necessary, on a relatively recent PC computer furnished with corresponding floppy disk drives. Longstanding forensic tools can help (eg Digital Intelligence's Image, an imaging tool specifically designed for floppy disks).

More challenging are the hundreds if not thousands of species of computer systems which were famously diverse during the 1980s and early 1990s before Microsoft DOS and Windows came to predominate (with varying combinations of processors, operating systems and ROM, and disk systems and actual hardware types) (Nadeau 2002).

Publishing and Typesetting

During and after this period there was a widely felt need to convert files from one type to another, as witnessed by guides to file formats such as Walden (1986, 1987), Kussmann (1990), Swan (1993) and the encylopaedic Born (1995). The need to create a degree of interoperability in order to move data between applications has long been one of the major motivations for the reverse engineering of software (eg Davis and Wallace 1997).

One community that required duplication and conversion technologies in the 1980s and 1990s were publishers and

their typesetters, who needed to read and convert files derived from diverse sources (ie writers) to a local standard that could be used by the in-house computer and printing equipment.

InterMedia was a UK company that specialised in supplying media and data conversion systems for over two thousand floppy disk and hardware and operating system combinations. The National Library of Australia has used the system for 5.25" and 3.5" floppy disks (Woodyard 2001).

The company has been bought up by a USA company, eMag Solutions, which retains offices in the UK. An InterMedia system, now renamed eMag Floppy Disk Conversion System Model MMC4000 has been obtained by the British Library with the InterMedia software and Stack-a-Drives for 8", 5.25", 3" and 3.5" floppy disks working with a proprietary floppy disk controller.

One success has been the capture and transfer of files to modern media from 8" floppy disks, dating from a quarter of a century ago. The equipment has also been used to read hundreds of files residing in 3" and 5.25" floppy disks dating from two decades or so. So far there have been relatively few cases where disks have been entirely unreadable: occasionally degradation can be seen in the physical condition of the disk, ie a light reddish brown surface apparently indicative of oxidisation.

Typically the system would have been used to read and convert files, derived from word processors, from one type to another that can be read by modern PCs, using basic Translation Tables as well as program Protocols that can handle pointers. There is a Disk Recogniser function that will sometimes though not invariably assist in identifying disk types. Original files would be converted to a proprietary InterMedia Internal Coding (IMIC), to be subsequently converted to a file with the desired format.

The later version of InterMedia for Microsoft Windows software (IMWIN, Windows XP) is convenient to use but the earlier version for Microsoft DOS (InterMedia, DOS version) is more powerful. It is geared towards a more complete analysis and replication of floppy disks at the most basic levels. Disks can be interrogated at the clock and bit level. In the reading and copying of sectors, disks with hard sectors as well as those with soft sectors can be addressed.

The approach adopted by the Digital Manuscripts Project is, as far as feasible: (i) to copy the individual files in their original format (file digital replicates); (ii) to copy the entire disk (disk digital replicate); and (iii) to create and retain converted files that provide the basic alphanumeric content as low fidelity copies (eg as Word documents) which are later converted to an interoperable form such as PDF.

One simple but useful extension of the overall workflow is to import these files into the forensic system, thereby creating hash values for all the files and integrating them with other files and providing an audit trail.

In addition, MediaMerge for PC (MMPC) has been obtained from eMag Solutions for reading and copying tapes. The user can view and duplicate at block level as well as copying the individual files. A series of 0.25" data cartridges derived from UNIX computers active in the 1990s have been copied by this means.

While it has been very satisfying to capture historically important files using these systems, relying (in the case of floppy disks) on proprietary technology that is no longer fully supported and developed, is clearly not a sustainable solution. The inherent knowledge in this and other data conversion systems is being pursued by several avenues.

Another key source of useful technology for the purposes of the Digital Manuscripts Project has been and will be the classic, retro and vintage computer communities.

Expert Enthusiasts

As a result of continuing enthusiasm for these ancestral computer systems, a small German company called Individual Computers has produced modern technology that enables the reading of early format floppy disks. Specifically, the Catweasel is a universal floppy disk controller that can be used with modern PCs and normal floppy disk drives for 5.25" and 3.5" floppy disks (and in principle others too).

The manufacturer has indicated that Catweasel will work with the following formats (many though not necessarily all variants): Amiga, Apple IIe, early Apple Macintosh, Atari, Commodore and PC, with more planned.

Its attraction lies in its flexibility and degree of openness. Catweasel MK 4 is a low profile PCI card that uses FPGA chips (Field Programmable Gate Arrays) that provide it with reconfigurable logic meaning that software drivers for currently unsupported disk formats can be downloaded when these become available (from Individual Computers itself or expert enthusiasts), and used to reprogram the Catweasel without removing it from the computer. With the appropriate software, it can be used with Linux computers, and use with Mac OS X is anticipated.

The Digital Manuscripts Project has installed the device and is currently exploring its capabilities.

Individual Computers is also involved with other developers in the Commodore One (C-One Reconfigurable Computer). This is a computer that began in 2003 as an enhanced adaptation of the venerable Commodore 64 (C64), one of the most prolific computers of all time. The current version of the C-One (actually a motherboard that can be used with widely available hardware components such as an ATX type computer case) is reconfigurable, again due to FPGA chips. This means that the same basic hardware system can be modified so that it can behave like another early computer such as the C64's sister, the VIC 20, or the Schneider CPC, Atari, or Sinclair Spectrum and others. Expert users are encouraged to create their own FPGA cores using the free development tool Quartus by Altera. Furthermore, with project Clone-A, Individual Computers is developing a cycle-accurate reproduction of original chipsets in Amiga computers.

Equally this hardware is being matched by software made available for and within the various classic computer communities. Copies of the original software may still be available, as is the case for LocoScript for use with the CP/M operating system running on the Amstrad PCW series of computers.

Other software seeks to emulate the original hardware, operating systems and applications. At the forefront is the Amiga community, for example, with Amiga Forever (preconfigured Amiga ROM, OS and application software files) running on Apple OS X (say) using a hardware emulation derived from the open source Ubiquitous Amiga Emulator (UAE) and Fellow emulator of Amiga hardware. Emulators of application and operating system software are also produced of course which allow early applications to run directly on modern operating systems: for example AppleWin emulates Apple IIe in Windows (available at berliOS).

There are essentially two sentiments in classic computing: (i) a desire to respect and maintain the original nature of the computer system of interest, down to the exact sounds emitted; and (ii) a desire to ensure the continuing and strengthened relevance of the system by adding modern and new features to it, not least in its interfacing capabilities. This observation and the varying extent to which high fidelity is achieved even when sought points to the crucial role for digital curation and preservation specialists in the certified authentication of these kinds of products. Key institutional resources in this endeavour (in the UK) will be the Science Museum, the National Museum of Computing at Bletchley Park, the Computer Conservation Society, and others, with their expertise and representatives of original equipment.

Along with originators' computers, personal archives frequently contain original software disks and manuals which are likewise retained and used, with permission.

Evolutionary Perspectives and Tools

There are many examples of engineers adapting or copying technologies from nature. Perhaps none is as profound in the digital context as the adoption of DNA itself as a tool.

Digital information, of course, lies at the heart of life in the form of DNA. This has led to the development of DNA computing. But of more direct interest to the present conference is the proposal to use DNA as a means of longterm storage of information (Wong, Wong, and Foote 2003). Three observations have been used to support the idea (Bancroft et al. 2001): (i) viable bacteria have been reported in salt crystals dating from 250 million years; (ii) DNA is the genetic information of humans and therefore will remain central to civilization and scientific progress; and (iii) enormous numbers of identical molecules can be created to ensure informational redundancy to mitigate against stochastic loss.

A vision of DNA encoded library and archival information is a fascinating one but although there are clear advantages to the use of DNA not least in its compact form, the real question to ask is how did it come to be? It is not just a matter of the medium — the molecule — concerned, but the evolutionary process.

Evolutionary Preservation and Capture

Evolutionary science can usefully contribute to digital preservation in a number of ways (John 2005). An aspect of natural selection that is often overlooked is that it reflects the need for diversity in solutions, in strategies. One finds in nature phenomenal amounts of variation; variation that continues to exist generation after generation. It exists because of the inherent unpredictability of nature. It is a recognition — an admission — that the future cannot be predicted. It reflects the existence of multiple strategies: diversity in the face of unpredictability.

It might seem counterintuitive to adopt an evolutionary perspective when striving to preserve something forever (the mission of the British Library's Digital Library Storage system). Quite understandably people tend to marvel at nature's capacity for change but the biological world is also capable of supreme constancy and conservatism. There is information in DNA that has remained the same not merely for thousands or hundreds of thousands of years, but for millions and hundreds of millions of years. It is a phenomenon that deserves the greatest respect of any digital preservation specialist. It confirms the feasibility of deep digital preservation but also points to the need for a humility that seeks more than one best practice, that seeks an evolving strategy incorporating dynamically diverse options.

Conversely, it can be expected that as fundamental advances in digital preservation emerge, it will have many contributions to make to understanding in evolutionary science.

Automation, Information, Personal Archives

Turning to the more directly practical, many powerful techniques of bioinformatics and phylogenetics have been developed where information science meets genome and genetic science. An illuminating example of adapting an existing approach in another field is to be found in the use of phylogenetic algorithms by manuscript scholars wanting to establish or corroborate the ancestry of surviving manuscripts (eg Barbrook et al. 1998, Spencer et al. 2004). These and other bioinformatic techniques will undoubtedly play an important role in authenticating digital files including eMSS; and indeed in forensic analysis of digital files more broadly (eg Goldberg et al. 1998, Carrera and Erdelyi 2004).

There are, however, other aspects of genomic technologies that could be useful in the context of personal archives. The emergence of high throughput gene sequencing capabilities for example has resulted in the production of vast volumes of information, which in turn have led to a demand for automated or supervised computer extraction and interpretation of pertinent information. As a result dedicated gene and genome databases have been established for remote analysis using GRID and

eSCIENCE technologies; but it is not just the protein and gene databases that are available for analysis. There is a burgeoning literature reporting findings from genetic analyses, and in part due to its size this literature too is subject to computational text analysis in its own right (Müller, Kenny, and Sternberg 2004, Raychaudhuri 2006).

Compared with the information in a database, the information in the academic literature (even peer reviewed) is barely structured. Significant advances have been made in the application of natural language processing (Manning and Schutze 1999). One instance is the natural language processing system GENIES which automatically identifies and extracts biomolecular pathways, and forms a key part of GeneWays a technology that processes many thousands of scientific papers and automatically produces a database that is able to identify and visualise molecular relationships and interactions in response to queries from a researcher (Friedman et al. 2001, Krauthammer et al. 2002).

Ontologies will play an important role in testing and training algorithms that provide automated functionality including through supervised machine learning. The expert and ongoing annotation of entities necessary for high quality function coding means, nonetheless, that automation is ultimately going to be necessary to make use of large scale research resources (Raychaudhuri 2006).

The ability to identify the names of genes in scientific literature is not trivial due to inconsistency and nonstandardisation. The most successful algorithms often combine different techniques for classifying documents: descriptive text, nearest neighbour, naive Bayes, maximum entropy and multivariate statistics (Raychaudhuri 2006). The open source software for the multifactor dimensionality reduction technique promoted by the Computational Genetics Laboratory at Dartmouth College, New Hampshire, USA, and used for analysing genes (Moore et al. 2006) has potential for being adapted to pattern search in text.

At first these kinds of technologies will serve less as a means of producing the definitive index, catalogue or ontology, and more as a means of providing pointers, suggestions and indicators for the examiner to confirm independently.

One of the most difficult curatorial challenges of personal digital archives is the need to check for confidentiality and data protection requirements, for copyright issues, for authenticity and provenance concerning all files. Software that was able to automatically search and identify these issues relating to digital rights would be beneficial. It might provide a first stage examination, highlighting likely issues and making suggestions to curators, complementing and strengthening the existing forensic use of GREP searching for example.

It is possible to anticipate in the not too distant future an ability to identify patterns that enable the eMSS to be provisionally classified according to key phases of a person's life: associated with childhood stages (eg starting school), coming of age, initiation rites, process of a job application, a resignation, a promotion, communications leading to weddings or partnership, professional collaborations, retirement, reminiscence and reflection, births and deaths, memories and remembrance, and so on.

Conclusions

The overall approach of the Digital Manuscripts Project has been in some sense an evolutionary one that allows for flexibility and diversity. It is essential, for example, not to rely on any single technology for digital capture. The adopting and adapting of existing technologies is likewise part and parcel of this approach.

There are a number of existing and evolving technologies that are proving to be useful in the digital curation of eMSS. Software and hardware from the forensic, ancestral computer and bioinformatic communities are evidently useful directly as tools and as sources of ideas and inspiration for digital curators and preservation specialists.

While these existing technologies are providing an urgently needed means of making progress with digital capture, this diminishes neither the need for detailed and extensive testing and certification of processes, nor the value of novel developments that are carefully focussed, practical and timely. A case in point is the exciting open source software developed by the Dioscuri Project, and capable of "emulating a modern x86 computer environment, while remaining durable and easy to configure" (van der Hoeven, Lohman and Verdegem 2007), work which has been supported by the pan-European PLANETS project since July 2007.

Acknowledgements

I am very grateful to Jamie Andrews, Stephen Bury, Katrina Dean, Frances Harris, Oliver Urquart Irvine, Scot McKendrick, Ronald Milne, Christiane Ohland, Richard Ranft, John Rhatigan, Elfrida Roberts, Matthew Shaw, John Tuck, Richard Wakeford and Lynn Young for their longstanding and essential support. Arwel Jones, Susan Thomas and Dave Thompson have been reliable and instrumental sources of enjoyable and comradely conversation and discussion.

Although the Digital Manuscripts Project is supported directly by the British Library, the writing of this paper was enriched and made possible by the Digital Lives Research Project, which is funded by the Arts and Humanities Research Council, Grant Number BLRC 8669, and is being led by the British Library. Special thanks to all Digital Lives team members including Andrew Charlesworth, Alison Hill, David Nicholas, Rob Perks, Ian Rowlands and Pete Williams, and most particularly to digital preservation experts Neil Beagrie, Peter Bright, Rory McLeod and Paul Wheatley; and, in addition, Adam Farquhar and Clifford Lynch.

References

ACPO. 2003. *Good Practice Guide for Computer Based Electronic Evidence*. National Hi-Tech Crime Unit: Association of Chief Police Officers.

Barbrook, A. C., Howe, C. J., Blake, N., and Robinson, P. 1998. The phylogeny of *The Canterbury Tales*. *Nature* 394: 839-840.

Bancroft, C., Bowler, T., Bloom, B., and Clelland, C. T. 2001. Long-term storage of information in DNA. *Science* 293: 1763-1765.

Born, G. 1995. *The File Formats Handbook*. London: International Thomson Computer Press.

Carrera, E., and Erdelyi, G. 2004. Digital genome mapping - advanced binary malware analysis. *Virus Bulletin Conference* September 2004: 187-197.

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

Casey, E. ed. 2001. *Handbook of Computer Crime Investigation. Forensic Tools and Technology*. London: Academic Press.

Crowley, P. 2007. *CD and DVD Forensics*. Rockland, MA: Syngress Publishing.

Davis, P., and Wallace, M. 1997. *Windows Undocumented File Formats. Working Inside 16- and 32-Bit Windows*. Lawrence, KS: R & D Books.

Friedman, C., Kra P., Yu, H., Krauthammer, M., and Rzhetsky, A. 2001. GENIES: A Natural Language Processing System for the Extraction of Molecular Pathways from Journal Articles. *Bioinformatics* 17 Suppl. 1: S74-S82. .

Goldberg, L. A., Goldberg, P. W., Phillips, C. A., and Sorkin, G. B. 1998. Constructing computer virus phylogenies. *Journal of Algorithms* 26: 188-208.

John, J. L. 2005. *Because topics often fade: letters, essays, notes, digital manuscripts and other unpublished works*. In *Narrow Roads of Gene Land. Volume 3. Last Words*, ed. M. Ridley, 399-422. Oxford: Oxford University Press.

Krauthammer, M., Kra, P., Iossifov, I., Gomez, S. M., Hripcsak, G., Hatzivassiloglou, V., Friedman, C., and Rzhetsky, A. 2002. Of truth and pathways: chasing bits of information through myriads of articles. *Bioinformatics* 18 Suppl. 1: S249-S257.

Kruse, W. G., II, and Heiser, J. G. 2001. *Computer Forensics. Incident Response Essentials*. Boston: Addison-Wesley.

Kussmann, R. 1990. *PC File Formats & Conversions. Essential Guide to Transferring Data between PC Applications*. Grand Rapids, MI: Abacus.

Manning, C. M., and Schutze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, MA: The MIT Press.

Moore, J. H., Gilbert, J. C., Tsai, C. T., Chiang, F. T., Holden, T., Barney, N., and White, B. C. 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology* 241: 252-261.

Müller, H.-M., Kenny, E. E., and Sternberg, P. W. 2004. Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology* 2 (11): e309.

Nadeau, M. 2002. *Collectible Microcomputers. A Schiffer Book for Collectors with Price Guide*. Atglen: Schiffer Publishing.

NIJ. 2004. *Forensic Examination of Digital Evidence: A Guide for Law Enforcement*. NIJ Special Report. U. S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Raychaudhuri, S. 2006. *Computational Text Analysis for Functional Genomics and Bioinformatics*. Oxford: Oxford University Press.

Sammes, T., and Jenkinson, B. 2007. *Forensic Computing. A Practioner's Guide*. Second Edition. London: Springer-Verlag.

Spencer, M., Davidsion, E. A., Barbrook, A. C., and Howe, C. J. 2004. Phylogenetics of artificial manuscripts. *Journal of Theoretical Biology* 227: 503-511.

Summers, A., and John, J. L. 2001. The W. D. Hamilton Archive at the British Library. *Ethology, Ecology & Evolution* 13: 373-384.

Swan, T. 1993. *Inside Windows File Formats*. Reading, MA: Addison-Wesley.

van der Hoeven, J., Lohman, B., and Verdegem, R. 2007. Emulation for digital preservation in practice: the results. *International Journal of Digital Curation* 2(2): 123-132.

Walden, J. 1986. *File Formats for Popular PC Software. A Programmer's Reference*. New York: John Wiley & Sons.

Walden, J. 1987. *More File Formats for Popular PC Software. A Programmer's Reference*. New York: John Wiley & Sons.

Wong, P. C., Wong, K. K., and Foote, H. 2003. Organic data memory using the DNA approach. *Communications of the ACM* 46: 95-98.

Woodyard, D. 2001. *Data recovery and providing access to digital manuscripts*. Information Online Conference 2001, Sydney.