Trends & issues
in crime and criminal justice

Australian Government
Australian Institute of Criminology

No. 480 September 2014

Data reduction and data mining framework for digital forensic evidence: Storage, intelligence, review and archive

Darren Quick and Kim-Kwang Raymond Choo

Foreword | *The volume of digital forensic evidence is rapidly increasing, leading to large backlogs. In this paper, a Digital Forensic Data Reduction and Data Mining Framework is proposed. Initial research with sample data from South Australia Police Electronic Crime Section and Digital Corpora Forensic Images using the proposed framework resulted in significant reduction in the storage requirements—the reduced subset is only 0.196 percent and 0.75 percent respectively of the original data volume. The framework outlined is not suggested to replace full analysis, but serves to provide a rapid triage, collection, intelligence analysis, review and storage methodology to support the various stages of digital forensic examinations. Agencies that can undertake rapid assessment of seized data can more effectively target specific criminal matters. The framework may also provide a greater potential intelligence gain from analysis of current and historical data in a timely manner, and the ability to undertake research of trends over time.*

Adam Tomison
Director

The increase in digital evidence presented for analysis to digital forensic laboratories has been an issue for many years, leading to lengthy backlogs of work (Parsonage 2009). This is compounded with the growing size of storage devices (Garfinkel 2010). The increasing volume of data has been discussed by various digital forensic scholars and practitioners such as McKemmish (1999) and Raghaven (2013). While many of the challenges posed by the volume of data are addressed in part by new developments in technology, the underlying issue has not been adequately resolved. Over many years, there have been a variety of different ideas put forward in relation to addressing the increasing volume of data, such as data mining (Beebe & Clark 2005; Brown, Pham & de Vel 2005; Huang, Yasinsac & Hayes 2010; Palmer 2001; Shannon 2004), data reduction (Beebe 2009; Garfinkel 2006; Greiner 2009; Keneally & Brown 2005; Raghaven 2013), triage (Garfinkel 2010; Parsonage 2009; Reyes et al. 2007), cross-drive analysis (Garfinkel 2010; Raghaven, Clark & Mohay 2009), user profiling (Abraham 2006; Garfinkel 2010), parallel and distributed processing (Lee, Un & Hong 2008; Nance, Hay & Bishop 2009; Roussev & Richard 2004), graphic processing units (Marziale, Richard & Roussev 2007), intelligence analysis techniques (Beebe 2009), artificial intelligence (Hoelz, Ralha & Geeverghese 2009; Sheldon 2005) and visualisation (Teelink & Erbacher 2006). Despite there being much discussion regarding the data volume challenge and many calls for research into the applications of data mining and other techniques to address the problem, there has been very little published work in relation to a method or framework to apply data mining techniques or other methods to reduce and analyse the increasing volume of data. In addition, the value of extracting or using intelligence from digital forensic data has not been discussed, nor has there been any research regarding the use of open, closed and confidential source information during digital forensic analysis.

The growth in volume and number of devices impacts forensic examinations in many ways, including increasing lengths of time to create forensic copies and conduct analysis, which contributes to the increase in the backlog of requests. Digital forensic practitioners, especially those in government and law enforcement agencies, will continue to be under pressure to deliver more with less especially in today's economic landscape. This gives rise to a variety of needs, including:

- A more efficient method of collecting and preserving evidence.
- A capacity to triage evidence prior to conducting full analysis.
- Reduced data storage requirements.
- An ability to conduct a review of information in a timely manner for intelligence, research and evidential purposes.
- An ability to archive important data.
- An ability to quickly retrieve and review archived data.
- A source of data to enable a review of current and historical cases (intelligence, research and knowledge management).

In this paper, a data reduction and data mining framework is proposed that incorporates a process of reducing data volume by focusing on a subset of information. This process is not designed to replace full analysis, but provide a method of focusing an investigation to review items of importance, reduce data storage requirements for archival and retrieval purposes, and provide a capability to undertake intelligence analysis of digital forensic data. Full analysis of digital evidence may still be necessary and the data reduction processes outlined in this paper serve to support analysis rather than replace it.

The contributions of the proposed framework are two-fold:

- a data reduction method to reduce storage demands, and
- a more efficient forensic data subset collection process.

The framework provides the capability to conduct a review of a subset of data as a triage process and to store subset data for intelligence analysis, research, archival and historical review purposes.

The next section explains the challenges (primarily costs) in storing evidential data, which highlights the need for a cost-efficient data reduction process. The proposed data reduction and data mining framework is then presented, alongside an explanation of how it can be applied, as well as its benefits. The *Case study* section outlines the results of a pilot study examining the data reduction and triage potential of Step 5 in the proposed framework (see Figure 1). The last section summarises the conclusions and highlights future research.

Research motivations

Increasing data volume and cost implications

The issue of the volume of data required to be analysed in a digital forensic examination has been raised over many years. In 1999, McKemmish (1999) stated that the rapid increase in the size of storage media is probably the greatest single challenge to forensic analysis. In the interim years, there have been many publications stating the increasing volume of data is a major issue for forensic analysis. However, there have been no overall solutions proposed and the problem is still discussed. Alzaabi, Jones and Martin (2013) discuss the ongoing trend of storage capacity increasing and the prices of devices decreasing, and while there are tools and techniques to assist an investigator, the time and effort to undertake analysis remains a serious challenge. For example, Raghavan (2013: 91) states that the 'exponential growth of technology has also brought with it some serious challenges for digital forensic research' and he suggests that this is the 'single largest challenge to conquer' (Raghavan 2013: 108). When discussing the challenges posed to the field of digital forensics, Dr Eugene Spafford (cited in Palmer 2001: 7) stated that

> [d]igital technology continues to change rapidly. Terabyte disks and decreasing time to market are but two symptoms that cause investigators difficulty in applying currently available analytical tools.

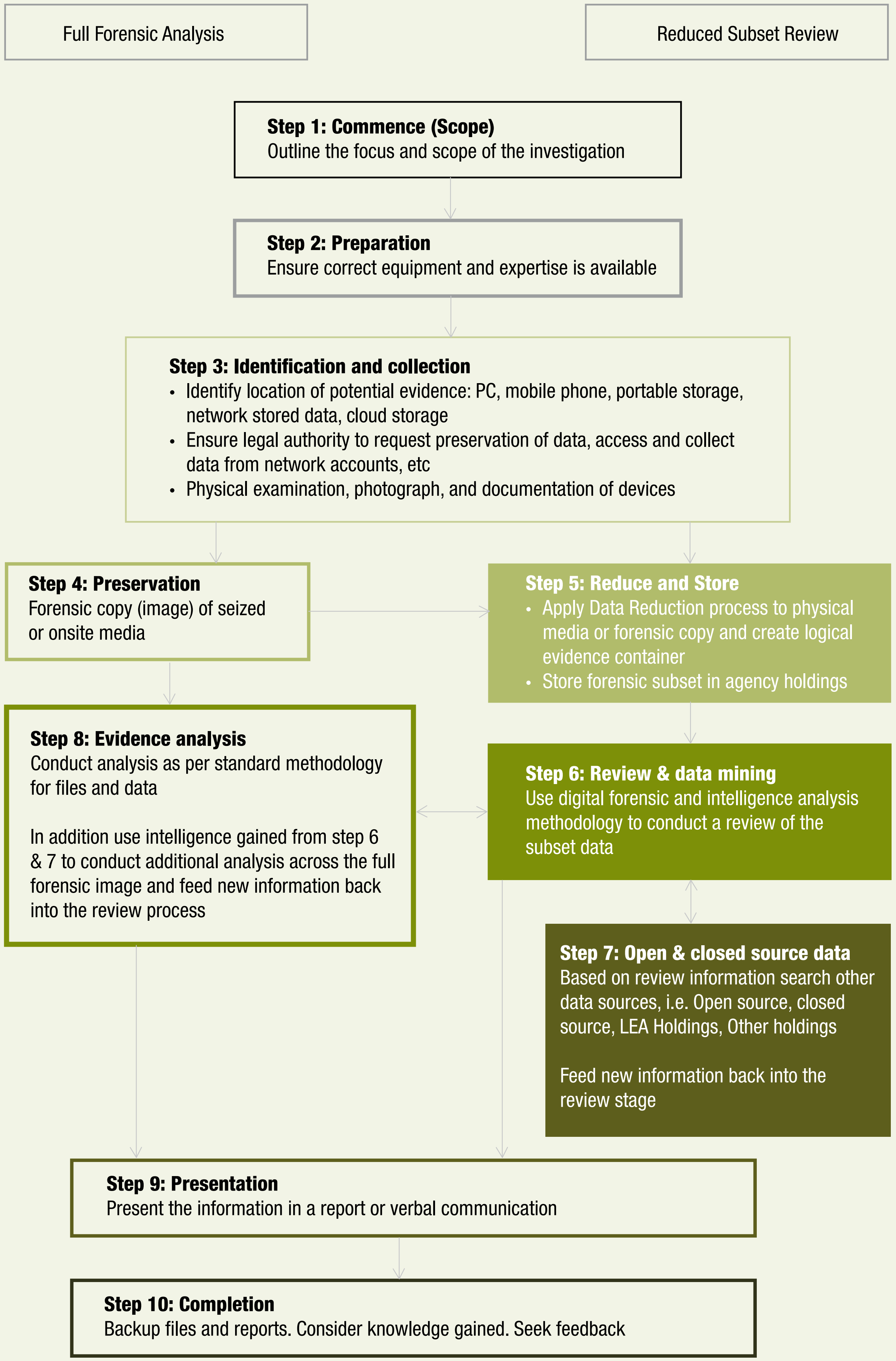
Moore's Law is the observation that the number of transistors on an integrated circuit doubles every 18–24 months and that this assists in predicting the development of technology (as cited in Wiles et al. 2007). Kryder (as cited in Walter 2005) observed that in the space of under 15 years, the storage density of hard disks had increased 1,000 fold, from 100 million bits per square inch in 1990, to 2005 when 110 gigabit drives were released by Seagate. Kryder's Law can equate to the storage density doubling every 12 months, holding true since 1995 (Wiles et al. 2007). This is about twice the pace of Moore's Law (Coughlin 2001). While storage capacity is doubling every year, the capacity to process data is only doubling every 18 to 24 months, leading to an ever-growing gap in the capability to process the volume of data seized using processing power alone.

To review the growth in digital forensic data, information from the US Federal Bureau of Investigation (FBI) Regional Computer Forensic Laboratory (RCFL) annual reports from fiscal year 2003 to 2012 (ie 1 Oct 2002 to 30 Sep 2012) were examined (FBI RCFL 2003–12). The data and figures in the reports were compiled and are summarised in Table 1.

Not surprisingly, the figures show an increase in the volume of data analysed each year, growing from 82 terabytes (TB) in fiscal year 2003 to 5,986TB (5.8 petabytes (PB)) in fiscal year 2012. This equates to an overall increase of an average of 67 percent per annum and 36 percent per annum average increase for the last five fiscal years.

Using the total volume of forensic data examined by the FBI RCFL of 20PB (see Table 1) as a baseline figure for calculations, the cost to store this volume of data uncompressed in a manner that is readily accessible is expensive. In 2011, to house 14PB of data, a commercial solution that had the ability to scale to 15PB cost an estimated US$18m.

Figure 1 Digital forensic data reduction framework



Step 1: Commence (Scope)
Outline the focus and scope of the investigation

Step 2: Preparation
Ensure correct equipment and expertise is available

Step 3: Identification and collection

- Identify location of potential evidence: PC, mobile phone, portable storage, network stored data, cloud storage
- Ensure legal authority to request preservation of data, access and collect data from network accounts, etc
- Physical examination, photograph, and documentation of devices

Step 4: Preservation
Forensic copy (image) of seized or onsite media

Step 5: Reduce and Store

- Apply Data Reduction process to physical media or forensic copy and create logical evidence container
- Store forensic subset in agency holdings

Step 6: Review & data mining
Use digital forensic and intelligence analysis methodology to conduct a review of the subset data

Step 7: Open & closed source data
Based on review information search other data sources, i.e. Open source, closed source, LEA Holdings, Other holdings

Feed new information back into the review stage

Step 8: Evidence analysis
Conduct analysis as per standard methodology for files and data

In addition use intelligence gained from step 6 & 7 to conduct additional analysis across the full forensic image and feed new information back into the review process

Step 9: Presentation
Present the information in a report or verbal communication

Step 10: Completion
Backup files and reports. Consider knowledge gained. Seek feedback

Table 1 Total volume of forensic data examined by the Federal Bureau of Investigation by fiscal year, 2003–12

	Fiscal year										
	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	Total
Service requests received	1,444	1,548	3,434	4,214	4,567	5,057	5,616	5,985	6,318	5,060	43,243
Examinations conducted	987	1,304	2,977	3,633	4,634	4,524	6,016	6,564	7,629	8,566	46,834
TB processed	82	229	457	916	1,288	1,756	2,334	3,086	4,263	5,986	20,397

Source: FBI RCFL 2003–12

A cheaper option in 2013 is to store the data using widely available 3TB removable hard drives, with the estimated cost of hard drives alone being US$922,292. This consists of 6,588 external hard disk drives purchased for US$139.99 from a consumer electronics store. However, the forensic data would be archived and not available for immediate review. Tape storage or other solutions would potentially be cheaper, but also require a method to retrieve the data from the stored medium prior to enabling access to the data for processing or searching. Consequently, the data is not readily available for review or analysis.

Forensic bit-for-bit copies of hard drives or other media (commonly referred to as forensic 'images') are often compressed, using containers such as the Expert Witness, E01, or other compressed formats. Data analysis was conducted on the figures for the volume of data comprising a range of forensic case types examined by the South Australia Police (SAPOL) Electronic Crime Section (ECS). The data examined for 43 cases involving 107 evidence items compared the size of the original media with the subsequent size of compressed E01 files. It was determined that the compression amount varied according to the data on each evidence item and ranged from 92 percent to two percent of the total volume. The average compression observed across 107 hard drives was 51.1 percent. When this compression percentage is applied to the FBI's 20PB of data, this reduces the storage requirement to just over 10PB of forensic images. Hence, using the compressed forensic image format would reduce the cost to store the data.

To summarise, it would be very costly to store the entire volume of digital forensic data examined by the FBI, either in an archived or accessible format. As discussed by Garfinkel (2006), government and law enforcement agencies rarely store or archive forensic copies, which limits cross drive analysis capability. Storing or archiving forensic data, such as on networked storage solutions, is beneficial; however, the rapidly increasing volume of data requires ever expanding network storage volumes, with the associated costs.

The need for a more (cost) effective approach

There is an opportunity to consider methods to reduce the volume of data at each stage of the forensic analysis process in relation to the seven needs listed in the introduction, namely faster collection, reduced storage, timely review, intelligence, research, knowledge management, archive and retrieval. Consideration can be given to the type of data collected, stored and reviewed, with a focus on data that will provide the greatest information. Keneally and Brown (2005) outlined a process for selective imaging to address the risks associated with collecting full forensic images for large drives, primarily the cost in time and resources, by selecting which data to image at the collection stage. The legal standards of reasonableness and relevance are raised to address concerns in relation to not undertaking analysis of a full forensic image. However, it could be argued that as the difference relates to hours or days, in a criminal or civil arena, it could be deemed reasonable to take a full bit-for-bit image and conduct analysis with all available and potentially relevant data. Hence, the proposed framework (see Figure 1) retains full imaging and analysis steps, with the reduced collection and review steps included to assist and support full analysis, rather than replace it.

Beebe (2009) proposed that a solution to the volume of data challenge is to strategically select a subset of data rather than an entire bitstream copy and that the subset could include portions of unallocated space. However, it was stated that further research is needed to determine the process to be undertaken.

As an example of subset data, files such as Microsoft Windows Internet Explorer Internet history 'index.dat' files and other browser history files and folders, can provide a great volume of information in a smaller size, when compared with other data, such as unallocated clusters, or 'Pagefile.sys' memory paging files.

Hence, collecting and storing Internet history files and not collecting or storing unallocated clusters, can reduce storage requirements and still retain information that is potentially important to an investigation. There are many file types of importance such as Log Files, Windows Registry Files, Windows Desktop Search database files, Prefetch files, email archival files and Word documents. The reduction process is undertaken on the understanding that by not collecting or storing all data, there is a subsequent risk that evidential information is potentially missed and therefore a subset of data may not be suitable for full or thorough analysis.

Turner (2005) introduced the concept of Digital Evidence Bags as a method to store a variety of digital evidence while retaining information relating to the source and location of the data subset. Schatz and Clark (2006) introduced the concept of a Sealed Digital Evidence Bag, providing for referencing between evidence bags. Commercial forensic software now provides the capability of selectively imaging files to support the collection of subset data into logical evidence files.

Garfinkel (2006) discusses Forensic Feature Extraction (FFE) and Cross Drive Analysis methods. FFE is outlined as a scan of a disk image for email addresses, message information, date and time information, cookies, social security and credit card numbers (Garfinkel 2006). The information from the data scan is stored as XML for analysis and comparison. However, as the original data is interpreted, there may be instances where new techniques are difficult to apply to the original or historical data. There have been many developments in recent years whereby additional information is able to be extracted from data holdings that were previously unknown. For example, Windows Registry analysis methodologies include newly discovered areas for locating information (Carvey 2011).

The proposed Digital Forensic Data Reduction and Data Mining Framework focuses on collecting and storing original files so that any future ability to extract information from data is retained (as the original file is retained and can be reprocessed with new methodologies or tools). The FFE and Cross Drive Analysis processes are valid and provide benefits based on current knowledge and capabilities. However, storing the original files should be undertaken where possible in an effort to future-proof data holdings, which could even lead to cold-case style analysis of historical cases with new techniques or methodologies.

Proposed digital forensic data reduction and data mining framework

The proposed Digital Forensic Data Reduction and Data Mining Framework (see Figure 1) applies to various stages of a digital forensic examination. This does not replace the need for full analysis and the framework is mapped to a common digital forensic framework with breakout steps for the reduction and review stages to maintain the distinction between full analysis and the data reduction and review steps (see Figure 1). This builds on a common digital forensic framework, listed on the left side of the framework, with the reduction and review steps highlighted on the right. Current digital forensic frameworks (ACPO 2006; McKemmish 1999; NIJ 2008, 2004) have a focus on conducting thorough analysis for evidence, which as outlined above, is not replaced with this framework. The steps are aligned with the digital forensic framework of Quick and Choo (2013a)—an extension of McKemmish's (1999) framework with the intelligence analysis cycle (see Ratcliffe 2003).

The following discussion outlines the steps of the proposed Digital Forensic Data Reduction and Data Mining Framework.

Step 1 Commence

The first step serves to outline the scope of an inquiry, including background information, analysis requirements and other material, and is not altered in the reduction framework.

Step 2 Prepare

This second step of the framework is again a common one and exists to ensure the correct equipment and expertise is available. This step is not altered from common frameworks.

Step 3 Identify and collect

The third step of the framework is the process of identifying the location of potential evidence, such as a personal computer, mobile phone, portable storage, network stored data, or cloud storage. This is undertaken with appropriate legal authority to collect media containing potential evidence. This step can also include the physical examination of devices and documentation of media, including source location, time and date accuracy.

Step 4 Preserve evidence

This step relates to the preservation of evidence and includes the process of making a full forensic bit-for-bit copy (image) of media and data using common forensic tools appropriate for the media. If a physical examination has not occurred, this would be the first part of this step to ensure detail about the source of the evidence is documented. This step is not outlined in depth as it is common to standard frameworks.

Step 5 Reduced data collection and storage

This is a new step that can be undertaken prior to, at the same time as, or subsequent to, the preservation of evidence. This is on the understanding that common forensic rules and practices are complied with, namely no change to the original media is made where possible (ACPO 2006). If changes to media are suspected to result from the subset reduction collection process, this should either not be undertaken, or be done subsequent to the evidence preservation process to ensure the evidence is not put at risk of not being accepted in court due to any changes made. The subset reduction process can be run across the original (write-blocked) media, or a full forensic image.

When working with electronic evidence, there is a potential to inadvertently change original data if agency or other guidelines such as ACPO (2006) are not adhered to. Hence, forensic guidelines need to be adhered to at all stages. The reduction process should not be undertaken to the detriment of the preservation process and hence, evidential and legal requirements take priority. Examiners must adhere to current best practice in relation to electronic evidence to ensure evidence is not at risk of not being accepted in court. However, examiners are not the only impacting factor in relation to acceptance of evidence in court. In *Roman & Anor v Commonwealth of Australia & Ors* [2004] NTSC 9 (11 March 2004), it is reported that the investigating officer spent an hour looking through a tower computer, which was subsequently seized and analysed. In *R v Ravindran* (No. 4) [2013] NSWSC 1106 (15 August 2013), the actions of the seizing member potentially affected the analysis of a computer, but not the acceptance of the evidence.

The reduction process is undertaken in a forensically sound manner using hardware or software write blockers and forensic software to enable the collection of data subsets. For example, connecting a SATA hard drive via a hardware write blocker to ensure data is not altered. Forensic software is then used to access the write-protected hard drive and pre-built conditions or filters used to display and select files containing potential data of interest, such as Windows Registry files, Internet browsing history, log files, documents, software initialisation files, software data files and other files of importance. The files of interest are selected and then preserved in a logical evidence container (L01). By focusing on files of importance rather than copying every bit of a hard drive, it is possible to substantially reduce the size of data preserved (see the *Case study* section for preliminary reduction figures observed).

While the reduction process will not alleviate the need to image everything for every case, there is a potential to speed up the overall process and reduce the need to image and store full forensic copies of every item seized. In practice, a triage process using a data subset to identify which items contain potential evidence can potentially reduce the need to image everything. While this process may initially identify data or evidence relevant to a case, there may still be a need to fully image and conduct analysis of a full forensic copy (depending on investigation need). One major benefit is that if items are identified at the triage stage with potential evidence, this may alleviate the need to image everything. Collecting a data subset and undertaking a rapid review may identify evidence on an item, allowing an examiner to produce a report and supply this to investigators or legal counsel and not require a full forensic image of every item seized.

To gain the greatest benefit from data mining and intelligence analysis across disparate cases, there is a need to collect similar data across cases. The process undertaken in the pilot study (outlined in the *Case study* section) included analysis of a digital forensic corpus and real world data to identify files with potential to provide the greatest information and exclude files with the least potential to provide information. Once the files with the greatest potential were identified, a filtering process was applied to a variety of cases and investigations types to collect the same or similar data from a variety of cases.

In practice, a data subset should be collected from each item (even if not analysed) and then archived. This should be undertaken to assist with any future questions that may arise, such as questions from prosecution or legal counsel prior to court proceedings.

The benefits of the reduction process will potentially be greatest when the original exhibit has been seized and can be imaged at a later stage (if required). In a situation where an item cannot be seized, there is still a potential need to take a full forensic copy. The reduction process can still provide benefits, in this case in relation to undertaking analysis, as a subset can be taken from the forensic copy and used to undertake a review to determine if the item has potential evidence. The subset process can also be used onsite to determine if an item contains potential evidence and assist in making a decision to seize or not.

Cloud storage provides users with an ability to store large amounts of data in remotely accessible storage locations (Quick & Choo 2014, 2013a, 2013b). This can cause issues for an examiner in relation to identifying the data, collecting the data and analysing the data (Quick & Choo 2013c; Quick, Martini & Choo 2014). A review of a data subset can potentially identify cloud stored data faster than waiting for a full forensic image to complete and process (indexing, metadata extraction and other processes).

There are a range of issues relating to the collection of data from cloud storage including legal issues, the time to access and preserve the data, and undertaking analysis of the preserved data. Collecting a data subset from cloud storage has potential time and storage size savings. This can be achieved by only collecting the data with potential to provide evidence, rather than collecting every byte of data stored remotely. Conducting a review of a subset of data will also be faster than undertaking a review of a full forensic copy. However, the needs of an investigation may dictate the need to collect and preserve every byte of data stored remotely and undertake full analysis.

It is also possible to apply a reduction framework to mobile phones or tablet computers; for example, using the option to only save call-related data, internet history, email and other software data files, with large files such as pictures and video not saved within a reduced subset (a full extract collection would be first undertaken for evidential analysis purposes).

In addition, video files can be converted to thumbnail snapshots for review purposes. As an example, software that takes consecutive interval snapshots of video frames can be used, whereby the storage requirements are vastly reduced.

The data subset files can also be stored with other data subset files; for example, in a structured manner in folder and sub-folders as per the work request number, by financial year, case number allocation, exhibit number or device information. As the reduced data subsets are vastly smaller than full forensic images, it is possible to store a considerable number of subset logical containers in a comparatively small storage space. The resulting subset files can then be reviewed for relevant information (see Step 6).

Step 6 Review and data mining

A review is then conducted using the smaller subset of data. As the data is substantially reduced, the time to process and review can be dramatically faster. The process used for undertaking forensic analysis (Bunting & Wei 2006; Carrier 2005; Casey 2011) can be used with the smaller subset of data and should result in a faster review of the information. The information review could consist of analysis of internet browsing history, filename information, a timeline review, Windows Registry analysis, keyword indexing and searching, hash analysis and other common forensic analysis techniques using a range of tools.

The ability to index forensic data prior to analysis has been available for many years. However, with the ever-growing size of data, the time to index the data is also growing. This is leading to longer times an examiner has to wait until the indexing is complete. The process of indexing by its very nature does not fully index every character or word and hence, searches undertaken across an index can potentially miss important evidence when compared with a full text search. By indexing a data subset, rather than the entire forensic image, there will be potential time savings in relation to processing and indexing (see *Case study*).

In addition, using the subset data for intelligence analysis and research of trends is an area that can provide substantial information to assist current and future investigations. Using an intelligence analysis methodology (as documented in a range of publications such as UNODC (2011) and Quarnby & Young (2010)) can assist to formalise the review process. When applying intelligence analysis practices to digital forensic data, expertise in relation to digital forensic analysis is beneficial to understand the relevance of information and to be able to extract meaningful inferences and hypotheses from the observed data.

The potential intelligence to be gained from digital forensic data holdings is an area that is rarely discussed in academic literature. However, there are vast potential gains to be made from undertaking analysis of these holdings for intelligence rather than just evidential purposes. Potential information can include names, addresses, vehicles, telephone numbers, associates and email correspondence. It is also possible for a psychological profile of the user to be built using the information stored within a user's computer or mobile device, such as common websites and interests of the user. This can potentially be determined from a variety of sources on a computer or mobile device, such as internet history, bookmarks, recent files viewed or multimedia played and a range of other intelligence.

There is also a potential to conduct analysis across a range of disparate investigations for common linkages, further providing valuable intelligence or evidence to assist in investigations and prosecutions. Additionally, researching trends over time can assist to provide information to investigators as part of focusing investigations to locate evidence earlier. For example, research of historical case data may highlight a trend showing the increased use of specific internet chat software among specific criminal offenders and as such, future investigations can first look for these data remnants rather than examining data from software that has declining use.

To undertake any use of collected data, an examiner must ensure they abide by all legal authorities relating to the collection and use of seized data. There must be legal authority to collect data and also examine data, and in particular, use the data for mining or analysis purposes. Anyone accessing the seized data, whether it is a full forensic copy or data subset, must ensure they have the legal authority to do so.

At this stage, the subset data is reviewed and the findings can be utilised with other information (Step 7) to provide information for evidential analysis (Step 8). Data can also be classified according to reliability and security, as per common intelligence practice (UNODC 2011).

Step 7 Open and closed source data

The information and intelligence from the Review (Step 6) can be used to further search other information sources, such as open and closed source data. Closed source data can include confidential internal reports and other information holdings. Open source data includes information gathered from internet sources such as publicly available Facebook information, Twitter data, media reports and Weblogs (blogs). The information gained from this and the Review stage can then be used to provide input to the Evidence Analysis Step (Step 8) and serves to further increase the knowledge base used to determine information of evidential value, or of relevance to an investigation.

Table 2 Data reduction applied to SAPOL ECS cases

Item	Number of drives	Hard disks HD (in GB)	E01 (in GB)	L01 (in GB)	E01:HD ratio	L01:E01 ratio	L01: HD ratio
Smallest	1	40	4.5	.0415	11%	0.92%	0.10%
Largest	1	1000	121	.0143	12%	0.12%	0.01%
Total (all cases)	212	102396.5					
E01	107	45388	22040.68		51.1%		
L01	144	66438.5		62.98			0.196%
E01 & L01	37	9430	5197.9	22	55%	0.423%	0.233%
Average (across all)		461.4	136.79	0.44	58.7%	0.705%	0.196%

Step 8 Evidence analysis

This step is common to digital forensic analysis and is well documented (Bunting & Wei 2006; Carrier 2005; Casey 2011). Evidence analysis is conducted as per standard methodology for files and data. In addition, the information gained from conducting the review (Step 6) and other source data (Step 7) can be used when conducting analysis of the full forensic image to locate data relating to an investigation, which may result in additional information being discovered. Evidential analysis can be undertaken to confirm the findings from the review of the subset data and to locate additional data of importance. Any additional data (not present in the subset files) can be preserved in a logical evidence container and included with the reduced subset store for archive or historical review.

Step 9 Presentation

At this stage, the findings of evidence analysis are outlined. This can be in a written report format, or a verbal communication with investigators, legal counsel and a formal presentation of evidence to a court. In addition, intelligence and other findings from the Review step can be disseminated as per the intelligence analysis process (UNODC 2011). Research findings can be communicated through academic or agency specific channels.

Step 10 Complete

The final stage of the framework is to complete the examination. Practitioners involved ensure all questions have been answered and seek feedback from those involved, such as investigators and legal counsel. At this stage, considerations are made in relation to initiating new investigations or inquiries. In addition, it is important to ensure all relevant files are backed up.

Case studies

The data reduction process (Step 5 of Figure 1) has been applied to a variety of digital forensic cases and has provided for a significant reduction in data storage and archive requirements. Using SAPOL ECS case files, the data reduction process was applied to a sample of full forensic images (see Table 2). The subsequent size of the reduced dataset files (L01 in Table 2) was then compared with the size of the forensic copy (E01 in Table 2) and the original media volume sizes (HD in Table 2). Across a sample range of 34 cases from financial years 2012 and 2013 (ie 1 July 2011 to 30 June 2013) comprising 144 hard drives and other media, the volume of data was able to be reduced to 0.196 percent of total evidence drive volume.

Table 3 Data reduction applied to Garfinkel (2009) digital corpora forensic images

Item	Hard Disks HD (in GB)	E01 (in GB)	L01 (in GB)	E01:HD ratio	L01:E01 ratio	L01:HD ratio
2008 m57 Jean	10	2.83	0.088	28%	3.11%	0.88%
4Dell Latitude	4.5	1	0.0735	22%	7.35%	1.63%
charlie-2009-11-12	9.5	3.02	0.185	32%	6.13%	1.95%
charlie-work-usb-2009-12-11	1	0.00883	0.0047	1%	53.23%	0.47%
jo-2009-11-12	12	3.06	0.0971	26%	3.17%	0.81%
jo-2009-12-11-002	14.3	5.53	0.312	39%	5.64%	2.18%
nps-2009-domexusers	40	4	0.084	10%	2.10%	0.21%
nps-2011-scenario1	74.5	34.5	0.613	46%	1.78%	0.82%
nps-2011-scenario4	232.8	18.1	0.668	8%	3.69%	0.29%
pat-2009-12-11	12.1	2.97	0.243	25%	8.18%	2.01%
terry-2009-12-11-001	19.1	7	0.157	37%	2.24%	0.82%
tracy-external-2012-07-03-initl	13.2	3.47	0.000518	26%	0.01%	0.00%
tracy-home-2012-07-03-initial	17.4	3.99	0.605	23%	15.16%	3.48%
tracy-home-2012-07-16-final	17.4	3.99	0.471	23%	11.80%	2.71%
Total	477.80	93.47	3.60	19.56%	3.85%	0.75%
Average	34.13	6.68	0.26	19.57%	3.89%	0.76%

Source: Authors' compilation

The reduction process was also applied to the forensic disk copies comprising the Digital Corpora (Garfinkel et al. 2009). The results are listed in Table 3. While these figures differ from the figures from the SAPOL ECS files, this can be explained in that many of the Corpora images are scenarios purposely built on smaller hard disk drives in a test environment, rather than larger hard drives observed in actual cases.

To highlight the figures in the Corpora (see Table 3), it can be seen that in the 'nps-2009-domexusers' case, from a 40GB hard drive, the E01 file is 4GB (10%) and the resulting data subset is an 84MB L01 file (0.21%). The 'nps-2011-scenario1' disk image is of a 74.5GB hard drive and the forensic copy is 34.5GB (46%), with the resulting data subset consisting of a 613MB L01 file (0.82%). By comparison, one of the SAPOL ECS cases comprised 6TB of hard drives, which when imaged comprised 3TB of E01 forensic copies (50%) and reduced to 1.6GB of L01 data subset files (0.03%).

Applying the 0.196 percent reduction percentage to the FBI data discussed in the earlier section could theoretically reduce the 20PB of total data to only 4TB as a reduced subset of the data comprising the cases from 2003 to 2012. The potential storage cost savings are quite significant and the ability to search the data would be considerably faster (resulting in more savings).

Also observed were benefits in conducting evidence analysis by initially collecting a reduced subset and conducting a review while waiting for the full forensic image to complete. Results observed included a subset collection only taking 79 seconds to collect the reduced dataset from a 320GB hard drive (Windows 7 Professional), compared with three hours to complete a full forensic copy and another three hours to verify the copy.

Using forensic software to process and fully index the reduced subset only took two minutes 53 seconds, compared with nearly six hours to process and index the full forensic copy. In relation to the storage requirements, the E01 images comprised 218GB compared with 687MB for the L01 file (0.215% L01:HD).

A review of the subset data located information of relevance in the internet history and the registry files (website listings and recent document entries), highlighting the need to conduct further analysis of the full forensic disk image. Had there been no information found in the review, the drive would still have been fully examined, but would have been undertaken subsequent to other items of a higher priority in the investigation.

When applied in a triage manner, the Digital Forensic Data Reduction and Data Mining Framework enables rapid collection, processing, indexing and searching of subset data to take place, which can quickly highlight devices that contain potential evidential material. Other devices can be then excluded or given a lower priority if there is less chance of evidential data being present.

During the review of the data, it was also observed there was information that can be utilised for intelligence purposes, including the internet history of the user, documents authored by the user (eg resume information detailing the person's work history and experience, and email communication with associates) and other information that would be relevant for intelligence purposes. This data would also be of potential use for researching trends over time, such as specific websites visited in relation to alleged offence typologies.

Long-term storage of the reduced subset of data could also prove to be of benefit, as important data in its original format can be retained. If questions arise from investigators, prosecutors and counsel (which can often be many months after the analysis is finalised), it can be beneficial to be able to access the case subset data, such as registry files or internet history, to promptly answer questions relating to user accounts, recent documents, or browsing history, without having to fully reimage or reprocess physical evidence to enable analysis of the information.

It is also possible to examine many subset data cases by loading them into forensic software and reviewing data across a range of cases. An example is loading multiple mobile phone subset datasets (without pictures or videos) into visualisation software to locate links between disparate devices and cases.

While the reduced subset does not store all data and hence, may not be as comprehensive for full evidential analysis, it serves a need for intelligence, research and knowledge management purposes. Consider that data subsets of every case and device examined by a law enforcement or government agency could potentially be stored on relatively small hard drives or network storage. The reduction process provides the ability to search this data quite rapidly (when compared with a potential cost of storing full forensic images and the amount of time it would take to search potentially many petabytes of data). There is potential intelligence and evidential benefits in relation to an understanding of historical cases, such as the use of a particular URL across historical investigations, or matching illicit file hash values among disparate and historical cases, potentially providing valuable intelligence.

The Digital Forensic Data Reduction and Data Mining Framework can be applied as either an addition to an evidence analysis process to gain a faster understanding of information as a triage process or be considered for archival storage, cross-case knowledge, research and intelligence review benefits.

Conclusion and future work

The growth in digital forensic data has been ongoing for many years and with the predicted ongoing growth in technology and storage, is estimated to become increasingly larger over the coming years. This has led to large backlogs of evidence awaiting analysis. By utilising the Digital Forensic Data Reduction and Data Mining Framework and a reduced subset of data, a greater understanding of data can be made at a substantially reduced cost, by comparison with storing full forensic images.

The data reduction subset process can be used to triage devices and media to quickly assess which devices may contain potential evidence and hence should be examined as a priority, and which devices have less potential evidence and can be given a lower priority for full analysis.

The findings of the pilot study have demonstrated that there are potential major benefits in the areas of data storage, as well as dramatic reductions in the time to process data subsets and gain knowledge and potential evidence from digital forensic data. Future research will be undertaken in relation to applying and refining the observed time and data storage reductions observed in the pilot study to a wider range of data in investigation typologies, as well as examining the benefits in relation to analysis and data mining timeframes.

As highlighted in the case study, the indexing time for the full forensic copy was six hours, whereas the time to index the data subset was two minutes 53 seconds. In real-world cases, indexing can sometimes take more than 12 hours, or many days to complete and with the size of data in some cases being so large (6TB or larger is not uncommon), the index and database can become too large for typical software to function. Indexing has a valuable part to play in the forensic process, but the increasing time to index cases is becoming problematic and as such, indexing a data subset can provide greater time savings.

Reviewing the subset data for information that may have potential use in intelligence holdings is another benefit of the subset process as this can be undertaken quite rapidly. There is potential to utilise data mining or intelligence analysis software to streamline and automate intelligence analysis of the subset data. The next aspect of this research to be undertaken will examine in detail which files to collect, as well as the process to collect standard files across a variety of cases, to ensure the greatest potential for data mining and intelligence analysis.

In addition, cross-case analysis can provide a greater understanding of criminal offending and networks, and potentially lead to disparate cases being linked and valuable intelligence gained. Research can also be undertaken to determine trends in relation to the data observed over time. Further research opportunities include outlining and refining the reduction process to a wider range and volume of data to determine the appropriate reduction, storage and analysis methodology to gain the greatest benefit from forensic images. Further research is to be undertaken to compare the information value of the data subset with full forensic images.

An agency that seizes and analyses digital evidence should consider the reduction framework to rapidly triage and review media prior to full analysis to determine if relevant evidence is potentially located on the media. This can be used to prioritise full imaging of media according to the knowledge gained from the reduced dataset review. Forensic practitioners should consider storing subset data with backups of notes, reports and other common analysis files to answer questions that may arise subsequent to full analysis.

Another benefit is that subset data can be stored in data holdings to enable research of historical case data and intelligence analysis, where legal authority exists. A future research opportunity is to examine the potential benefits of having dedicated intelligence analysis and research of digital forensic data including the use of data mining techniques to extract intelligence from the structured and unstructured data common to digital forensic data subsets.

References

All URLs correct at April 2014

Abraham T 2006. *Event sequence mining to develop profiles for computer forensic investigation purposes*. ACSW Frontiers ‘06: Proceedings of the 2006 Australasian workshops on Grid computing and e-research: 145–153

Association of Chief Police Officers (ACPO) 2006. *Good practice guidelines for computer based evidence v4.0*. www.7safe.com/electronic_evidence

Alzaabi M, Jones A & Martin TA 2013. An ontology-based forensic analysis tool. *Journal of Digital Forensics, Security & Law* vol. 2013 Conference Supplement: 121–135

Beebe N & Clark J 2005. Dealing with terabyte data sets in digital investigations, in Pollitt M & Shenoi S (eds), *Advances in digital forensics:* 3–16

Beebe N 2009. Digital forensic research: The good, the bad and the unaddressed, in Pollitt M & Shenoi S (eds), *Advances in digital forensics*: 17–36

Brown R, Pham B & de Vel O 2005. Design of a digital forensics image mining system, in Negoita M (ed), *Knowledge-based intelligent information and engineering systems*: 395–404

Bunting S & Wei W 2006. *EnCase computer forensics: The official EnCE: EnCaseCertified examiner study guide*. Indianapolis, IN: John Wiley & Sons

Carrier B 2005. *File system forensic analysis*. Boston, NJ: Addison-Wesley

Carvey H 2011. *Windows registry forensics: Advanced digital forensic analysis of the Windows registry*. Burlington, MA: Elsevier

Casey E 2011. *Digital evidence and computer crime: Forensic science, computers, and the internet*. Burlington, MA: Elsevier

Coughlin T 2001. High density hard disk drive trends in the USA. *Journal Magnetics Society of Japan* 25(3/1): 111–120

Federal Bureau of Investigation Regional Computer Forensic Laboratory (FBI RCFL) 2003–12. FBI regional computer forensic laboratory annual reports 2003–2012. Quantico, VA: Federal Bureau of Investigation

Garfinkel S 2010. Digital forensics research: The next 10 years. *Digital Investigation* 7(Supplement 0): S64–S73

Garfinkel S 2006. Forensic feature extraction and cross-drive analysis. *Digital Investigation* 3: 71–81

Garfinkel S, Farrell P, Roussev V & Dinolt G 2009. *Bringing science to digital forensics with standardized forensic corpora*. DFRWS 2009. Montreal, Canada. http://simson.net/clips/academic/2009.DFRWS.Corpora.pdf

Greiner L 2009. Sniper forensics. *Networker* 13(4): 8–10

Hoelz B, Ralha C & Geeverghese R 2009. *Artificial intelligence applied to computer forensics*. SAC ‘09: Proceedings of the 2009 ACM symposium on Applied Computing, ACM: 883–888

Huang J, Yasinsac A & Hayes PJ 2010. *Knowledge sharing and reuse in digital forensics*. Systematic Approaches to Digital Forensic Engineering (SADFE), 2010 Fifth IEEE International Workshop on IEEE: 73–78

Kenneally E & Brown C 2005. Risk sensitive digital evidence collection. *Digital Investigation* 2(2): 101–119

Lee J, Un S & Hong D 2008. High-speed search using Tarari content processor in digital forensics. *Digital Investigation* 5: S91–S95

Marziale L, Richard G & Roussev V 2007. Massive threading: Using GPUs to increase the performance of digital forensics tools. *Digital Investigation* 4: 73–81

McKemmish R 1999. What is forensic computing? *Trends & Issues in Crime and Criminal Justice* no. 118. Canberra: Australian Institute of Criminology. http://aic.gov.au/publications/current%20series/tandi/101-120/tandi118.html

Nance K, Hay B & Bishop M 2009. *Digital forensics: Defining a research agenda*. System Sciences, 2009. HICSS’09. 42nd Hawaii International Conference on IEEE: 1–6

National Institute of Justice (NIJ) 2004. *Forensic examination of digital evidence: A guide for law enforcement.* http://nij.gov/nij/pubs-sum/199408.htm

National Institute of Justice (NIJ) 2008. *Electronic crime scene investigation: A guide for first responders, 2nd Ed.* http://www.nij.gov/pubs-sum/219941.htm

Palmer G 2001. *A road map for digital forensic research*. Report From the First Digital Forensic Research Workshop (DFRWS), August 7–8

Parsonage H 2009. *Computer forensics case assessment and triage—Some ideas for discussion*. http://computerforensics.parsonage.co.uk/triage/triage.htm

Quarnby N & LJ Young 2010. *Managing intelligence—The art of influence*. Sydney: The Federation Press

Quick D & Choo K 2014. Google drive: Forensic analysis of data remnants. *J Network and Computer Applications* 40: 179–193

Quick D & Choo K 2013a. Digital droplets: Microsoft SkyDrive forensic data remnants. *Future Generation Computer Systems* 29(6): 1378–1394

Quick D & Choo K 2013b. Dropbox analysis: Data remnants on user machines. *Digital Investigation* 10(1): 3–18

Quick D & Choo K 2013c. Forensic collection of cloud storage data: Does the act of collection result in changes to the data or its metadata? *Digital Investigation* 10(3): 266–277

Quick D, Martini B & Choo K 2014. *Cloud storage forensics*. Waltham, MA: Syngress Publishing

Raghavan S 2013. Digital forensic research: Current state of the art. *CSI Transactions on ICT* 1(1): 91–114

Raghavan S, Clark A & Mohay G 2009. FIA: An open forensic integration architecture for composing digital evidence, in Sorell M (ed), *Forensics in telecommunications, information and multimedia*: 83–94

Ratcliffe J 2003. Intelligence-led policing. *Trends & Issues in Crime and Criminal Justice* no. 248. Canberra: Australian Institute of Criminology. http://aic.gov.au/publications/current%20series/tandi/241-260/tandi248.html

Reyes A et al. 2007. Digital forensics and analyzing data, in Reyes A, Brittson R, O'Shea K & Steele J (eds), *Cyber crime investigations: Bridging the gap between security professional, law enforcement and prosecurors*. Massachusetts: Syngress: 219–259

Roussev V & Richard G 2004. *Breaking the performance wall: The case for distributed digital forensics*. Proceedings of the 2004 Digital Forensics Research Workshop

Schatz BL & Clark A 2006. *An open architecture for digital evidence integration*, in AusCERT Asia Pacific Information Technology Security Conference. Refereed R&D Stream. 21–26 May 2006. Gold Coast, Queensland

Shannon M 2004. Forensic relative strength scoring: ASCII and entropy scoring. *International Journal of Digital Evidence* 2(4): 151–169

Sheldon A 2005. The future of forensic computing. *Digital Investigation* 2(1): 31–35

Teelink S & Erbacher R 2006. Improving the computer forensic analysis process through visualization. *Commun. ACM* 49(2): 71–75

Turner P 2005. Unification of digital evidence from disparate sources (digital evidence bags). *Digital Investigation* 2(3): 223–228

United Nations Office on Drugs and Crime (UNODC) 2011. *Criminal intelligence manual for analysts*. New York: UNODC

Walter C 2005. Kryder's Law. *Scientific American* 25 July. http://www.scientificamerican.com/article/kryders-law/

Wiles J et al. 2007. *Techno security's guide to e-discovery and digital forensics*. Burlington, MA: Elsevier

The authors are with the Information Assurance Research Lab, University of South Australia.

General editor, *Trends & issues in crime and criminal justice* series:
Dr Adam M Tomison, Director,
Australian Institute of Criminology

Note: *Trends & issues in crime and criminal justice* papers are peer reviewed

For a complete list and the full text of the papers in the *Trends & issues in crime and criminal justice* series, visit the AIC website at: aic.gov.au

ISSN 0817-8542 (Print)
1836-2206 (Online)

GPO Box 2944
Canberra ACT 2601, Australia
Tel: 02 6260 9200
Fax: 02 6260 9299

Disclaimer: This research paper does not necessarily reflect the policy position of the Australian Government