Home Office

eDiscovery in digital forensic investigations

D Lawton

R Stacey

G Dodd (Metropolitan Police Service)

September 2014

CAST Publication Number 32/14

Contents

1 Summary

Digital information, such as that found on computers, mobile devices and storage media can be relevant in the investigation of a wide variety of crimes including the most serious. A variety of enforcement agencies are responsible for investigating these crimes and securing the prosecution of suspects. However, the widespread uptake of sophisticated mobile devices, coupled with the affordability of storage has resulted in huge growth in the volume of digital information being created and stored.

Conventional law enforcement approaches rely on digital forensic examiners interrogating seized devices and providing their findings to an investigator. Further avenues are then typically identified by the investigator and the examiner re-examines the data in light of this new information. Both the examiner and investigator have relevant skills and knowledge to progress the investigation but they are frequently applied independently in a protracted, to-and-fro process.

Given the proliferation of digital data, it is no surprise that the challenges brought about by large data volumes are also of relevance to other professions. Corporate lawsuits are a natural parallel, where the prohibitive cost of searching and reviewing substantial company archives has driven the need for software tools to be developed. The requirement for such tools has come from both the defence and prosecution as the cost of data review is inevitably a point of contention. The use of tools to facilitate this electronic discovery process has become an accepted approach to managing the cost of lawsuits and the market for these 'eDiscovery' tools was estimated at $1.8 billion in 2014 and is expected to grow to $3.1 billion by 2018[1] (approximately £1.1 billion and £1.9 billion respectively).

The eDiscovery approach would appear to be applicable to the world of digital investigations where digital evidence from a number of devices or systems needs to be sifted, interpreted and acted upon in a rapid manner.

In conjunction with the Metropolitan Police Service (MPS), the Centre for Applied Science and Technology (CAST) recently conducted an assessment of commercial products offering an eDiscovery approach to reviewing large volumes of data. The assessment revealed some interesting differences between the way that eDiscovery tools approach an investigation and the standard workflow in a criminal investigation involving digital evidence.

On the basis of this limited assessment, an ideal tool to support the needs of both the technical and investigative elements of digital investigations does not appear to exist. However, the tools assessed did meet many of the key requirements and could be a significant part of a combined solution. In addition, development of the tools has continued since the assessment and is bringing helpful improvements and new features to the market.

There are examples of large law enforcement and regulatory bodies in the UK using eDiscovery techniques as part of their investigations. There are a wide range of tools available and although

[1] Magic Quadrant for E-Discovery Software, Gartner, 2014.

some will be beyond the budget of all but the largest units, more moderately priced tools, and free tools in some situations, are available to deliver some of the benefits of eDiscovery to a wider audience.

This document serves as an introduction to the area of eDiscovery, a survey of typical functionality available and a guide to options for introducing eDiscovery into wider use in criminal investigations.

2 Introduction

Digital information, such as that found on computers, mobile devices and storage media can be relevant in the investigation of a wide variety of crimes including the most serious. Policing is responsible for investigating these crimes and securing the prosecution of suspects. However, the widespread uptake of sophisticated mobile devices, coupled with the affordability of storage has resulted in huge growth in the volume of digital information being created and stored.

Working practices are struggling to keep pace with this trend. The historical practice of examining every exhibit in detail requires considerable time and, with limited ability to direct more resources at the task, either exhibits have to wait to be examined or the amount of work performed on each exhibit must be rationed.

The extraction of digital information from devices is a technical task that requires appropriate tools, training and experience if it is to be performed correctly. However, the examiner performing the task may not be fully aware of the details of the investigation so their work needs to be steered by information from the investigating team. This division of skills and knowledge between the examiner and investigator can result in an inefficient process, lengthening the investigative process. Giving investigators rapid access to the digital information in a form that they can understand and work with has the potential to significantly enhance an investigation.

CAST believes that there are useful lessons to be learnt from the way the legal profession has dealt with the increasing quantity of electronically stored corporate information. This field is termed, within the legal profession, as electronic discovery or eDiscovery.

In conjunction with the Metropolitan Police Service (MPS), CAST recently conducted an assessment of commercial products offering an eDiscovery approach to reviewing large volumes of data. The assessment revealed some interesting differences between the way that eDiscovery tools approach an investigation and the standard workflow in a criminal investigation involving digital evidence.

This report does not focus on the detailed performance of a few tools but on the ways in which the eDiscovery approach can assist the world of digital forensics and investigations.

Section 3 contains an introduction to the area of eDiscovery including how it developed, examples of its use and efforts to standardise the process in the form of the widely accepted Electronic Discovery Reference Model (EDRM).

The following section provides a brief overview of digital investigations with a focus on digital forensics and then looks at the common ground between this area and eDiscovery.

The general lessons learnt from the assessment with the MPS are summarised in section 5 before examples of ways of integrating eDiscovery techniques into a digital investigation workflow are outlined in section 6.

Section 1 concludes this report and is followed by a brief glossary which explains common terms in their digital investigation or eDiscovery context.

3 Overview of eDiscovery

3.1 What is eDiscovery?

With the world's increasing reliance on digital media and the decreasing cost of data storage, many organisations have accumulated an extensive digital archive, storing far more data than previous paper-based systems. In 2003 it was estimated that 93% of documents are created electronically of which over 70% are never converted into hard copy[2].

Trying to retrieve information from this digital archive in a systematic way, for example in response to a Freedom of Information request or a legal proceeding, can be time-consuming and so a method of finding and reviewing potentially relevant material is required.

This is not a new problem for large organisations. Taking a step back, when companies' records were largely paper-based, responding to a disclosure request would have involved the manual review of large volumes of paper records. Relevant material would be duplicated and then delivered to the party requesting the information. For a large case, it could be more efficient to scan the paper records and deliver them to the requestor electronically. In many cases, it was more efficient to pay for all the possibly relevant documents to be scanned and significant keywords listed, use the keywords to highlight potentially relevant documents, review the potentially relevant ones and then pass on the final set of documents electronically. From these initial steps, electronic discovery has grown to a billion-dollar industry.

Electronic discovery, or eDiscovery as it is termed in this report, is a process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a legal case.

The processes within eDiscovery are relatively straightforward but their application can require substantial technical solutions due in the main to the quantity of data being stored and processed. Functions that are taken for granted within eDiscovery such as keyword searching and distributed review are the basics around which a whole industry has grown, developing and refining eDiscovery tools for both legal and investigative professionals.

3.2 How does it work?

Once potentially relevant sources of information have been collected, the data is extracted, indexed and placed into a database within the chosen eDiscovery tool. Some tools refer to the data in terms of matters, a legal term for discrete causes or claims to be resolved.

The data can be initially reviewed to remove irrelevant documents and so reduce the volume of data. Two typical methods employed by eDiscovery tools to reduce the overall data volume are removing duplicate documents and known files. A side effect of modern electronic communication is the frequency with which documents are duplicated, an email that is 'cc'd' for

[2] Computer Technology Review, Sharon Isaacson, March 2003.

example. This can be of assistance when the original source has been lost; however, such duplication also adds to the amount of data recovered as part of an investigation. eDiscovery tools identify duplicate documents via hashing and some can also detect near-duplicates (such as different drafts of a document). This process is referred to as de-duplication or near de-duplication. Some files are straightforward to hash but collections of files such as email archives can be trickier where vendors use different combinations of email content and metadata to produce a vendor-specific hash value for each email and to identify conversation threads.

The identification of known files is performed by comparison to white lists containing the hashes of common files that will have no relevance to an investigation. This is often referred to as de-NISTing as NIST (National Institute of Standards and Technology) produce the National Software Reference Library (NSRL) of hashes of standard files from operating systems and common applications.

The tools identify duplicates within the data when it is first imported. This has an initial time cost but enables virtually instant de-duplication when the data is interrogated during the subsequent investigation. Note, you would not always want to de-duplicate so the option is normally presented to the user of the tool.

The data then passes through multiple stages of manual review in order to highlight items of interest for further examination. This may be to determine if the item is truly relevant or, as a separate issue to check for privileged material. This sort of material comes from various sources including legal, medical and journalistic. This could arise in an investigation where a lawyer may be advising some clients legitimately but also providing a criminal service to others. Redaction may be used in order to shield the particulars of either individuals or companies who do not form part of the investigation itself but whose details appear in the raw data.

The electronic data under scrutiny in any investigation will invariably contain metadata including information such as time stamps and details of the document's author. In some cases more specific metadata such as geographical location and the make and model of camera used to create a photograph may be present; all information that may provide valuable insight for an investigation or help further reduce the volume of data to be considered.

Another frequently used method of data sifting is keyword searching. Depending upon the tool in use, such searches can range from single word searches through to more complex Boolean logic searches where multiple searches are combined by the use of AND, OR and NOT functions, for example searches for 'Fraud' AND 'UK' NOT 'VAT'. Keyword searches need some consideration as they do not discriminate between words that are spelt the same but have different meanings, for example 'Bow' can be on a present, the front of a ship, used to play a stringed instrument or a district of London amongst other definitions.

As eDiscovery tools have developed, so has the complexity of their search functionality. Concept searching moves on from the multiple meaning problem of keyword searching to try to understand the concept being conveyed rather than a specific set of letters. For example, a keyword search for 'gun' might return both gun and guns whereas a concept search for the same single word could potentially return documents that contained terms such as 'shooter', 'piece' or 'sawn-off'. Predictive coding is a feature of a limited number of eDiscovery tools where documents that have been manually reviewed are automatically analysed to identify key distinguishing features and then the rest of the material is searched to identify similar documents.

A possible feature of advanced search methods is detecting emotion. The advertising industry relies heavily on such algorithms to deduce whether a brand is currently in favour with consumers by trawling online comment. At least one tool is currently available which offers this functionality but it has not been widely adopted yet.

The advanced searching capabilities some of the tools offer can come with a significant processing requirement and have not been widely tested in court yet. As such, they may actually present an extra cost to the process as they consume time or hardware resources and may be challenged.

Throughout the eDiscovery process, the implementation of various searches ultimately reduces the volume of material and so reduces the cost and time of an investigation. The goal of eDiscovery is the production of a concise data set related to a line of enquiry. Extending this to application in a criminal investigation, the aim should be for those documents to not just be relevant but for the process to assist in building a forensically sound case.

3.3 eDiscovery in context

The case of *Zubulake* v. *UBS Warburg* is frequently referenced with regard to eDiscovery and is considered a landmark eDiscovery case in the United States. The case centred on an employee's claim of sex discrimination against their employer. As the case developed, "all documents concerning any communication by or between UBS employees and the plaintiff" were requested. In response UBS produced approximately 100 emails claiming that this was the extent of the data held. However, it was discovered that back-up tapes had not been searched. At this point the case turned from a conventional discrimination dispute into a test of disclosure which established responsibilities on the various parties involved and resulted in one of the highest awards to an individual employee in history.

The court stated that "a party or anticipated party must retain all relevant documents (but not multiple identical copies) in existence at the time the duty to preserve attaches, and any relevant documents created thereafter," and outlined three groups of interested parties who should maintain Electronically Stored Information (ESI).

- Primary players: Those who are likely to have discoverable information that the disclosing party may use to support its claims or defences.
- Assistants to primary players: Those who prepare documents for those individuals that can be readily identified.
- Witnesses: The duty also extends to information that is relevant to the claims or defences of any party, or which is relevant to the subject matter involved in the action.

The jury heard testimony of the missing data and returned a verdict for $29.3 million (£17.6 million), which included $20.2 million (£12.1 million) in disciplinary damages.

An illustration of the scale of damages that can be involved when two corporations go to court and fail to meet their disclosure responsibilities is provided by the case of *Coleman (Parent) Holdings, Inc.* v. *Morgan Stanley & Co*. Multiple errors in Morgan Stanley's attempts to produce email archives resulted in a claim for $2 billion (£1.2 billion) in punitive damages on top of the original claim at the heart of the lawsuit.

The increasing importance of eDiscovery techniques was reinforced in the case of *Anti-Monopoly, Inc.* v. *Hasbro, Inc*. The court concluded that data which had been stored electronically may be subject to discovery and therefore Hasbro were required to produce their material electronically, even though they were already producing it in hard copy.

> *"The law is clear that data in computerized form is discoverable even if paper 'hard copies' of the information have been produced, and that the producing party can be*

> *required to design a computer program to extract the data from its computerized business records, subject to the Court's discretion as to the allocation of the costs of designing such a computer program."*[3]

An example of the scale of paperwork in a case comes from the International Criminal Tribunal case against former Serbian prime minister Slobodan Milošević. The paperwork and documentary evidence for the case amounted to over 1 million documents of which multiple sets were required.[4]

3.4 Electronic Discovery Reference Model

eDiscovery was previously defined as a process where electronic data is sought, located, secured, and searched with the intent of producing it as evidence in a legal case. Attempts to formalise this process have resulted in the Electronic Discovery Reference Model (EDRM)[5] which has been in development since 2005 with many contributors including both users and suppliers of tools. Of note within the EDRM, shown in Figure 1 below, are the diagonal lines that sit below the flow diagram and emphasise that as the process progresses, the volume of data is reduced and its relevance increases.

The EDRM is a conceptual view of the eDiscovery process and is not intended to be a rigid flow diagram. Indeed, adopters of the model may decide not to employ all of the actions indicated or may choose to tackle the events in a different sequence. Furthermore, the model is designed to be iterative, whereby an individual action or event can be repeated any number of times in order to refine the output to the next stage or to incorporate new data.

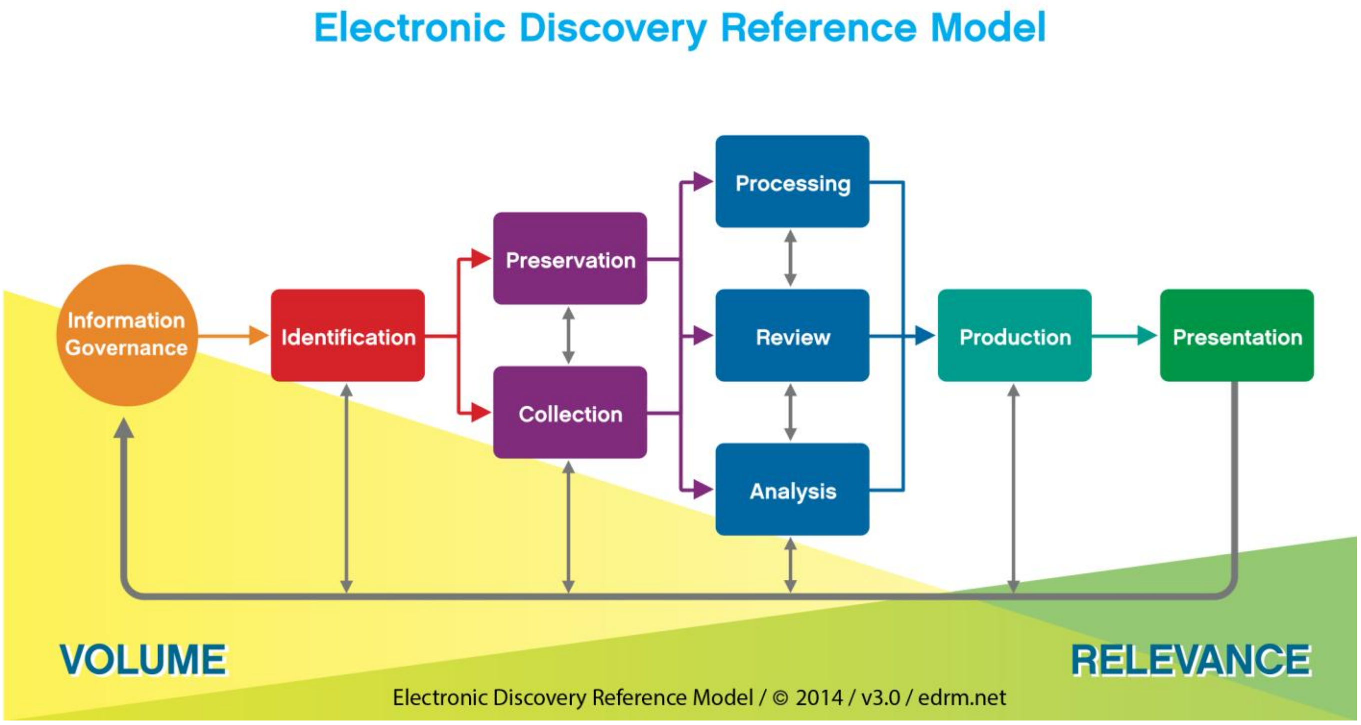

Figure 1: The standard Electronic Discovery Reference Model.

[3] Opinion retrieved from http://cyber.law.harvard.edu/digitaldiscovery/library/process/antimonopoly.html on 19/8/2014.

[4] Milošević trial transcripts, 29 November 2005, p46701.

[5] Used under a Creative Commons Attribution 3.0 Unported License from EDRM (edrm.net).

Expanding on the various headings within Figure 1, the following definitions are designed as starting points for discussion when organisations move towards adopting the EDRM for eDiscovery.

Information governance: Taking steps to ensure an organisation is ready for eDiscovery in order to mitigate risk and expenses, from initial creation of ESI through its final disposition.

Identification: Locating potential sources of ESI and determining the scope, breadth and depth. A potential challenge in this area is the increasing adoption of bring-your-own-device (BYOD) policies which encourage employees to use their personal devices in their business roles to access company files and applications.

Preservation: Ensuring that ESI is protected against inappropriate alteration or destruction.

Collection: Gathering ESI for further use in the eDiscovery process.

Processing: Reducing the volume of ESI and converting it, if necessary, to forms more suitable for review and analysis.

Review: Evaluating ESI for relevance and privilege.

Analysis: Evaluating ESI for content and context, including key patterns, topics, people and discussion.

Production: Delivering ESI to others in appropriate forms and using appropriate delivery mechanisms.

Presentation: Displaying ESI before audiences (at hearings, trials, etc.), especially in native and near-native forms, to elicit further information, validate existing facts or positions, or persuade an audience.

The general concepts behind the EDRM can be translated into a law-enforcement scenario and the overlap will be explored more in the next section of this report. However, considering a scenario such as an abducted person illustrates some of the stages.

The investigating officers visit the victim’s home and discover personal computers, mobile phones, digital cameras and a number of USB sticks and SD cards. Potentially all of the devices mentioned may contain clues as to the whereabouts of the abducted individual. Given that a modern home computer could easily have a 1TB hard drive, phones can hold many GBs of personal data, SD and micro SD cards are typically at least 8 or 16GB and that it is difficult to obtain USB sticks with less than 4GB of storage, the potential volume of data that could be recovered from just one crime scene becomes apparent.

With regard to eDiscovery, the relevant sources of ESI have been ‘identified’. Their removal and imaging with forensic tools would be the ‘collection’ stage. Going beyond the crime scene, surveillance photographs and CCTV might be suitable for inclusion as could records from phone companies.

The following stages of ‘processing’ and ‘reviewing’ are where the investigator starts to see a benefit. The ‘processing’ of the data includes converting the original material into more user-friendly formats so that it can be viewed with standard software such as a web browser or word processor. This provides investigators with simple, rapid access to the information so that they can start to ‘review’ it, using existing intelligence about the case to inform keyword searches on the data. For a large investigation, the eDiscovery tool can be used to create batches of information for multiple officers to review in parallel. Once irrelevant data has been excluded, the investigator can ‘analyse’ the data with tool options such as concept groups and automatic detection of elements such as phone numbers. Key items can be ‘produced’ as briefing material for others involved in the investigation and, when required, details can be exported so that the items can be evidenced in a forensically sound way.

3.5 eDiscovery summary

As a technique, eDiscovery has developed from a process of scanning paper documents into a billion-dollar industry due to its ability to search large volumes of electronically stored corporate data in a cost-effective way which offers companies a proportionate, defensible approach to information requests.

Companies have seen the benefit of adopting eDiscovery techniques, both in response to disclosure requests and before this occurs so that they are in a strong position to respond rapidly. Failure to engage with a request for disclosure can incur heavy penalties but cooperation with the request, and providing information which helps identify other transgressors, can earn a reduced penalty.

The EDRM has been developed as a framework for parties engaging in eDiscovery and describes standard steps in an eDiscovery process which have obvious parallels to the stages in a digital investigation.

4 Digital investigations and eDiscovery

4.1 What is a digital investigation?

In its broadest sense, digital investigations concern the gathering and analysis of any relevant digital data to provide both evidence and intelligence to assist with an investigation.

The field of digital investigation has grown over the past 30 years in both scope and complexity as access to computing resources has broadened and the applications diversified to influence many aspects of modern life. There have been numerous attempts to create definitions with one of the formative statements coming from the Digital Forensics Research Workshop (DFRWS) in 2001 that defined digital forensics as:

> *"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 facilitation or furthering the reconstruction of events found to be criminal …"*[6]

In the early days, the area was known simply as 'computer forensics' as computers were the common source for digital information. As mobile devices and connectivity between devices became more common so 'phone forensics' and 'network forensics' were added to the area. The modern term, 'digital investigations', can include a wide range of data including communications data, cell site analysis and open source intelligence and acknowledges that some of these activities may be aimed at gathering intelligence rather than evidence.

Digital investigations are common across all types of crime, from volume crime where a phone might be examined to determine contacts and messaging, through to the most serious of crimes where all the elements of a digital investigation may be brought to bear on a case.

4.2 Stages in gathering digital information

Returning to the definition from DFRWS, several key stages can be extracted (the influence of this can also be seen in the EDRM in the previous section).

> **Identification:** A key point of an investigation where the potentially relevant sources of information are identified. Without this stage the chance to preserve and collect relevant material can be lost. This stage could also inform other activities including gathering information about possible passwords and attempts to attribute the sources to individuals as ownership of a device or a document can be a point of contention later on.
>
> **Preservation:** This can generally be thought of as removing external influences which might alter the data held on a digital device. At its simplest, this could be removing individuals from the vicinity of their device but could equally be isolating a networked

[6] Digital Forensics Research Workshop. "A Road Map for Digital Forensics Research" 2001. www.dfrws.org.

system to prevent external access. It could also relate to issuing instructions to preserve online accounts and storage. Circumventing encryption can occur in both this stage and the next, if a machine can be preserved in an unlocked state there may be a chance to collect memory or files which will be unavailable or unintelligible once the machine is locked or turned off.

Collection: The process of gathering the data from wherever it resides. The most common collection approach is to create an image of a target device which can then be examined without altering the original exhibit. In a wider sense, this could also apply to aspects such as requesting and receiving communications data. Cloud storage is an increasing concern and whilst the forensic recovery of files stored remotely is possible, the subsequent analysis may require detailed knowledge of the application used. Complications can also arise from the data being held in a different jurisdiction.

Examination: Making sense of the diverse digital data collected. A range of tools and techniques will be used for this in an effort to ensure that as much data as possible is available for review. A lot of this data will be of no relevance to the investigation but it may take considerable effort to get a good understanding of the relevance of material and to present it in an intelligible form.

Analysis: The process of putting the different pieces of evidence together to allow conclusions to be drawn and ideas tested. Some units will have dedicated analytical support available which is a useful resource but many investigators will not have routine access to analysts so it can be helpful for the investigator to be able to conduct their own analysis.

Presentation: The examination and analysis can be conducted at a highly technical level but the information will ultimately need presenting to other individuals, either elsewhere in the investigation or the legal process, who are not so familiar with the detailed processes used and are more concerned with the usefulness of the information provided.

(A common addition to the DFRWS model is to include a ‘preparation’ stage which looks at wider issues such as having appropriately skilled staff and preparing the correct resources before taking any actions but this is less relevant to this document.)

4.3 Common ground with eDiscovery

It should already be obvious that, at least in terms of themes, eDiscovery shadows the key aspects of digital forensics fairly closely.

This report is primarily concerned with the potential of eDiscovery tools to assist with investigations once the information has been collected. However, it is worth looking at all of the stages within eDiscovery to understand the context in which it developed.

It is illustrative of the mindset that accompanies eDiscovery that the definitions for the first two stages of the EDRM (see section 3.4) include the phrases ‘mitigate risk and expense’ and ‘determine the scope, breadth and depth’. These illustrate one of the key points with eDiscovery which is that companies engaging in eDiscovery are attempting to meet their legal obligations whilst keeping a strict control of the costs, i.e. if you take measures to ensure your data is stored in an eDiscovery-friendly way (Information Governance stage) then the cost of future eDiscovery activities is minimised. Similarly, being able to understand the scope, breadth and depth of the material involved helps a company to estimate the costs of reviewing that material and to track progress with reviewing.

Digital forensic stages	EDRM stages
(Preparation)	Information Governance
Identification	Identification
Preservation	Preservation
Collection	Collection
Examination	Processing
	Review
Analysis	Analysis
	Production
Presentation	Presentation

Table 1: Comparison of stages in digital forensics and eDiscovery

For a digital investigation, there is a similar requirement to understand where information may be located and the volume of material present in order to understand how many people may be required to seize, process and review that information. Cost can also be relevant in criminal cases, particularly where the case is complex or requires extensive translation of documents.

The preservation and collection stages have the same intent in both eDiscovery and digital investigations but the manner in which they are carried out can vary significantly. In eDiscovery, a preservation order would normally be issued to inform the owner of an information source that they must preserve the information pending an investigation. Failing to engage appropriately in the discovery process can carry a significantly higher penalty than the actual dispute in question, whereas in a criminal situation the preservation may well require physical intervention as the penalty for the offence being investigated may be more significant than any penalty for interfering with the evidence.

Collection in eDiscovery may be challenging in terms of scale and accurately identifying sources of information but it is generally from compliant computing systems. In a digital investigation there may be various authorisations required to obtain data and the actual extraction of data may be complicated by the variety of devices encountered in a range of physical conditions and the lack of methods to make identical copies of the data without altering the original source.

Following collection, the Examination stage in a digital investigation is carried out by a specialist (forensic examiner, cell site analyst etc.) processing the data and trying to make sense of as much of it as possible. This overlaps with the Processing and Review stages of eDiscovery where the processing is largely automatic and the review is aiming to robustly separate relevant from irrelevant files. Significant techniques in eDiscovery are dividing the information into bundles for review by members of a team and the methods to do the dividing such that the items in a bundle have a common theme. At this point, the intention is not to try and find ‘smoking-gun’ material but to ensure a proportionate effort has been made to find relevant material.

A range of individuals can be involved in the analysis stage. The examiner may be working from information provided by the investigator or the investigator may be using a report from the examiner to identify key items and links. In the eDiscovery process, this could be the final stage of review before a set of documents are provided to the requestor or, similar to the investigator, they could be using eDiscovery tools to investigate a set of documents they have requested.

eDiscovery has an explicit production stage where the relevant material is exported for a third party. Digital investigations also have this stage, for example the forensic examiners report to the investigator or a telecoms company producing call record data but there tend to be many small productions rather than one obvious moment.

Both processes ultimately need to make their results available in a form which can be understood by a broad audience. For digital investigations this may be for an internal briefing or when the case crosses into the prosecution phase. A similar range of audiences are likely for the eDiscovery process. In both cases, there will be the need to document a clear link between the information being presented and the original source although this may need to be more robust in the criminal investigation side.

The key similarities and differences from the above discussion are summarised in Table 2.

Key similarities	Key differences	
	Digital investigation	**eDiscovery**
Working with large volumes of material	Broad range of content	Focus on text content
Trying to make best use of valuable resources	Also interested in how activities were conducted	Interested primarily in matters of record
Using case information to refine search	Key driver is locating evidence	Key driver is managing costs
Need to make material reviewable	Information is normally organised by source	Information is normally organised by owner
Need to share the workload	Relates to criminal proceedings	Primarily used in civil proceedings
Automation used where possible	Must maintain evidential chain	Attribution of documents and devices is not normally an issue

Table 2: Similarities and differences between digital investigations and eDiscovery.

4.4 Digital investigation summary

Digital investigation has become a popular term for modern investigations where digital data is a key component. The approach is very close to original ideas around digital forensics and the stages from digital forensics are equally applicable to some of the extra elements such as communication data. The stages also correspond closely to those in standard eDiscovery work although the emphasis is subtly different. Whilst both start with large volumes of data, eDiscovery seeks to reduce the volume to a manageable amount by using multiple individuals to screen the material and separate the potentially relevant from the irrelevant. This can occur over multiple stages of review and analysis. With a digital investigation, the aim is to find material that will assist the investigation and this is largely conducted by experts who can understand not only the content of the digital files but also the more obscure contextual information.

5 Assessment of eDiscovery options

5.1 Types of eDiscovery tool

Given the need for techniques to cope with large quantities of data within digital investigations, the eDiscovery approach seems a sensible one to explore. The marketplace for eDiscovery tools is competitive with a wide range of offerings at different scales. Some will support the whole eDiscovery process whilst others focus on specific functions or parts of the customer base. The bulk of the tools are designed around the 'corporate' or civil lawsuit situation but there are some which are attempting to bridge the gap to the 'forensic' or criminal investigation world.

Some of the functionality is common across the tools although the depth to which it is implemented varies. A key element is the ability to distribute work to different reviewers whilst retaining oversight of their progress. Distributing the review of files may not be current practice for some investigations but can be very useful for large cases. The tools can usefully put the material into a common reviewing format so the reviewer is not required to launch files in a variety of different programs with varying interfaces. This can also help with investigations of devices which are not familiar to the reviewer as they do not need to work with unfamiliar software and operating systems. This is normally achieved via a web browser interface which opens up the options for the review to take place remotely.

Another common element is the ability for the tool to search and sort the information and feed back results to the user. At its simplest, this could be using a single keyword to filter a set of documents. A more sophisticated tool might offer the ability to highlight a group of documents concerning the same topic and the most advanced options can work with a reviewer to predict which items may be relevant to an enquiry based on the reviewers grading of other items.

The tools also have some standard visualisation features which are normally focused on understanding the flow of emails or showing when items were created.

5.1.1 Forensic model key points

One of the major benefits of the eDiscovery tools which are forensically aware is the ease with which they can be integrated into existing investigative workflows as they work with the same concepts and artefacts that standard forensic tools use, for example, forensic image formats and deleted files. Compatibility with forensic images is particularly helpful given that these are produced as a standard part of the forensic workflow. A forensic eDiscovery tool would be able to read the image file and index the contents (possibly even carving for additional file fragments) without significant user input. Lacking this ability, an intermediate step (e.g. using other software to mount the images or manually extracting the files) would be necessary before the eDiscovery tool could start indexing the information.

However, given that most units will already have at least one specialised tool which handles forensic images as a core part of its functionality, the forensic eDiscovery tool still needs to offer useful functionality for non-forensic staff in order to justify its use. This requires making as much of the information as possible understandable to a reviewer and, where the more complex forensic aspects are being conveyed, doing so in a simple way. For example, presenting carved fragments, deleted files and live files without clear differentiation could be misleading. Similarly, forensic examiners will be familiar with a hex view of a file's content and want this functionality but it may not be suitable to provide that view to a reviewer.

Most of the eDiscovery tools use hashes to identify duplicates and may make use of resources such as NIST’s National Software Reference Library (NSRL) to remove common system and installed files. The forensic eDiscovery tools go beyond this by allowing for custom hash sets to be used which can be invaluable in some areas of investigation.

5.1.2 Corporate model key points

eDiscovery tools were developed to perform document indexing and keyword searching. The corporate tools have taken this aspect and honed it. The focus on keywords has some logical consequences, for example paper documents are of no use to the tool unless they can be scanned and converted so the tools smoothly support Optical Character Recognition (OCR) techniques and are pushing on to audio transcription. The ability to build and save complex searches combining multiple metadata fields and keywords is standard and the tools build on this with options for concept searching and predictive coding approaches.

The corporate eDiscovery tools are able to extract and understand information from a wide range of data repositories (e.g. electronic document management systems, email servers and corporate databases) and standard business productivity tools but can struggle to display the wide range of other files found on personal computers and mobile devices including browser histories and less common video and image formats.

The corporate tools offer more configuration options for distributing work including options to automate the distribution of new material according to saved searches and multiple levels of management. The tools are more compartmentalised than the forensic options allowing the customer to pick which elements of the eDiscovery process they want to implement. Separate applications for each stage allow a high degree of control and auditing but can also be awkward to navigate when rapidly moving between stages, e.g. processing and review.

Performance was not being assessed in the CAST/MPS testing as optimising the hardware for each tool would have been very time-consuming and the results would have had limited read across. However, it was noticeable that there was a higher level of complexity in the ingestion and processing stages with the corporate tools.

The tools also offer a wide range of functions to support the earlier stages of eDiscovery which are outside the scope of this report.

5.1.3 Niche tools

The tools assessed offered a broad coverage of the stages of the eDiscovery process. However, there are many more tools available which will cover some of the elements of the process, often developing sophisticated capabilities but in a narrow field, for example, the review of large numbers of images or processing of obscure file archives. Any process will need niche applications at times to cope with items which fall outside the abilities of the main tool. It is conceivable that a combination of niche tools could be assembled to cover the entire eDiscovery process or at least the elements of interest to an organisation. Given that a lot of the niche tools are less expensive or available as academic and open source projects, this offers a attractive lower-initial-cost option. The drawback of this approach is likely to be the requirement for extensive interventions in the process to move data from tool to tool in a reliable way and the need to become familiar with a multitude of different interfaces and ways of working.

5.2 Digital investigation requirements for eDiscovery

To enable CAST to conduct a meaningful look at the market, it was important to have some key elements to investigate. These elements took the form of requirements which were based on considering a platform which both the forensic examiner and the investigator could work with. This would start at the point of an exhibit having been imaged (hence the consideration is after the point of collection), would allow the forensic examiner to influence how data is processed, incorporate prior knowledge in the form of keywords and hash sets, have a simple but powerful interface for reviewing, incorporate some degree of automated analysis and visualisation and allow both informative and evidential reports to be created.

A draft list of requirements was created at CAST and then reviewed by the project partners to check it covered their specific use cases. Almost 100 requirements were defined with 39 of them being selected as top-level operational requirements. More details on the assessment of the tools is included in Appendix A and a full set of the top-level requirements is given in Appendix B.

5.2.1 Ingestion

Ingestion does not feature explicitly in the eDiscovery process but was an important element to consider for the assessment. There is a distinction between being able to get data into a tool from a wide range of sources preserving associated metadata (Ingestion) and being able to make the information content of that data accessible (Processing).

Requirement	Priority	Rationale
The tool shall be able to open and interpret standard forensic file formats	Essential	The process will be seizure of digital media, imaging of digital media, (potential processing) and use of eDiscovery tools to search the information recovered. As such, the tool must be able to work with forensic file formats.
The tool shall be able to open and interpret standard outputs from tools that process mobile devices	Desirable	Different forensic tools have different capabilities and, generally, mobile devices and smaller embedded electrical items are best dealt with by tools other than the mainstream computer forensic tools.
The tool shall be able to ingest files directly from different file systems	Desirable	It may be useful at times to simply ingest files directly from e.g. a DVD or memory stick.
The tool shall operate with forensic soundness, not altering sources and preserving metadata on extracted items	Essential	Metadata can be crucial in establishing connections and sequences of events and must not be altered by the tool. In addition, it must be possible to still produce the information evidentially. If the tool alters the information it is working on, it will be more difficult to defend.

Table 3 : Ingestion requirements used for the tool assessment.

The most proficient tools in this instance were, unsurprisingly, those with a forensic outlook. The corporate tools used an intermediary product to mount forensic images as accessible drives which could then be crawled. This worked for a wide range of file formats but causes problems working across different operating systems e.g. mounting the image of an HFS+ system via a tool running in Windows leads to a visible but inaccessible drive.

One of the requirements called for the tool to have the ability to ingest data in its native form from standard mobile device tools. Whilst this was generally unsupported, the forensic tools were able to interpret XML and so using the XML export from the phone tools would probably lead to a greater ability to ingest.

The preservation of metadata was generally good across the range of tools and, where checked, hash values pre- and post-ingestion tallied.

Of noticeable difference was the complexity of the ingestion process used by each of the tools. In some instances the process was sufficiently complex that representatives from the tool's supplier had to complete the ingestion. At the other end of the scale were tools where the ingestion procedure had an obvious flow to it. The user could make modifications for improved efficiency but not doing so would produce the same output, just not as rapidly.

Such variance between tools may raise issues with training and the speed with which operators can become fully proficient in operating a new tool.

Overall, there was a clear split between the tools with a more forensic focus and the tools that required intermediaries to ingest data. Using an intermediary product provides reasonable coverage and fits a model where specialised external tools perform dedicated tasks (in a similar way, most tools do not have their own OCR capability but can integrate with external tools). However, one tool showed that it was able to accept forensic images, gain access to elements such as deleted or carved data and understand a wide range of file systems resulting in a smoother and more complete ingestion process, demonstrating the benefit of having these facilities within the tool.

5.2.2 Processing

The processing stage requires the actual extraction of content from files so that it can be indexed and made reviewable by a user. This can involve unpacking archives and crawling databases. It can also involve automatic filtering to remove known irrelevant files (for example, by the use of white lists).

The way in which data is loaded onto some systems and divided up for review can have a significant impact on the computing resources used. Some tools lent themselves to simpler compartmentalisation of the information, which in turn allowed only the data relating to the current case to be worked on. The tools are designed with scalability in mind such that additional memory, processors or servers can be added to a core system to assist with different elements of processing.

Overall results from the processing tests showed varying levels of success with regard to tools identifying abnormal files such as encrypted or archived material. Variable results were also found around concepts such as hashing and deleted files. Standard white lists were commonly implemented and the tools were all able to identify some file types by their signatures rather than extensions, a useful feature for where a file extension has been changed.

The way in which the tools handled structured data, such as databases, was largely limited to simple views of data, without the ability to show links. Tool performance with unstructured data, such as video, audio and scanned documents varied from unable to display through to OCR, audio transcription and picture galleries for video.

Requirement	Priority	Rationale
The tool shall be capable of supporting large operations, scaling to meet the users' needs	Essential	eDiscovery tools are being suggested for large and complex investigations where multiple teams of investigators may be working on the same case.
The tool shall be able to process data having a structured format	Desirable	Investigations may involve data contained within databases, whether internet browsing artefacts in an SQLite database or business applications in Access/Oracle etc.
The tool shall be able to process data having a semi-structured format	Desirable	Modern office documents, web pages and other files are self-describing in various ways and this information should be used by the tool to assist in understanding the data.
The tool shall be able to process unstructured data including OCR and media transcription	Desirable	Data may be present in forms other than straight text so interpretation of graphics, audio, video etc. would be useful.
The tool shall be able to filter data by use of hashes	Desirable	Hashes provide a quick method for filtering out known material and avoid the need to spend time reviewing it.
The tool shall make the contents of containers searchable	Essential	Important information may be inside containers such as archives, compound files or encrypted folders.
The tool shall be scalable and extendable (e.g. via API)	Desirable	Investigations may grow beyond initial expectations and the expectation is that the system will be able to be upgraded easily to match this. In addition, it may be that additional functionality is required which can be implemented by a local expert.

Table 4: Processing requirements used for the tool assessment.

5.2.3 Review

Reviewing starts once the initial processing has completed and the full set of files is made available to the reviewer. The main aim of the review stage is to rapidly filter data down to a relevant set of files and then distribute them to a team of reviewers for more detailed examination.

Tools from the corporate backgrounds tended to perform better in this area as a result of their initial searching capabilities and highly configurable workflows. As would be expected, Microsoft Office files, PDFs and standard picture formats were all supported within the tools with some tools being capable of audio and video playback across a range of formats.

The actual process of using the tools to preview files rapidly was generally good for documents and spreadsheets but some tools lacked the ability to scale pictures or present multiple pictures for review which made them less effective. Identification of location and time data was generally limited to very specific examples e.g. EXIF data in photographs. Duplicate and near duplicates could be readily identified by most tools.

Requirement	Priority	Rationale
The tool shall be able to filter by keyword	Essential	The tool must give the user the ability to search information based on their own knowledge which will include key words.
The tool shall facilitate tagging/categorising items	Essential	A key part of review is being able to assign items to different categories to assist future work.
The tool shall be capable of creating collections of files for review	Essential	Rather than have all the investigators accessing the central data repository for a case, it may be that collections of files are hived off for external review e.g. to CPS.
The tool shall support standard fields for sorting/filtering files	Essential	Standard fields such as file type/size/name, file timestamps, original location etc. are key ways to focus down a review of material.
The tool shall support advanced sorting/filtering methods including regular expressions	Desirable	The tool should be able to support users who can conduct more advanced searches.
The tool shall be capable of performing complex searches such as extended Boolean searches or similarity searches	Desirable	The tool should be able to support more advanced users who can conduct more advanced searches.
The tool shall support different user privileges and workflows	Essential	In a collaborative environment where multiple investigators are working in different ways, it is important that a manager can set up privileges and processes for different types of users.
The tool shall display contents from a wide range of files	Essential	The tool must facilitate the easy review of material so must be able to display the contents.
The tool shall be capable of effectively managing Legal Professional Privilege (LPP) material	Essential	It is always possible that LPP material will be discovered during an investigation due to the bulk ingestion of data. This information must be handled correctly.

Table 5: Review requirements used for the tool assessment.

All the tools supported the entry of individual keywords for searches and some tools could use keyword lists either as one-offs or as persistent lists for use whenever required. All the tools had the ability to create more complex filters via a tailored interface but it was sometimes difficult to obtain access to the desired fields to filter by. Proximity searches (allowing for keywords within a set distance of each other to be found) and Boolean combination of keywords were also common features.

In terms of workflow, all of the tools permitted the creation of various levels of users who could be assigned different privileges and the ability to customise tags or categories. Some tools also allowed for specific files to be assigned to different batches allowing different rules or review options to be associated with them. All the tools had the ability to add comments to individual files and several had the ability to redact or directly highlight sections of a file, for example, a relevant paragraph from an email.

All the tools allowed a senior user to monitor progress and most could be set up to accommodate some method of excluding privileged material. None of them had an option to spawn a self-contained review package to be used by an external party which could be a useful feature. However, the tools all have a web interface available so, with the appropriate security, could support remote viewing of material.

5.2.4 Analysis

Whilst the review stage deals with the initial assessment of material in a case, the analysis stage focuses on understanding the detail within the data and helping the investigator to discover connections between items.

This stage is where each tool had at least one major capability gap in its functionality.

Data visualisation ability, in the context of assisting investigators in identifying patterns in the data, was generally weak across all tools. Functionality for graphically displaying times and dates was mostly limited to analysing email correspondence with a few exceptions. One expectation that was not fulfilled was tools having the ability to display temporal or spatial information in innovative ways.

One area where tools have significant analytical capabilities is text analysis. This ranges from more advanced forms of text searching such as fuzzy searches (finding matches to minor variations of the keyword) and stemming (reducing a keyword to its root term) to the ability to automatically identify entities (such as email addresses, locations, phone numbers etc.) and concepts. Some tools have the ability to be trained to identify relevant documents but this does put a onus on the user teaching the tool over a period of time or the use of concept groupings which can be easily swamped by irrelevant features of the material.

Requirement	Priority	Rationale
The tool shall assist investigators in identifying new lines of enquiry	Essential	Although hard to quantify, an underlying assumption is that providing the investigator with the ability to search information on their own terms should allow them to develop their investigation faster and in more productive areas.
The tool shall display spatial connections	Desirable	Spatial information is easier to process when it is presented graphically to illustrate the locations.
The tool shall display the relationships between items based on temporal information	Desirable	Temporal information can be used to sort information but becomes more powerful when displayed such that aspects such as the time between items is made obvious.
The tool shall support advanced search methods such as thesauri and taxonomies	Desirable	The tool should assist the user in forming better searches by suggesting alternatives and understanding the relationships between concepts.
The tool shall be able to construct timelines from both metadata and file content	Desirable	Temporal information might come from the metadata of a file or from times discovered inside a file. Both may be useful.
The tool shall identify entities and key topics without user input	Desirable	A tool's independent identification of key entities/topics can assist the investigator in identifying new dimensions to an investigation.

Table 6: Analysis requirements used for the tool assessment.

5.2.5 Production

The production stage is entered once the user has completed an area of enquiry and intends to produce records of what has been completed with the relevant files identified. It is important at this stage that the tool can provide a clear and concise audit trail allowing results to be traced back to their source in such a way that the investigator can provide a meaningful commentary on the files identified.

It was possible to incorporate comments, typically by reviewers or analysts, into a report to accompany a selection of files. The majority of tools enabled the user to select types of metadata to include in reporting. This allows for information on the original exhibit and contextual information (such as being an attachment to an email or coming from an archive) to be included.

Requirement	Priority	Rationale
The tool shall provide auditing functionality	Essential	There is a need for a range of information to be gathered to inform management, investigative and oversight activities. E.g. case processing information for the High tech lab or searches conducted for case review.
The tool shall ensure continuity throughout the system (ability to identify source of evidentially relevant data)	Essential	After an investigator has highlighted an artefact as being relevant to the investigation, the digital examiner must be able to link that back to the data from the original device.
The tool shall have understandable evidential and presentational reports	Essential	The tool should produce information about the investigation, both for presentation during the investigation and for submitting as a formal document.

Table 7: Production requirements used for the tool assessment.

With regard to auditing functionality all the tools offered basic logging of user logins and case access. Some tools complemented this simple functionality with logs for each case recording searches, items being tagged etc., and the level of recording was generally sufficient to be able to follow a user's actions.

5.3 Significant gaps in functionality

The major eDiscovery tools are based around searching and analysing text which is often sufficient for their primary market in civil cases. Their principal objective is to assist in producing a set of material which is relevant to a line of enquiry. That does not necessarily include attempting to find the key set of documents or helping the investigator explore relationships between the data in order to build a compelling picture of events which could prove decisive in the outcome of the enquiry.

Criminal investigations consider a wide range of material from sources as diverse as cloud storage, communication records, surveillance material, covert intelligence sources, traditional forensics and digital data from modern devices. It is important to maintain the distinction between intelligence and evidence but there would be advantages to bringing as much of the information together as practical. The crimes investigated may require in-depth examination of the details of pictures, audio, video, documents and location data amongst other types of data. Currently the tools do not provide innovative methods to search the wide range of non-textual data involved in investigations.

The tools will readily deal with a wide range of text material and allow an investigator to filter the material based on keywords. They may show groupings of material based on concepts or machine learning and, if the material is typical email correspondence, the tools can show the flow of messages between individuals and potentially illustrate how connected individuals are to each other. These same ideas could be applied to mobile communication data or instant messaging to immediately improve the range of material handled.

Visualising data in terms of its time or location information was generally poor or not present at all. It could be argued that this is better performed in dedicated tools but given the ease of pushing geographical data into a tool such as Google Earth it seems a missed opportunity. It is possible to add additional toolkits to some of the eDiscovery products which could help with the visualisation but these will come with additional complication to the user and increased overall cost on what are already expensive pieces of software (and supporting hardware).

The tools are able to identify some entities and have the potential to allow the user to define new types of entity that are relevant to their situation. However, they need to support this further by allowing the user to correct or supplement entities whilst they are progressing their investigation in a similar manner to the way users can train and correct the predictive coding to return related documents. It should be possible for a user to highlight a piece of data and associate it with a new entity or a concept such as time or location. Once the connection is made, the tool then needs a range of abstract visualisations to allow the user to display the data in ways they find helpful. Some of the tools from the academic and open source community are leading the way in producing tools which assist with the visual analysis of data.

Whilst the more forensically minded tools provided good compatibility with forensic processes, they struggled to strike an appropriate balance with the user interface. One was very similar to a standard forensic tool which would be off-putting for a non-forensic user and the other was so restricted, in an attempt to simplify the interface for the reviewer, that the scope was severely limited.

The facility for comparing key data between different cases was regularly discussed during the assessment. Although there are good reasons for limiting the scope of an investigation when attempting to resolve legal disputes in the corporate world, the situation is somewhat different in law enforcement units with an ongoing caseload where the same individuals are present in multiple investigations. In this situation, identifying links between apparently unrelated cases can provide new insights into the cases and a deeper understanding of the criminal networks involved. For this reason it would be very helpful if the tools could provide some method to compare key features from one case to another. At a very simple level, this could be a question of identifying duplicate files but more helpful would be if this was at the level of entities. This form of comparison is starting to appear in phone analysis tools but still has to be initiated by the user. Ideally, the collection of entities would occur without intervention and new cases would automatically be compared against past cases to identify linkages. Whilst this may be a step too far removed from the standard scope of an eDiscovery tool, at least having the facility to export key entities, particularly those identified within reviewed, relevant documents, would be a helpful start.

5.4 Assessment summary

CAST, in conjunction with the MPS, conducted an assessment of four commercial products offering an eDiscovery approach to reviewing large volumes of data. The assessment allowed both parties to gain an understanding of what is possible with today’s tools and was based around a set of high-level requirements.

Two of the tools were targeting a corporate eDiscovery process and two were more oriented towards the forensic workflow and digital investigators. CAST and the MPS already had some familiarity with the smaller and open source tools for eDiscovery and visualisation.

Although the commercial products offer a strong capability in text searching and distributing review to multiple groups, they were generally poor at helping the investigator to understand the links between their data and spot new investigative leads. The forensic tools met more of the requirements and would be easier to integrate with standard forensic workflows but still need to work on presenting data appropriately to individuals with differing levels of technical knowledge.

On the basis of this limited assessment, a single tool to support the needs of both the technical and investigative elements of digital investigations does not appear to exist. However, the tools assessed did meet many of the key requirements and could be a significant part of a combined solution. Development of the tools has continued since the assessment and is bringing helpful improvements and new features to the market.

6 eDiscovery in practice

The work performed for this report was based on a desire to investigate the potential benefits of a platform on which both forensic examiners and investigators can easily apply their expertise and knowledge, from the point of an exhibit having been imaged through to evidential reporting,

Previous sections have focused on describing the similarities and differences between traditional eDiscovery and digital investigations and providing an overview of how a sample of tools performed when tested.

This section looks at examples of how eDiscovery can be incorporated into existing investigative practices and the roles and processes that may be necessary to get the best from the combined approach.

6.1 What opportunities does eDiscovery offer?

There are several examples of eDiscovery tools being used in law enforcement and regulatory units in the UK. These range from individuals using an open source tool for specific types of investigation through to entire agencies basing their investigative process around eDiscovery.

Some of these examples are illustrated in the following pages.

The flow charts used to illustrate the processes are fairly simple using only three symbols. Diamonds mark a decision point, skewed rectangles represent inputs and outputs and rectangles indicate a process.

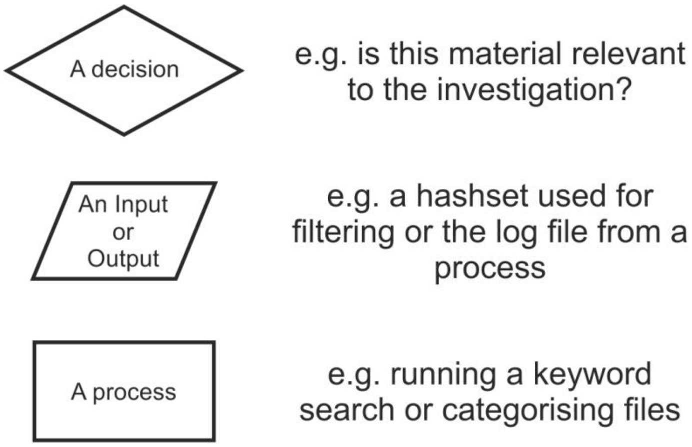
A decision
e.g. is this material relevant to the investigation?
An Input or Output
e.g. a hashset used for filtering or the log file from a process
A process
e.g. running a keyword search or categorising files

Figure 2: Guide to symbols used in diagrams.

Colour coding in the following examples reflects that employed by EDRM for the different stages of eDiscovery.

6.1.1 Investigator aid

Probably the simplest way of incorporating an element of eDiscovery into a normal investigative process is to use it as a way for an investigator to carry out their own interrogation of the data in a case, without the need to go through a complex forensic tool or requesting the work from an examiner or analyst.

The initial stages are standard starting with imaging the devices and beginning the forensic examination with one of the standard forensic or phone tools.

If the investigation only requires easily retrievable material then it may be sufficient to simply extract the live files and push them into the eDiscovery tool. However, there are normally some elements which will require more detailed examination which can be dealt with by a forensic examiner and added to the material for the tool to process.

For computer/media exhibits, the forensic tool would normally generate a keyword index. As this is also a key function that the eDiscovery tool will perform, pushing the material into the eDiscovery tool as soon as possible will allow both tools to index in parallel rather than waiting for the forensic tool to build its index (which the eDiscovery tool will not make use of) before transferring.

With the information indexed in the eDiscovery tool, the investigator can now apply any case knowledge they have in terms of keywords, significant dates or names. Different crime types will lend themselves towards different ways of dividing up the data, for example, fraud cases may focus on corporate email accounts whereas a drugs investigation may make use of dictionaries of slang terms or search for common phone numbers to narrow the focus.

Particularly in cases with multiple exhibits, the use of the eDiscovery tool helps the investigator see the overall picture and spot patterns across devices which may be significantly easier than trying to find relevant information in a large written report from a forensic examiner.

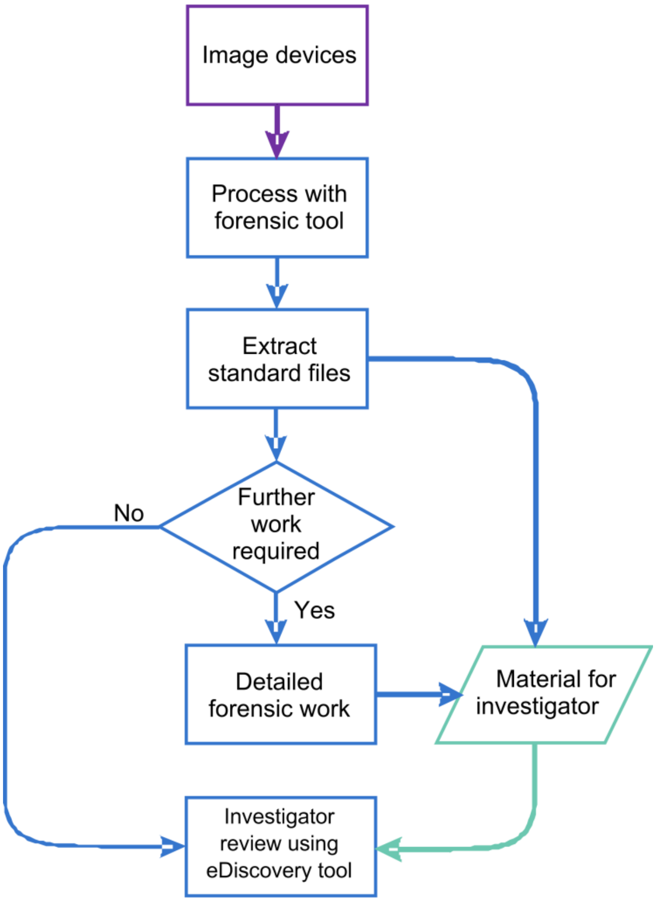
Image devices
Process with forensic tool
Extract standard files
No
Further work required
Yes
Detailed forensic work
Material for investigator
Investigator review using eDiscovery tool

Figure 3: eDiscovery as a review platform for the investigator.

6.1.2 Distributed review

The model from the previous example can be extended to use some of the standard workflow options available within eDiscovery tools to help manage larger cases.

In this example, as before, the initial work is performed with forensic tools to provide a pool of reviewable material for the investigation.

After ingestion into the eDiscovery tool, various methods (expanded on in the next example) can be used to create bundles of material which share a common thread. Examples could include a bundle per individual in a complex fraud case, bundles for different subject areas or bundles of material of various formats (audio/video, pictures, spreadsheets etc.) which need to be reviewed differently.

Rather than having one big team looking at all the material, dividing it into bundles allows appropriate personnel to be assigned to each bundle with only the relevant material from each bundle then being passed on to the lead investigator. This mirrors the way the tools are used by legal teams with junior members narrowing down the initial set of data to a more manageable, relevant set for the lead counsel.

The eDiscovery tool can be used to set up different teams of reviewers. It will also manage the distribution of material to them, the recording of comments and categorisations made by the reviewers, and monitor their overall progress. More advanced tools institute rules to ensure that the review is conducted in a standard manner (e.g. the material must be assigned to one of a set of categories before it can be returned).

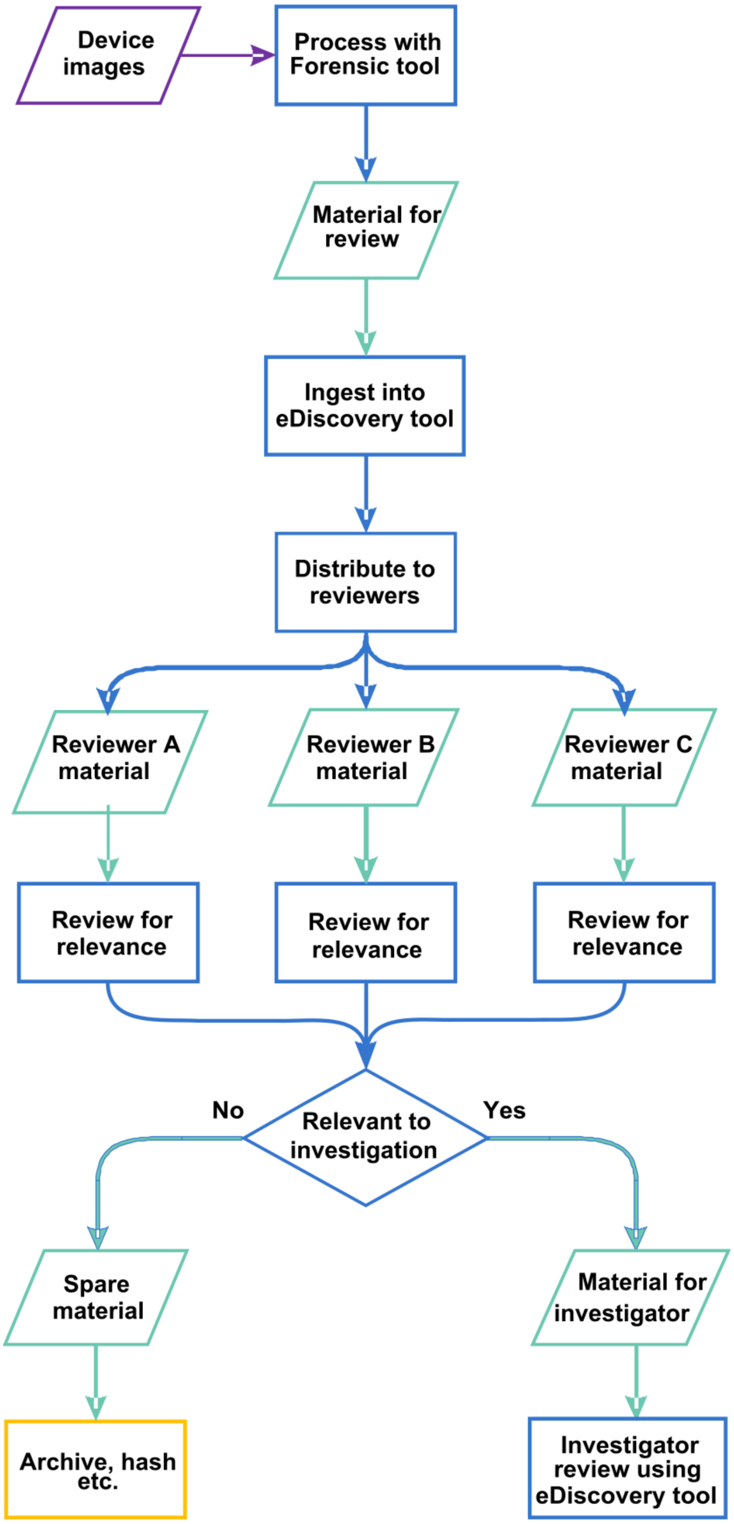

Figure 4: eDiscovery tool being used for management of the review process.

Generally, all the search and visualisation options that the tool offers will be available to the user irrespective of the stage reached in the process.

The lead investigator benefits from receiving a smaller set of data to review which should all be relevant. Analysis by the investigator may then reveal new leads which can be passed back to reviewers to process.

6.1.3 Creating review bundles

As discussed in section 5.2.4, the eDiscovery tools have a range of methods available to assist the user in understanding their data. These are shown, in no particular order, in Figure 5. These methods can be used to help decide how to divide up the material for review.

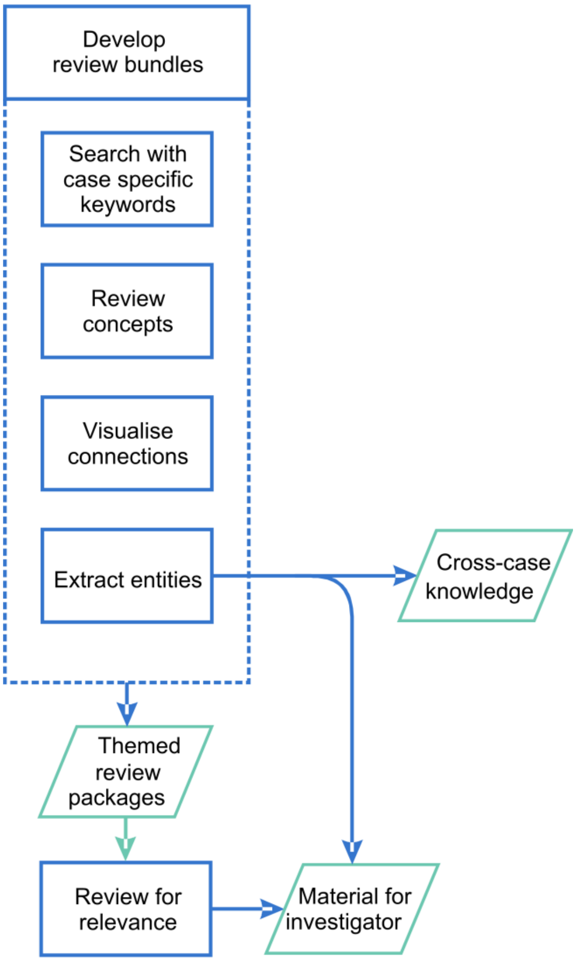

Figure 5: Using eDiscovery features to create themed review bundles.

Keyword searching is already a feature of forensic tools and similar methods are applied in eDiscovery tools. The benefit is in making the keyword searching accessible without the potentially off-putting interface of a forensic tool. The comparison is often made to using an online store or search engine which is designed to be used by anybody.

Some of the tools will attempt to group material by concept, producing bundles of assorted file type but all with a link to a theme or idea. This is normally an automated process with the tool deciding what the important concepts are but there are also examples of tools which allow the user to specify in advance which concepts are of interest (as a more advanced form of a keyword list).

Visualising the patterns in the material, for example, communication connections between individuals, may make some obvious groupings apparent which could then be separated off into different bundles.

The tools also have the ability to automatically identify entities (currencies, internet addresses, named people and places are just a few examples). This is normally achieved by either comparison to an in-built keyword list (useful for detecting entities such as cities or countries) or looking for a specific pattern, for example four groups of four numbers each, which could correspond to credit card numbers or text either side of the @ symbol. All material with potential credit card numbers in could be passed to one team to review whilst obvious emails could be supplemented with the material containing @s and given to a team reviewing communications.

A potential development for the tools would be to make these identified entities available as a store outside of the specific case being worked on so that they could be compared against other cases.

6.1.4 End-to-end eDiscovery

The idea of using an eDiscovery tool from the beginning to the end of an investigation may not be entirely practical but it is not so far away as to be ignorable. This example is close to realising an end-to-end approach whilst still incorporating the key role of forensic images in the criminal process. Having an auditable process will be vital in defending the approach taken.

It is difficult to imagine removing the need to capture a forensic copy of the relevant devices and information but the forensic eDiscovery tools are capable of ingesting this directly rather than taking the extracted files from a forensic tool. There will be artefacts that fall outside of the scope of the eDiscovery tool so these will still require manual intervention from an expert before including in the material to be investigated.

Skipping the detail of reviewing and analysing the material, the eDiscovery tool will log the actions taken by the user, recording key aspects such as searches run and the assigning of categories. The combination of this log with the logs from the original imaging tool, the log concerning the ingestion and the log from the forensic tool used to deal with exceptions raised during the ingestion provide a good audit trail from the original exhibits through to the selection of relevant material.

A forensic tool could then be used to evidence the material or it could be produced from the eDiscovery tool depending on the requirements of the case. It might be that the eDiscovery tool output offers a clearer explanation of the relevance of the material whilst the forensic tool might provide a better reproduction of key metadata.

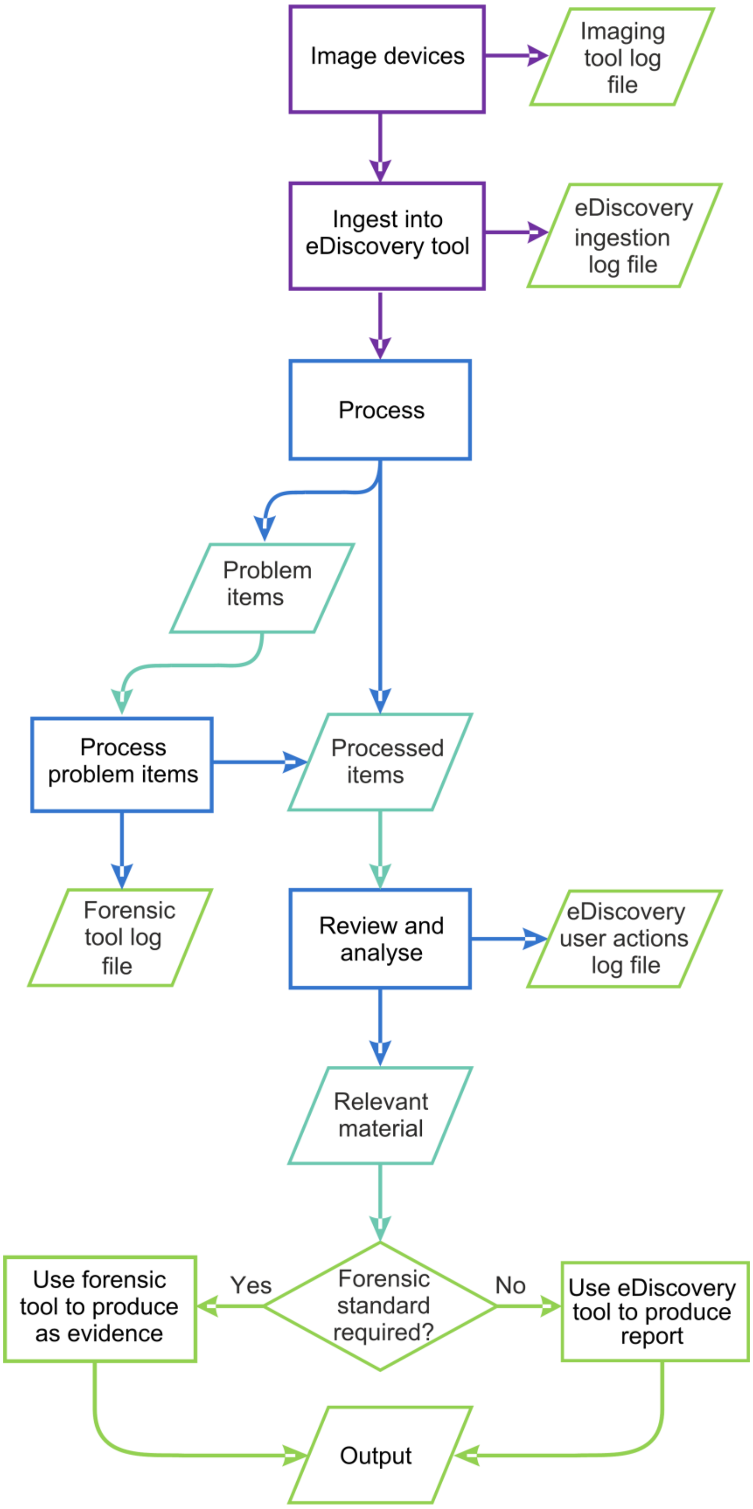
Image devices
Imaging tool log file
Ingest into eDiscovery tool
eDiscovery ingestion log file
Process
Problem items
Process problem items
Processed items
Forensic tool log file
Review and analyse
eDiscovery user actions log file
Relevant material
Use forensic tool to produce as evidence
Yes
Forensic standard required?
No
Use eDiscovery tool to produce report
Output

Figure 6: Near end-to-end eDiscovery.

6.2 Implications of adopting eDiscovery

Some units may already be following a process which is very similar to the examples above, for example using a forensic tool to allow investigators to review the content of forensic images, without thinking about it as being eDiscovery. Other units may have identified problem areas around the scale of investigations and getting data to investigators in a timely fashion but not yet taken steps to resolve them. Adopting an approach closer to eDiscovery will therefore present a greater or lesser challenge but there are some common elements which will need to be considered.

6.2.1 Single provider or multiple elements

It is possible to imagine two models for fulfilling the software requirements of the process: a single tool which is closely aligned to the steps in an investigation or a series of tools which excel at their individual functions but require integrating. The forensically focused tools are nearer to the single provider model whereas the tools with a corporate background already represent something of a modular approach.

A single provider solution has an appealing simplicity and should allow data to flow through the system with minimal management. This frees up time for the forensic examiner and also makes it easy for the investigative team to review packages and conduct their own searches. The downsides to this approach are not necessarily getting the best overall functionality and being tied to a single supplier.

A modular system is likely to involve additional steps to achieve the same functionality as a single provider's system. Forensic examiners may need to extract collections of files from forensic images and apply hash sets before ingestion. They may need to export collections of files from the main repository to assign to investigators and incorporating the findings may be more difficult. It may also make it more difficult to link back to original exhibits and understand how an investigation reached its conclusions. However, the advantage this approach has is being able to pick the most appropriate tool for each stage, in the same way that the tools already use external OCR engines rather than develop their own. The abilities of the tested tools in areas such as visualisation were generally poor but in a modular system this could be handed on to one of the niche packages which specialises in this area. The adoption of standards, such as the EDRM XML data format, can, if implemented in the tools, solve many of the issues around ensuring data can flow smoothly and also provide compatibility when changing supplier at the end of a contract.

6.2.2 Changing processes

Many units will be working to standard processes that determine what steps will be undertaken during an investigation. There will also be working relationships that have developed between the forensic staff and the investigators. Both of these elements may need adjusting to accommodate changes in the way of working and that can often be a source of tension. Some staff may need convincing of the need to change and others may see their role as being undermined or overburdened. The forensic examiner is discussed below but investigators may also have concerns. They may want the security of having a forensic examiner performing the searches or simply be concerned about having to use computers for analysis. The key advantage for the investigator is being able to get access to data about their investigation at a much earlier point by removing some of the waiting period whilst exhibits are with the forensic team. Features such as distributed reviewing may be helpful to some investigators and the potential for remote reviewing (with appropriate safeguards) could put the information in the hands of the investigator no matter where they are based.

6.2.3 Disclosure considerations

There may be concerns about complying with relevant legislation such as the Criminal Procedure and Investigation Act (CPIA). CPIA requires all material collected in an investigation to be assessed, graded, tested for relevance and either listed or disclosed in full to defence teams. The Attorney General’s Office has issued guidance[7] on the principles of disclosure with a section dedicated to digital material. The objective of the guidance on digital material is:

> *“to set out how material satisfying the tests for disclosure can best be identified and disclosed to the defence without imposing unrealistic or disproportionate demands on the investigator and prosecutor”*

The guidance reinforces the ACPO principles[8] around handling digital evidence and stresses that:

> *“It is not the duty of the prosecution to comb through all the material in its possession - e.g. every word or byte of computer material - on the look out for anything which might conceivably or speculatively assist the defence*.”

Sections A41 to A43 of the guidance outline different approaches including manually reviewing material but also the use of sampling, key words or other “appropriate search tools or analytical techniques”. For large cases, it suggests that:

> “*it will usually be appropriate to provide the accused and his or her legal representative with a copy of reasonable search terms used, or to be used, and invite them to suggest any further reasonable search terms.“*

The guidance also considers what records should be kept including:

- logs of all material seized or imaged,
- the search/disclosure strategy and techniques employed,
- the searches carried out including who conducted them and when,
- how the search strategy developed as material was reviewed and
- if material was highlighted by a search but not examined, why that decision was taken.

In many ways, the guidance outlines an approach which is in keeping with eDiscovery. Whilst the tools may not cover all aspects of the disclosure regime, they can be of assistance in both the proportionate searching and the documentation aspects. A disclosure or case officer could be given access to the tool in the same way as an investigator and could see the log files concerning material ingested, searches conducted and comments from reviewers, all of which would be of assistance in fulfilling their responsibilities.

[7] Attorney General’s Guidelines on Disclosure, December 2013.

[8] ACPO Good Practice Guide for Digital Evidence, v 5.0, 2012.

6.2.4 The role of the forensic examiner

Implementation of an eDiscovery process does not remove the need for skilled forensic examiners. If implemented in a considered way, the examiner retains their role in ensuring the forensic soundness of the evidence, uses software tools to help refine the material for the investigation, and tackles the more challenging artefacts which are periodically going to fall outside of the scope of the eDiscovery process.

The diagram here expands on the involvement of the forensic examiner in the earlier examples and shows more detail on the processing which can be conducted within the eDiscovery tool. The examiner is heavily involved in the initial imaging of devices, ingesting the data into the eDiscovery tool and monitoring and aiding the processing. The process does put an emphasis on imaging the devices as quickly as possible which may require some changes in when exhibits are processed.

The benefit for the examiner is having others conduct the searching and review of data which frees up time for the examiner to investigate items such as encrypted files, work on novel devices and analyse unusual data so that it can later be fed into the eDiscovery tool.

They may have a direct role in managing the processing that occurs in the eDiscovery tool and building relevant exclusion sets and subject-matter-specific hash sets.

The higher burden of proof in criminal vs. civil investigations will require the forensic examiner to produce the relevant material to a forensic standard and address the critical questions around how the material came to be present within the exhibit.

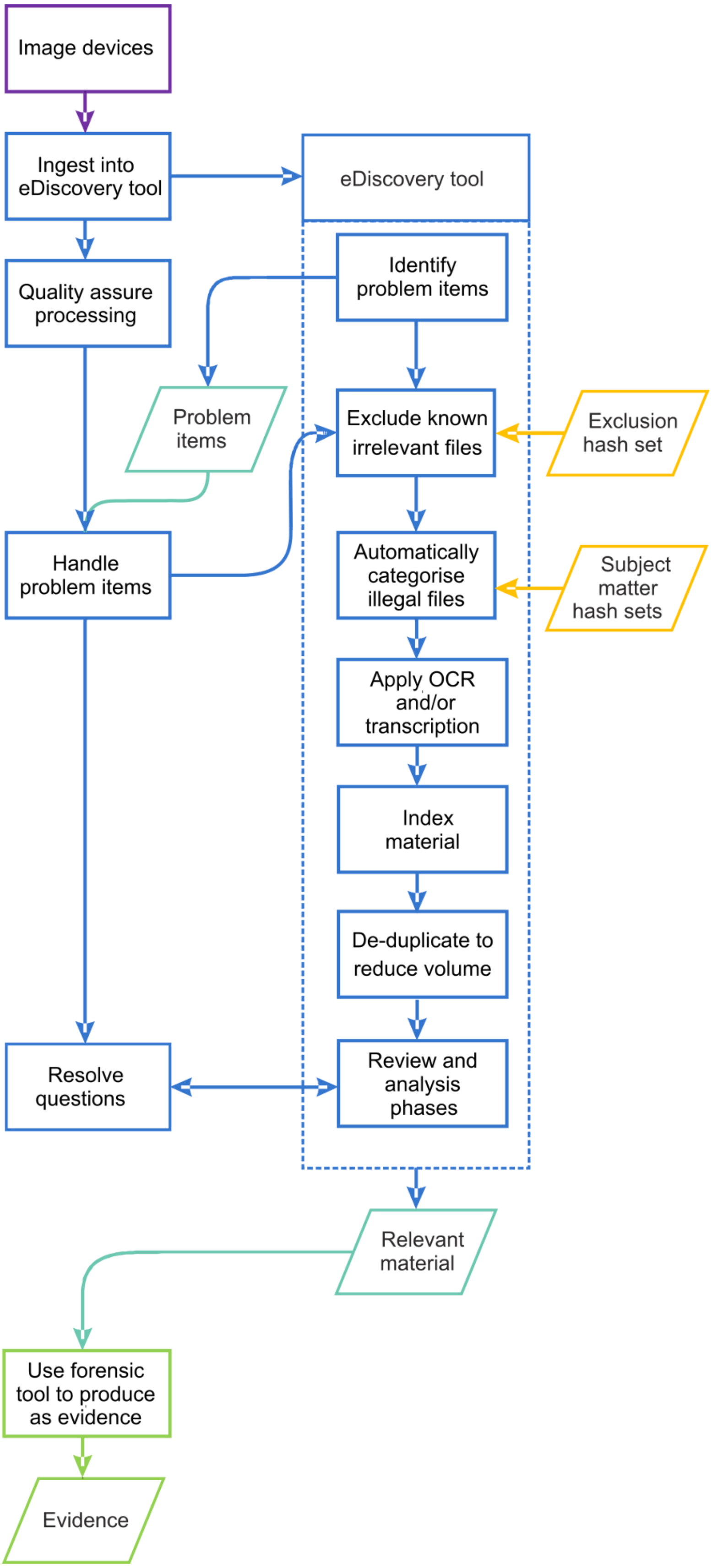
Image devices
Ingest into eDiscovery tool
eDiscovery tool
Quality assure processing
Identify problem items
Problem items
Exclude known irrelevant files
Exclusion hash set
Handle problem items
Automatically categorise illegal files
Subject matter hash sets
Apply OCR and/or transcription
Index material
De-duplicate to reduce volume
Resolve questions
Review and analysis phases
Relevant material
Use forensic tool to produce as evidence
Evidence

Figure 7: The forensic examiner’s role.

6.2.5 Licences, hardware and training

A major consideration is the cost of implementing any new approach. This can be largely broken down into three components: cost of the tool, infrastructure to support the tool and training costs.

There are several pricing models used for eDiscovery tools.

- Some tools are available for a standard licence fee supplemented by maintenance options to pay for support and upgrades. This tends to apply for the smaller tools and can see prices in the order of a few thousand pounds per licence. This puts them on a similar level to standard forensic tools.
- Some tool providers charge depending on the storage or hardware details of the installation, for example, charging per processor core in the system or by volume of data to be worked on. After the initial up-front cost, a smaller fee for annual maintenance is also typical.
- Some tools are sufficiently complex and broad in scope that there is no standard pricing; it will depend upon which software modules the user needs and any additional services required such as assistance ingesting data into the tool. This model tends to apply for the larger tools and can see total costs in the hundreds of thousands of pounds.

The smaller tools tend to be delivered as software which can be installed on standard desktop computers. For the larger tools, the hardware requirements are more demanding and are better suited to deployment on servers with high-speed networking and a combination of high-speed storage for current processes and large volumes of storage for case data. The tool suppliers will be able to advise on suitable set-ups or may also offer an appliance option where they provide both the software and a tailored set of hardware. They may also offer a Software as a Service (SaaS) model which can be an appealing option for units which only wish to use eDiscovery techniques occasionally but there are issues to consider including the security set-up, where the data will be hosted and the bandwidth available to upload data.

Training costs will vary from supplier to supplier but as a general guide, user training commonly takes less than a day and costs around £500 per user. This would cover using the tool for the review and analysis stages. A deeper level of training for individuals administering the system would typically be two or three days, cost around £1,500 and would cover the ingestion and processing stages as well as the configuration options of the tool.

6.3 eDiscovery in practice summary

eDiscovery tools are already in use in a range of law enforcement and regulatory agencies in the UK. Their application varies from providing a simple review client for a non-technical investigator through to the entire lifecycle of complex cases with multiple teams working in a coordinated manner on an investigation.

Whilst the former is relatively simple to implement with limited cost implications and changes to existing processes, the latter would be a major shift in ways of working and require significant investment. Discussing the implications with investigators and technical staff would also be an important element of achieving a working solution.

7 Conclusions

This document has looked at eDiscovery from its beginnings in indexing scanned documents through to its application in complex cases involving teams of investigators. The typical stages in an eDiscovery process are a good match to the stages in traditional digital forensics but with subtle differences in approach which provide some of the benefits but also highlight some of the additional work required to produce forensically sound evidence in a criminal case.

There are clear benefits to investigators if they can access the data relevant to their case faster and see all the relevant data in one common format rather than separate reports or platforms for data from different sources. If the investigators can be enabled to conduct their own searching of digital information then the technical staff can also benefit through having more time available to focus on the technical issues which will continue to emerge as technology progresses. Other benefits can accrue from the ability to direct reviewing towards relevant investigators and the visualisations available (albeit currently limited) to help investigators identify key patterns in the data.

There are multiple examples of investigative units within the UK using an eDiscovery approach to help manage the growing volumes of digital data involved in investigations. Whilst the small study of current tools conducted for this report showed some gaps in capability, there is a competitive marketplace with regular improvements in the tools being announced. There is also an active research base pushing the development of innovative features.

For those looking to gain a better understanding of the eDiscovery tool market, Gartner's eDiscovery Magic Quadrant document[9] may be useful as may DCIG's 2012 survey of eDiscovery Early Case Assessment software[10].

[9] Magic Quadrant for E-Discovery Software, Gartner, 2014. Available to buy from Gartner.com of free from various eDiscovery tool suppliers (registration required).

[10] eDiscovery Early Case Assessment software buyer's guide, DCIG, 2012. Available from DCIG.com (registration required).

8 Glossary

Algorithm A step-by-step procedure to perform a calculation or solve a problem. Often implemented in the form of a piece of computer software.

Analysis (EDRM) Evaluating **ESI** for content and context, including key patterns, topics, people and discussion.

Boolean logic The use of terms such as 'AND', 'NOT' and 'OR' to combine search terms.

Carving A process where a **forensic image** file is searched to find fragments of data which are not part of any recognised files.

CAST Centre for Applied Science and Technology.

Categorising Indicating that a particular item is related to a distinct, wider group of items.

Cloud A term used to describe computing resources which a user can access via the internet. This can include storage, processing and software.

Collection (EDRM) Gathering **ESI** for further use in the **eDiscovery** process.

Concept searching A method of searching which attempts to group related keywords into a concept. E.g. ‘Coke’, ‘heroin’ and ‘smack’ could all be in a ‘Drugs’ concept.

CPIA Criminal Procedure and Investigation Act 1996

Crawling A process for systematically moving through a large set of data and extracting relevant information.

De-duplication Removing duplicate items from a large set of material to reduce the volume to review.

Deleted files Files which are no longer accessible to the user of a device but which may still be intact and accessible with appropriate software.

De-NISTing A form of **white listing** using a specific library created by **NIST**.

DFRWS Digital Forensics Research Workshop.

Digital forensics 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 facilitation or furthering the reconstruction of events found to be criminal.

Digital investigation The gathering and analysis of any relevant digital data to provide both evidence and intelligence to assist with an investigation.

Disclosure The obligation on parties involved in a legal dispute to make available all material relevant to the dispute.

Discovery See **Disclosure**.

Distributed review Methods to distribute material to different teams or individuals to review, a key part of **eDiscovery**.

eDiscovery A process in which electronic data is sought, located, secured, and searched with the intent of using it as evidence in a legal case. Also referred to as eDisclosure in the UK.

EDRM Electronic Discovery Reference Model.

Electronic discovery See **eDiscovery**.

Entities Items, people, locations and other elements which have a discrete, independent existence.

ESI Electronically Stored Information.

EXIF Exchangeable Image File Format.

Forensic image To avoid altering evidence, exhibits containing digital data are ‘imaged’ or ‘cloned’ to produce an identical replica of the data on the device. The replica is known as an ‘image’ of the original.

Fuzzy searches A technique for finding matches to minor variations of a search term.

GB Gigabyte, a volume of data equal to 2^{20} bytes. Corresponds to approximately 3,000 typical documents.

Hash value A short string of letters and numbers which provide a virtually unique reference for a discrete set of data. Comparison of hashes can then identify duplicate items.

Hashing Applying an **algorithm** to a discrete set of data, e.g. a document, which results in a **hash value**.

Hash sets A collection or library of **hash values** corresponding to a specific area of interest.

Hex view A method of visualising the individual bytes in a file by representing them in hexadecimal notation.

HFS+ A file system commonly used on Apple computers.

Identification (EDRM) Locating potential sources of **ESI** and determining the scope, breadth and depth.

Indexing Constructing a list of the words that occur in a set of documents and where they occur.

Information governance (EDRM) Taking steps to ensure an organisation is ready for **eDiscovery** in order to mitigate risk and expenses, from initial creation of **ESI** through to its final disposition.

Ingestion The process of getting data, which can be from a wide range of sources, into a tool whilst preserving **metadata**.

Live files Files which are present normally in a storage device’s filing system as opposed to **deleted files** or fragments in **unallocated** space.

Matter Legal term for discrete causes or claims to be resolved.

Metadata Data about data, e.g. the date a file was created or a document’s author.

Mount A term used to describe a process which makes a **forensic image** file appear as a normal drive to an operating system.

MPS Metropolitan Police Service.

Near de-duplication Identifying items which are similar to each other, e.g. draft versions of the same document.

NIST National Institute of Standards and Technology.

NSRL National Software Reference Library – see also **White list**.

OCR Optical Character Recognition.

Predictive coding A process where an **algorithm** can analyse documents which have been manually **categorised** and attempt to detect key features which will allow related items to be automatically highlighted.

Presentation (EDRM) Displaying **ESI** before audiences, especially in native and near-native forms.

Preservation (EDRM) Ensuring that **ESI** is protected against inappropriate alteration or destruction.

Privileged material A term to cover material which has come from a protected source such as legal, medical and journalistic sources.

Processing (EDRM) Reducing the volume of **ESI** and converting it, if necessary, to forms more suitable for review and analysis.

Production (EDRM) Delivering **ESI** to others in appropriate forms and using appropriate delivery mechanisms.

Proximity search Searching for keywords within a set distance of each other.

Redaction The process of removing data from a set of material, this can be both entire items and elements within an item.

Regular Expression An advanced form of text searching which allows the searcher to specify features and structure of the text to be found rather than the specific letters and symbols.

Review (EDRM) Evaluating **ESI** for relevance and **privilege**.

SaaS Software as a Service.

Signature (file) Specific bytes of data in a digital file which are characteristic of a specific file type, e.g. a jpeg or Word document.

Spatial Relating to space, in this context, often geographical information.

Stemming Reducing a search term to its root and then using this to broaden a search.

Structured data Data, commonly stored in databases, with a strict, well-defined structure and relationships between individual elements.

Tagging See **Categorising**.

TB Terabyte, a volume of data equal to 2^{30} bytes. Corresponds to approximately 3 million typical documents.

Temporal Relating to time.

Timestamp A piece of **metadata** about a file containing information on when it was created, modified or last accessed.

Unallocated space Areas of a storage device which are not currently assigned for containing a file or files.

Unstructured data Data which has no obvious structure or text content that an **algorithm** can process.

White list A set of **hash values** corresponding to common files which will not be relevant to an investigation.

XML eXtensible Markup Language.

Appendix A. Overview of assessment

In partnership with two units within the Metropolitan Police Service (MPS), CAST conducted an assessment of commercial products offering an eDiscovery approach to reviewing large volumes of data. The assessment was conducted to allow all parties to gain a deeper understanding of what is possible with today's tools. It was not a formal performance assessment but was rigorous enough to provide a wide range of realistic tests for the tools.

Ten suppliers replied to an open call for proposals and from those, four tools were selected for assessment based on the relative merits of their proposals and an attempt to assess different kinds of tools. Through three months of testing, the tools were assessed and operational staff from both MPS units were able to gauge the suitability of the tools for their operational set-up.

It was decided that three weeks should be sufficient for the installation, training and assessment of each tool. Although the process for each tool was a little different, the first week was generally spent on installation, training and ingestion of data. The second and third weeks offered CAST the opportunity to conduct approximately five days' worth of testing and provided time for users from the MPS to run sample data to get a feel for how the tool fitted into an investigative setting.

The assessment was based upon the requirements shown in Appendix B.

Appendix B. Operational requirements

IDs	Requirement	Priority	Rationale
eDisc - 01	The tool shall be able to open and interpret standard forensic file formats	Essential	The process will be seizure of digital media, imaging of digital media, (potential processing) and use of eDiscovery tools to search the information recovered. As such, the tool must be able to work with forensic file formats.
eDisc - 02	The tool shall be able to open and interpret standard outputs from tools that process mobile devices	Desirable	Different forensic tools have different capabilities and, generally, mobile devices and smaller embedded electrical items are best dealt with by tools other than the mainstream computer forensic tools.
eDisc - 03	The tool shall be able to ingest files directly from different file systems	Desirable	It may be useful at times to simply ingest files directly from e.g. a DVD or memory stick.
eDisc - 04	The tool shall operate with forensic soundness, not altering sources and preserving metadata on extracted items	Essential	Metadata can be crucial in establishing connections and sequences of events and must not be altered by the tool. In addition, it must be possible to still produce the information evidentially. If the tool alters the information it is working on, it will be more difficult to defend.
eDisc – 05	The tool shall provide auditing functionality	Essential	There is a need for a range of information to be gathered to inform management, investigative and oversight activities. E.g. case processing information for the High tech lab or searches conducted for case review.
eDisc – 06	The tool shall have a rapid and intuitive ingestion process	Desirable	When used on big cases, the tools could be handling many TBs of data. The tool should not be slowing the process down any more than is necessary.
eDisc - 07	The tool shall be capable of supporting large operations, scaling to meet the users' needs	Essential	eDiscovery tools are being suggested for large and complex investigations where multiple teams of investigators may be working on the same case.
eDisc - 08	The tool shall be able to process data having a structured format	Desirable	Investigations may involve data contained within databases, whether internet browsing artefacts in an SQLite database or business applications in Access/Oracle etc.

eDisc - 09	The tool shall be able to process data having a semi-structured format	Desirable	Modern office documents, web pages and other files are self-describing in various ways and this information should be used by the tool to assist in understanding the data.
eDisc - 10	The tool shall be able to process unstructured data including OCR and media transcription	Desirable	Data may be present in forms other than straight text so interpretation of graphics, audio, video etc. would be useful.
eDisc - 11	The tool shall identify, extract and present spatial identifications	Desirable	Modern information is often geotagged or contains geo location data which can assist in investigations.
eDisc - 12	The tool shall identify, extract and present temporal information	Desirable	All files will have time information in terms of metadata but some files will also contain timestamped entries such as chat logs.
eDisc - 13	The tool shall be able to filter data by use of hashes	Desirable	Hashes provide a quick method for filtering out known material and avoid the need to spend time reviewing it.
eDisc - 14	The tool shall be able to filter by keyword	Essential	The tool must give the user the ability to search information based on their own knowledge which will include key words.
eDisc - 15	The tool shall understand forensic concepts such as deleted files, files in unallocated space and file signatures and present such information to the user appropriately	Desirable	The tool will be working on the output of forensic tools which will recover information beyond standard live files. This may be confusing to the investigator unless appropriately identified.
eDisc - 16	The tool shall make the contents of containers searchable	Essential	Important information may be inside containers such as archives, compound files or encrypted folders.
eDisc - 17	The tool shall detect the presence of malware or viruses and protect against them	Desirable	Seized devices may contain malicious software which could interfere with the investigator’s computer if they run the software whilst trying to preview material.
eDisc - 18	The tool shall have an intuitive interface which facilitates collaborative work, remote working	Essential	For a large team, the tool needs to assist the investigators to collaborate from a range of different locations.
eDisc - 19	The tool shall facilitate tagging/categorising items	Essential	A key part of review is being able to assign items to different categories to assist future work.
eDisc - 20	The tool shall be capable of creating collections of files for review	Essential	Rather than have all the investigators accessing the central data repository for a case, it may be that collections of files are hived off for external review e.g. to CPS.
eDisc - 21	The tool shall support standard fields for sorting/filtering files	Essential	Standard fields such as file type/size/name, file timestamps, original location etc. are key ways to focus down a review of material.
eDisc - 22	The tool shall support advanced sorting/filtering methods including regex	Desirable	The tool should be able to support users who can conduct more advanced searches.
eDisc - 23	The tool shall be capable of performing complex searches such as extended Boolean searches or similarity searches	Desirable	The tool should be able to support more advanced users who can conduct more advanced searches.

eDisc - 24	The tool shall support different user privileges and workflows	Essential	In a collaborative environment where multiple investigators are working in different ways, it is important that a manager can set up privileges and processes for different types of users.
eDisc - 25	The tool shall display contents from a wide range of files	Essential	The tool must facilitate the easy review of material so must be able to display the contents.
eDisc - 26	The tool shall assist investigators in identifying new lines of enquiry	Essential	Although hard to quantify, an underlying assumption is that providing the investigator with the ability to search information on their own terms should allow them to develop their investigation faster and in more productive areas.
eDisc - 27	The tool shall display spatial connections	Desirable	Spatial information is easier to process when it is presented graphically to illustrate the locations.
eDisc - 28	The tool shall display the relationships between items based on temporal information	Desirable	Temporal information can be used to sort information but becomes more powerful when displayed such that aspects such as the time between items is made obvious.
eDisc - 29	The tool shall support advanced search methods such as thesauri and taxonomies	Desirable	The tool should assist the user in forming better searches by suggesting alternatives and understanding the relationships between concepts.
eDisc - 30	The tool shall display internet browsing details, cookies, searches etc. in a systematic manner	Desirable	Internet artefacts are typically complicated and numerous so the tool needs to present the information sensibly.
eDisc - 31	The tool shall be able to construct timelines from both metadata and file content	Desirable	Temporal information might come from the metadata of a file or from times discovered inside a file. Both may be useful.
eDisc - 32	The tool shall identify entities and key topics without user input	Desirable	A tool’s independent identification of key entities/topics can assist the investigator in identifying new dimensions to an investigation.
eDisc - 33	The tool shall ensure continuity throughout the system (ability to identify source of evidentially relevant data)	Essential	After an investigator has highlighted an artefact as being relevant to the investigation, the digital examiner must be able to link that back to the data from the original device.
eDisc - 34	The tool shall have understandable evidential and presentational reports	Essential	The tool should produce information about the investigation, both for presentation during the investigation and for submitting as a formal document.
eDisc - 35	The tool shall have good system performance (reliability, speed of operation)	Essential	The tool is supposed to be improving efficiency but to do this it must operate on the same timescales as the operators or it will become a bottleneck. It must also be robust to minimise downtime.

eDisc - 36	The tool shall be scalable and extendable (e.g. via API)	Desirable	Investigations may grow beyond initial expectations and the expectation is that the system will be able to be upgraded easily to match this. In addition, it may be that additional functionality is required which can be implemented by a local expert.
eDisc - 37	The tool shall keep data securely	Essential	Access to the data in a case must be controllable within the law enforcement environment and only available to authorised individuals.
eDisc - 38	The tool shall have a clear cost/support model	Essential	Cost of the system will influence final decisions on procurement and inform the discussion on relative benefits. If the costs are not clear then neither can the discussion be.
eDisc - 39	The tool shall be capable of effectively managing Legal Professional Privilege (LPP) material	Essential	It is always possible that LPP material will be discovered during an investigation due to the bulk ingestion of data. This information must be handled correctly.

Table 8: Operational requirements.

Centre for Applied Science and Technology
Sandridge
St Albans
AL4 9HQ
United Kingdom

Telephone: +44 (0)1727 865051
Fax: +44 (0)1727 816233
Email: CAST@homeoffice.gsi.gov.uk

Website: https://www.gov.uk/government/organisations/home-office/series/centre-for-applied-science-and-technology-information

ISBN: 978-1-78246-465-5
Published by: Centre for Applied Science and Technology (CAST), Home Office