

Digital Forensics for Preservation

Workshop Report

Rachel Beagrie

The Oxford Centre- 28th June 2011

Digital forensics lie at the intersection of many of the core challenges of digital collections management, especially for those collecting institutions that deal in the papers and correspondence of personal and public life. How do we cope with the growing scale and complexity? How do forensics relate to more familiar concepts like cataloguing and characterisation? How can we make our workflows more efficient and our collections more manageable? What tools do we need for discovery and what are the limits of reasonable deployment? What advice should we give to depositors and what restrictions might we put on users?

This DPC briefing day provided a forum for members to review and debate the latest developments in the use of digital forensics for preservation. Based on commentary and case studies from leaders in the field, participants were presented with emerging policies, tools and technologies and were encouraged to propose and debate new directions for research.

PRESENTATIONS

The Nature of the Problem (Jeremy Leighton John, British Library)

Preservation issues are not new, however the nature of the problem has changed. Many cases now involve information from individuals who are still alive- a change from the past. Protecting privacy has been a longstanding issue and the variety of media used to store information is constantly expanding.

We should be able to capture information without changing it, demonstrate that it has not been changed and continue to analyse it without changing it. We're not only interested in text but also the style and layout of a document, how the ideas progress and the techniques used. Forensics involves using traces to work out what happened in the past and defend it in

a public way (this could be a court of law or a scholarly publication). Once documents are acquired an inventory must be made and a hash library to identify software and keep for future use. We have to look out for forgeries and fakes. This is more difficult with solitary files as if you have an archive it can be used to help back up a document. It is best to use a multi-evidential approach (evidence from a wide variety of sources). It is important for systems to have some flexibility built into them as we do not know what they may encounter, as well as being capable of quickly adding and breaking metadata down into smaller units. Related files can be detected using fuzzy hashes. There must also be clear practises for dealing with privacy issues.

The term forensic analysis has become unpopular, which may indicate the need for a wider meaning. It has become linked to privacy issues and is generally thought to be directed at 'finding the smoking gun' when in fact it is more like story telling- details and evidence are important.

E-Discovery and Sense Making: Tools Techniques and Processes (Simon Attfield, Middlesex University)

How do we make sense of large amounts of information? E-discovery is the process in which e-data is sought, located, secured and searched. In a legal context the goal is to make sense of facts relevant to a case based on documents recovered. This process was studied by interviews with lawyers, observing their pattern of behaviour in a criminal investigation. The quantity of information received is constantly increasing, creating serious problems for the legal industry. It is necessary to narrow the field to a reasonable set of documents and there is pressure on representations to get a sense of what is going on in order to develop more specific questions for the investigation. The problem is that you have to define what you are doing as you go along, although must understand what you are looking for before you can begin.

Each line of enquiry has a lower line and each line may throw up information for others. Communication can be vertical (to senior members) or lateral. The idea of social translucence can enter here- making participants and their activities visible to each other.

Cues can be used to coordinate activities. Computer systems are often opaque and we must programme translucence into them. Clustered representations of documents are helpful as when there is linear representation it is more difficult to understand relevance. Clustering brings similar documents together so it is possible to coordinate selection of interesting documents. One goal should be to look at a representation and instantly be able to understand the underlying social reality as well as visualise the information.

Mobile Forensics: A Case Study (Brad Glisson HATII, University Of Glasgow)

What does your mobile phone reveal about you? This is an area of research becoming pervasive in digital forensics. Mobile phone subscriptions are on the rise and people may have more than one. The smart phone market share is increasing. Over 30 million devices were recycled in 2009 in Western Europe and many are thought to go into secondary markets.

This study investigated data retained on resold mobile phones to find what information is left and how consistent forensic software applications are in recovering it. Location was limited to the UK, models divided into price ranges and 3 toolkits used. 49 devices were used, 44 from eBay and 5 from a local pawn shop. In all 11,135 artefacts were recovered, which included personal identifiable information, account numbers, sort codes and sensitive information. It was possible to tie information together to profile people and some data could be recovered despite deletion. More information could be collected by a manual extraction which shows some issues with reliability of toolkits. As mostly old handsets were used this was not fully representative of the move to smart phones but in that case it is often also possible to deal with phone backups on computers.

The Stanford Forensics Lab: A Case Study (Michael Olson, Stanford University)

Collection acquisitions come in many forms and can have interesting privacy concerns- for example in emails. Other forms include floppy disks, USBs and even iPods and a way is needed to preserve these. Lots of computers no longer have floppy disk drives and we must be able to process older technologies even if the hardware is obsolete.

There is variance in capture failure statistics. Multiple attempts of capture using different drives, hardware and software give different results. We cannot guarantee there is not more information left on disks and it takes large amounts of time to make good attempts to remove information. An example is the Project Xanadu hard disk drives from the late eighties, where data could only be accessed on 2 out of 6 disks.

One end goal of the AIMS Born-Digital Collections project (<http://born-digital-archives.blogspot.com/p/links-etc.html>) is to have collections available via a software application (Hypatia, an application for arranging, describing, and delivering born digital archival content). It is important to separate capture from analysis by building separate capture units. In the future manuals could be created for people to process information themselves.

Trends and Tools 1 (Gareth Knight, CERch, King's College London)

Processing of the great number of files stored on digital media is largely done manually with little documentation. The FIDO project (Forensic Investigation of Digital Objects <http://fido.cerch.kcl.ac.uk/>) looks specifically at open source and free software to capture and curate archival digital records. Broad issues to consider are the working environment (appropriate hardware and software must be used), who is performing the work (what knowledge and training do they require?) and how to communicate intent (as this has ethical issues).

Data acquisition is commonly achieved through creation of a disk image or clone. An appropriate format must be chosen for data imaging- FIDO built on file formats assessment criteria for choosing disk formats to investigate criteria such as the level of analysis supported and ability to embed metadata. Acquisition tools depend on the booting device used, for example from floppy disk or from CD. Processes depend on the type of data to be captured and the level of analysis; whether active or inactive data. This must be specified in the donor agreement to avoid ethical concerns. Hashsets can then be used to identify origin and purpose of files. Data carving from larger data files can allow extraction from damaged files. There are a number of methods with varying levels of success but false negatives and

positives can be produced and different tools are able to retrieve different numbers of files. Future challenges include dealing with multiuser systems, archiving data on third party systems and dealing with diverse devices and media types.

Trends and Tools 2 (Kam Woods, University of North Carolina at Chapel Hill)

Retaining forensically packaged images can reduce loss but it is still necessary to analyse them. Archived disk images are subject to legal arrangements and software shouldn't add to expensive and challenging decisions. Open source formats allow simple extraction and can point to hotspots in the data to give context. They can also process data directly and find private and sensitive information. These programs can be run as the drive is being imaged so it has been indexed by the time imaging has finished.

The immediate goal of the tool shown (BitCurator) is to reprocess data and produce a one page report which an untrained person could use. Another aim is that tools could be used to allow fast and accurate triage on data- informing the user when a disk has damage. There are many reasons you may want to stick with existing tools and workflow, however the main advantage to these tools is when you receive large amounts of data and don't know what they contain or have untrained staff.

PANEL SESSION AND DISCUSSION

- Question: What of the things you've heard today have specific implications on what you're doing?

It is incredibly valuable to be able to do very quick processing as data streams through.

There is a question of what constitutes an appropriate level of digging down into a binary image that has been brought into an archive. What if there is malware on the image and you are ingesting something that can harm users? Can the tools seen today give protection?

We can identify known good and bad files but not the unknown. Hash tags would not tag these due to antifoensics practices. We can never offer a 100% guarantee that something is clean though you are less subject to malware if you are just analysing data rather than

mounting it. We should be overly cautious and anything in doubt should not be included in an image sent out to users. To summarise we can never be sure.

Attacks could potentially become live again- you can no longer recognise viruses for Windows 95 with virus checkers. There is a point where you can't trust your own repository or assume the process is perfect so must put an element of protection on display processes. At the end of the day we can only write that we have taken reasonable precautions- we can't stop a disaster but can reduce its probability.

- Question: Does anyone have practical experience of exotic file formats/devices of anticipation of them coming their way?

Phones are very hard to deal with.

Will the next generation see a new proliferation of formats? It is dependent on the devices used. User generated sound content comes in 5 different file types. Is there more commonality now than 5 years ago? Fewer specialist tools are needed making it easier to get at data. However android customisable content could create a problem.

Is this a question of designing for the exceptional case? This could be seen as a policy decision- is taking a mobile phone worth the investment? Can the problem be eliminated by just saying no? New phones can be backed up so many can be accessed from a computer, which is an easier process.

- Question: Observation on digital forensics and record management

Historically we are not just interested in the document but in where it is made and details such as the pen used. Is information being thrown away with digital documents? Information was being thrown away before, what we have now has been kept by chance. Conversely are we in danger of over-keeping things? Just because we can doesn't mean we should. It is often difficult to determine the boundaries of a record- the problem is that you don't know in advance what's going to be useful/what constitutes value. If we preserve the bits better technology will come along to analyse them.

Instead of worrying about the process of capturing data should we use forensics? Essentially this is spying. Would it give us a richer history or too much information?

The unit of focus is always the single file. The problem is that mental models and metadata models build around the concept of a document; it is computationally difficult to be able to operate on metadata consistently all the way up and down the chain. It is unfeasible to archive everything you are given so there has to be a half way house. Only so much content can be put in.

We are playing catch up with the industry as there is no economic drive for them to store the data themselves.

- Question: Has this encouraged or discouraged you to engage with the tools?

Forensics tools are constantly moving forwards- the implication of this are that we should just make images and come back to them down the line when the technology has improved. However there are worries about functionality being retired (for example with floppy disks), so we must do some work now. Also if we do not use the tools then where is the incentive to improve them? They improve because people try them and find flaws. Some aspects are worth running now and others we don't need to because they can't give useful information immediately. We must keep bits safe so if something else proximate comes along we can use them. The focus should be on what we need to know now.

In some cases the tools discussed are way above the level needed. Simplistic methods are needed to compare data such as hash libraries. Bigger institutions must focus on making tools available to smaller companies. If tools are good enough then they should be able to scale.

As a final note- is thinking about tools all the time productive?