CARI Project

Automated Digital Forensics

Emlyn Butterfield, Mark Dixon, Stephen Miller, and Z. Cliffe Schreuders

Leeds Beckett University and West Yorkshire Police

2018

This is a *pre-print*, in the process of undergoing academic publication.

The CARI Project

The CARI Project is a large-scale collaboration between West Yorkshire Police and the Cybercrime and Security Innovation Centre (CSI Centre) at Leeds Beckett University. The CARI Project aims to improve and incorporate an evidence-based approach into the policing of digital forensics and cybercrime investigations. An extensive needs assessment of UK policing and cybercrime and digital evidence was conducted to understand the current situation, and to identify needs across the force. The CARI Project also involved implementing a training and research programme that has impacted the capability of the digital forensics and cyber units within West Yorkshire Police to engage in research. This needs assessment and research training led to the development of a set of research proposals, which were scored and selected. Subsequently, academics and police staff co-produced 9 research and development workstreams: a framework for seizure, preservation and preservation of cloud evidence; automated forensic analysis; image linkage for victim identification and framework for image fingerprint management; automated grooming detection; frontline officer awareness development and decision support mobile app; assessment of methods of cyber training; an evaluation of the role of the Digital Media Investigator within WYP; and characteristics of victims of cybercrime. Each of these projects were designed to address needs within law enforcement and outputs include evidence-based procedures, new capabilities such as software/algorithms, and actionable intelligence.

This work was supported by a Police Knowledge Fund grant, administered by the Home Office, College of Policing, and the Higher Education Funding Council for England (HEFCE).

Executive Summary
The objective of this research was the investigation and development of a standardised data storage format for digital forensic evidence data that can be queried to allow for links to be created between cases and exhibits, both historic and current.

The project started through the literature review of ontological data, and its use within cognate areas. This identified a recent interest in the use of ontological data for the representation of shared data within security and general computing, but with limited use within Digital Forensics. Based upon this research and discussions with the DFU within a UK police force key areas of evidence and sources of evidence were identified. In particular how this data was linked with interrelationships was developed to form a standardised way of storing and representing the key data within a forensic investigation. Storing this in a centralised database type system allows for multiple cases to be uploaded and the data to be accessible to the whole of the DFU.

Through discussion with the DFU the key software packages used within a forensic investigation were identified. The outputs of these tools were identified and researched. To facilitate the extraction from these outputs, software parsers were developed. These parsers output the data from the main forensic tools into the designed standardised format.

To facilitate extraction of data from the database system, a series of queries were created. These queries allow users to ask for cases and exhibits within the system that are linked to their specific attributes, such as what names are associated with a telephone number, or what exhibit a certain file can be found upon. Initial tests have been successful.

The ability to identify the exhibits, current and historic, that may be linked to information of interest within an investigation will allow for a more focused investigation for the DFU, potentially saving time and money. The project has future development requirements, including testing on real world data, development of a graphical user interface, and the implementation of more advanced queries and ways for queries to be easily designed.

The project has delivered on its objectives by providing software tools which enable the force to extract data from the key forensic tools in a standardised format to allow the expedited analysis of exhibits and the output of data from the key tools used within the DFU, and serves as a proof of concept for a new paradigm for processing and analysing digital forensic evidence.

Results and discussion

Introduction

The prolific reliance on technology in everyday life is now ingrained into society to such an extent that there are very few activities that can be performed without the use of a digital device in one form or another - from a satellite navigation system to find a route or a watch to time a distance. The field of digital forensics, which on inception was tasked with the analysis of single standalone devices - home PC's and mobile phones - the advancement of technology and the 'socialisation' of technology now means that no longer is a single device seized as part of an investigation (where the data storage would be in the 10's of GB's). There is now the potential for data to be spread across phones, tablets, laptops, computers, gaming consoles, satellite navigation systems, and the list goes on (where the data storage could be in the 10's of TB's). This data storage explosion now means that the traditional forensic techniques and tools employed to analyse a single device are no longer as

effective. Backlogs of 6-12 months are now the norm in most Digital Forensics Units (DFU) worldwide, with the trend now moving towards dismissing devices from an investigation, or performing a cursory analysis - using techniques such as triage - instead of the full investigation of their contents. This has the potential to lead to numerous false negatives with people freed when there is actual data available. Forensic analysts are now being forced to become button pushers, allowing certain automation techniques with the tools currently used to produce the data, with no time to contextualise or fully interpret the findings. There is therefore a need to review the current use of technology and forensic processes to reduce the investigation overhead, through the reduction of the data to be reviewed or the intelligent analysis of the data. This workstream project is looking towards the use of alternative data storage formats and analysis using intelligence based analysis to provide some form of automation of at least some of the analysis of the data.

This document contains an overview of the workstream project, an overview of techniques currently applied in cognate areas as well as the techniques currently applied in the same domain. The report defines the design, implementation and evaluation of the workstream and its generated products. The document completes with a summary of future development and next stages with the generated material.

Scope of the Project

The focus of this workstream was the development of open source software to enable Digital Forensics Units to perform automated analysis of cybercrime-related data from multiple seized devices. The intention of the project will be to investigate, and where appropriate, produce the following components:

- Data extractors: smaller software packages that have the ability to parse data from some of the forensic file formats found within a UK police Digital Forensic Unit (DFU).
- Forensic ontology: a novel standardised data storage format for combining sources of information and evidence from multiple forensic tools - this would act as the intermediary storage format to allow further analysis and investigation
- Data storage: investigation of data storage in a database esque system suitable for the generated standardised format; this would allow the historic view and analysis of data
- Output: generation of results and output from the system based upon current and historic cases gathered from a set of queries that will allow some form of automation of digital investigation which are currently time consuming.

Following the investigation, and development, of the above these will be provided to the DFU to enable further in-house development.

Literature Review

The development of technology, and its integration within society, in recent years has led to the mass market of relatively inexpensive digital devices; capable of storing terabytes of data. The cost of storage has fallen from an average of $437,300 per GB in 1980 to $0.05 per GB in 2013 (Anon, n.d.); with a GB of data having the capacity to hold approximately 260,000 pages of text or about 20,000 images (Casey et al., 2009). It is now possible to purchase an 8TB hard drive as a single drive; but clearly capacity can be significantly increased and expanded through RAID (Redundant Array of Inexpensive Disks) configurations. The storage cost of data is now so negligible that people do not even stop to consider purchasing additional storage when their current device has reached capacity. This means that no longer is there just the difficulty for forensic examiners in terms of larger data storage on single devices, there's the added dimension of multiple devices. With multiple devices comes the requirement to consider the impact and the additional context of data across all of these, and their links and connections that are not obvious through a standalone investigation.

The increase in data has led to a vast increase in workload for forensic examiners, in particular in the law enforcement arena, meaning that backlogs are now present, in instances up to 3 years (James & Gladyshev, 2013) and many with 6 months to 1 year (Casey et al., 2009). Not only does a backlog impact on the criminal justice system of any country, it also negatively impacts on the life of the individual under investigation; in multiple instances such delays in an investigation, and the fact an accusation was publicly made, has led to suspects committing suicide even though some were proven to be innocent when their devices were later analysed (Palmer, 2009). Law enforcement organisations are looking to reduce this backlog by outsourcing to specialist digital forensic providers: as way of example, Dyfed Powys Police spent £128,000 outsourcing mobile phone examinations in 2012/13 (Dyfed Powys Police, 2014) and Northumbria Police £45,340.30 in 2010/11 (Police, 2010). The costs have, in the past, spiralled out of control, as in the case of Vogon International Limited and The Serious Fraud Office in which the estimated bill of work performed was £22,500 but the final invoice was £314,375 - leading to complex legal disputes (ACPO, 2011). The process of outsourcing is a difficult one, there are many questions to be asked of the contracted company as to their expertise, quality of work and the continuity of the evidence.

With the uptake of devices and data capacity only expected to increase, it is imperative that current and proposed techniques and processes are critically appraised to enable a solution to be identified. Even without outsourcing there will be continued pressure on the forensic analysts to perform forensic examinations faster in an attempt to reduce the problem. It is suggested (Shaw & Browne, 2013) that the experience of the examiner and the amount of time they have to conduct an examination significantly impacts on the thoroughness of the investigation; digital forensics is a forensic discipline and therefore it should not be possible to 'cut corners' as it has the potential to significantly impact on people's lives (Palmer, 2009).

Arguably, the two most widely used, and accepted tools within digital forensics, are EnCase by Guidance Software (Guidance Software, n.d.) and Forensic Toolkit (FTK) by AccessData (AccessData, n.d.). These two tools offer a range of support for the analysis of computers and digital devices; with a relatively recent extension into mobile device forensics. Although these tools are widely used and accepted it does not mean they are the most appropriate tool for a particular investigation. Ayers (2009) refers to such tools as first generation computer forensic tools and Garfinkel (2010) suggests that many of the tools used today are actually designed for the investigation of child pornography; which is a throwback to the early years of digital investigations where the main workload came from international law enforcement operations such as Operation ORE (Palmer, 2009). Nowadays, child pornography still accounts for a large percentage of digital forensics investigations but the scope has increased to include any and all digital based crimes: including hacking and intellectual property theft. Specifically the tools are designed, in the main, to identify single pieces of evidence and not to explicitly assist in the investigation side. Guidance Software and AccessData are attempting to expand their tool's abilities through the integration with secondary tools, such as Passware Kit Forensic (Guidance Software, n.d.), and the incorporation of additional functionality that can be programmed by users through an integrated scripting environment.

van Baar et al. (2014) and InformationWeek (2009) propose a new approach to the traditional standalone forensic process; in which an examiner images and analyses the data in a relatively isolated manner, delivering results to the instructing parties. However, Baar and Dell state that the whole analysis process can be conducted in a collaborative manner, using a shared centralised server based environment. Whereby all parties can be involved in the analysis and review of the data; with the forensic expert available for the more 'expertise' areas and the instructing parties, with the greatest amount of case specific knowledge identify the data of interest. Whilst this was found to expedite the process, it still does not solve the problem of vast quantities of information — simply

sharing the workload will speed up a process but more manpower is required — and in an environment of reduced budgets (Sommer, 2013) this is not always possible, nor is it necessarily the best solution as there is still a reliance on the examiner to deal with requests from the instructing party and the ubiquitous nature of the data.

A relatively recent attempt to speed up investigations, through the identification - and subsequent prioritisation or dismissal of exhibits - has come about with the introduction of forensic triage (Garfinkel, 2010); forensic triage attempts to reduce the number of exhibits, and subsequently data, that an examiner must examine. This is achieved through a relatively automated process, that will be implemented by less qualified individuals, allowing the 'experts' to concentrate on the investigation. The main areas of focus have been in relation to the prosecution of indecent images of children and the use of hash sets; skin tone detection and keywords are traditionally used for this. This has been relatively successful with a recent study run by the National Police Improvement Agency (Anon, n.d.) identifying that this process can vastly reduce the backlog of a police force. However, the risk with triage is that not all data is analysed, only the data that is known. Hash sets have a known flaw in that if a single bit is modified the entire hash is changed, meaning a negative result. Some work was conducted in trying to solve this problem through the use of fuzzy hashing and small block analysis. The issue with all of these solutions is that they cannot deal with the unknown and therefore anything new in the evidence could mean that it is overlooked, and the evidence dismissed as irrelevant.

Richard III & Roussev (2006) suggest a number of potential solutions to the data explosion and the problems encountered during analysis, which include:

- The redesign of forensic tools - which are suggested to have been created when data sizes were only small and have not received a significant update to reflect the current state of data and devices, as supported by Garfinkel (2010)
- Distributed Computing - the use of multiple computers and powerful servers to spread the computer intensive processing actions of a forensic investigation
- Data Reduction - the utilisation of hash sets, keyword searches and other data reduction techniques to reduce the actual amount of data to be reviewed; by removing the known only the unknown is left. Some work has occurred in the reduction of data using hash sets and a targeted analysis using keyword searches (Casey, 2013). Such techniques have no intelligence utilising mathematical algorithms and straight comparisons, such as the use of hash sets to reduce and identify images under investigation in indecent images of children cases.

Although the techniques discussed can all be applied, and clearly some have been applied, to analyse the data, what is actually left is still a significant manual process that takes a significant amount of time to complete. What is actually needed is the ability to remove the significant manual aspect of forensics, allowing the examiners to practise their expertise interpreting and analysing data (Richard III & Roussev, 2006).

This naturally leads to an argument for the use of automated software, which could ensure a more efficient and effective use of time in a digital forensic investigation - freeing investigators to perform the advanced functions, with the argument that the thoroughness of an investigation can be determined by the capabilities of the software used (James & Gladyshev, 2013; Shaw & Browne, 2013). Through the use of software automation (Thibault, 2014; Mora & Kloet, 2010; James & Gladyshev, 2013; Watson & Jones, 2013) it may be possible to automate certain aspects of digital forensics to reduce the human effect - in which errors can occur when performing repetitive and mundane tasks.

Automation has been used in numerous technical arenas, such as engineering and computing. Within the realm of Information Technology (IT) automation can be seen to link various systems and process in such a way that they may perform repetitive actions without user intervention (CCSK Guide, 2013).

There is, however, concern that through software automation forensics may suffer from 'PBF', or Push Button Forensics (James & Gladyshev, 2013), in which the forensic investigator is no longer aware of what a tool is doing or how it functions - they simply press a button and produce results. It has been argued that this approach allows a forensic examiner to place too great a dependence on a tool, and detracts from the field of digital forensics and the requirement of an 'expert' - the expertise does not come from using a tool it comes from an ability to interpret and analyse data (Kovar, 2009; Slovenski, 2014). However, James & Gladyshev (2013) go on to state that it is not ideal to automate the whole forensic process, seeing as much of the forensic analysis process relies on the interpretation of the data presented and providing contextualisation of the data. There have also been instances of people exploiting PBF, maliciously creating data that is designed to modify and exploit the popular forensic tools (Garfinkel, 2007). This malicious action can modify the findings of the forensic tool beginning to undermine the confidence in its abilities and maybe its ability to stand up in a court of law - within the user's expertise to identify and know this PBF can lead to false positives and false negatives in terms of the investigation, or a delay in an already delayed investigation process. Another issue with PBF is the fact that much of the software is proprietary, meaning that forensic investigators are unable to review the code to determine its full functionality (Meyers & Rogers, 2005). This prevents full error and suitability testing of the software, with a reliance placed on the testing conducted by the manufacturer or the Computer Forensics Tool Testing (CFTT) Project maintained by NIST (NIST, n.d.).

The collection phase of digital forensics (referenced from figure 1) has traditionally been a relatively manual process (Al-Fedaghi & Al-Babtain, 2012; Gladyshev, 2004; Basis Technology, 2013; Garrie, 2014; Forensic Science Regulator, 2012), with some automation utilised to using techniques that have not changed significantly since the birth of digital forensics.

However, automation of something does not necessarily solve the problem of increased data submitted for examination, this would allow the reduction of data to be looked at and move a significant chunk of the time to computing rather than manual time; a significant improvement on what is currently the case. However, automation can also be applied in an intelligent way allowing for more than data reduction, making inferences and analysis of the data under investigation - this is referred to as expert system in the field of artificial intelligence (Liao et al., 2009).
An area of recent research for the application of expert systems within the digital forensics arena is semantic reasoning, with ontologies – within the fields of Knowledge Engineering & Expert Systems. This area has been applied to digital forensics at various stages, with varying degrees of success. But in most instances the technique has only been applied to lexical analysis, involving the interpretation and analysis of textual information available within the data. Whilst textual information is important, and can be the lynchpin for an investigation, there are many other mediums of data, such as images, that are missing from such an investigation.

The use of ontology allows for a clear specification of a particular domain; this includes a definition of concepts and relationships of the data; classification of possible relationships as well as the constraints of the data. Through the use of an ontology (or multiples thereof) it is possible to define all data stored on a device (digital evidence) and annotate all aspects of the data. There are three families of languages (ontologies) formally defined, known as Semantic Web Languages (SWL) (Straccia, 2013):

1. Triple Language: every subject has a property and a value, such as is found in Resource Description Framework (RDF) or further extended with additional keywords with RDFS "stating that subject *s* has property *p* with value *o*"
2. Conceptual Language: attempts to define subject, or entities, through a relationship with each other; found within the Web Ontology Language (OWL). This allows for a more expressive decision problem than within triple languages.
3. Rule Based Language: generally easier to term a rule based exchange system, rather than a language, as the system does not define a single one-fits-all rule language; as found within the Rule Interchange Format (RIF)

On top of the ontology sits the semantic reasoner, this is utilised to conduct the 'searches' of the data, drawing inferences and links in the data that would not appear with a 'flat' search. Some work has been carried out in recent years applying semantic reasoning and ontologies to the automation of the creation of timelines (Chabot et al., 2014), linkage of data (Kahvedžić & Kechadi, T., 2009; Garfinkel, S., 2012) and network intrusion analysis (Saad & Traore, 2010). Previous work would suggest that this a viable field of study, and it is unclear as to why only limited research has currently been done, with much of the focus currently placed on the analysis of complex intrusion detection systems (Avancini & Ceccato, 2013; Avancini & Ceccato, 2013; Wang, J. et al., 2014; Razzaq, Anwar et al., 2014; Razzaq, Latif et al., 2014).

Methodology

Design Science has been selected as the methodology to be used for this research, literature points to the use of design science as an appropriate methodology for the development of an artefact (Kaza et al., 2011; Hevner, A. R., 2007; Leonard & Ambrose, 2012; March, S. T. & Storey, 2008; PEFFERS et al., 2007), in this instance automation system for digital forensics; and subsequently proof of concept application development. Design science is grounded in the arena of information science (IS) and is meant to address problems - solved or otherwise - more effectively utilising new or existing techniques. In particular design science is seen as a way to solve a problem that is in conjunction with people and organisations and not something separate (Hevner, A. et al., 2004).

Hevner et al (2004) provide seven guidelines for design science research which clearly covers the process of design, through to evaluation and communication of research. This was refined in the Design Science Research Process Model (DSRP) developed by Peffers et al. (2006) which contains six clearly defined stages allowing for a structured approach to the whole research:

1. Problem identification and motivation
2. Objectives of a solution
3. Design and development
4. Demonstration
5. Evaluation
6. Communication

The project has successfully completed steps 1-4 with future work needed to fully complete steps 5-6. This is mainly down to the requirement of a process change and development of a Standard Operating Procedures (SOP) before it can be implemented.

Design

When thinking about the design of an ontology it is important to first clearly define its focus and have a clear idea in terms of its future use and audience. Due to the lack of collaboration between mobile forensic tools such as XRY and the more entrenched computer forensic tools such as EnCase, the project makes the focus on the corroboration of data across devices. In particular with a focus on

the strength and focus of evidence available from mobile devices such as contacts and communication methods

Also with a view for the integration with other workstreams the project looked to incorporate evidence that may be outputted from these. This includes Sensor Pattern Noise and the results of grooming analysis on chat logs.

When looking at mobile phones there are four potential sources of evidence:

1. The mobile device, such as the handset, and its onboard memory
2. The SIM card
3. Expandable memory, usually an SD card
4. Information from the network provider, typically in the form of Call Detail Records (CDR's)

Although the wealth of information from mobile devices is restricted to a handful of sources, there is a large variety of evidence available on these devices. The primitive view of the data on these devices can be classified as seen in Figure 1 and their potential digital hosts are depicted in Figure 2.

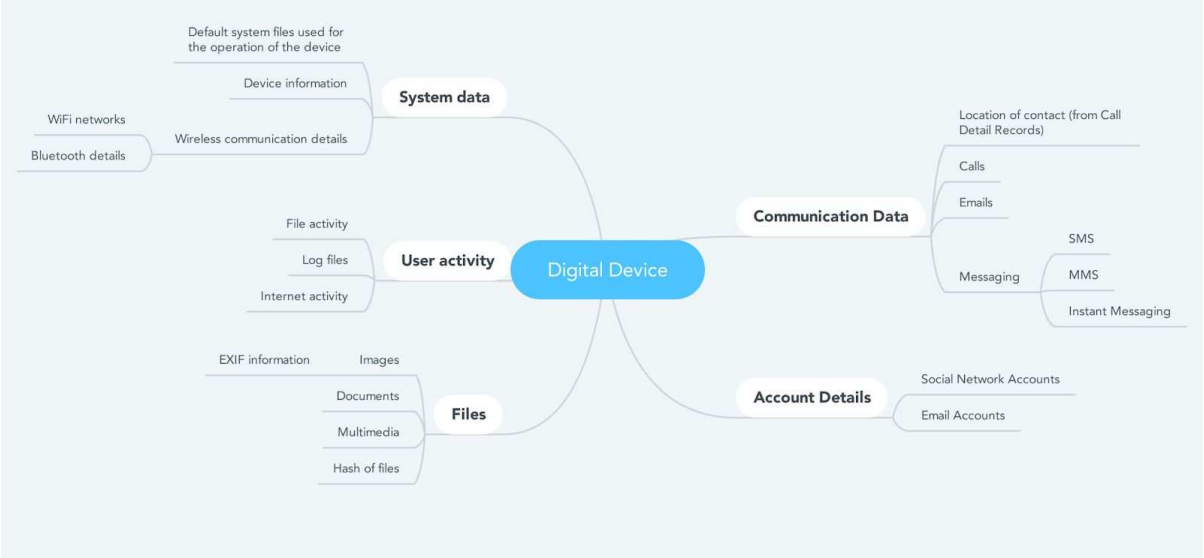

Figure 1 Potential Information from Digital Devices

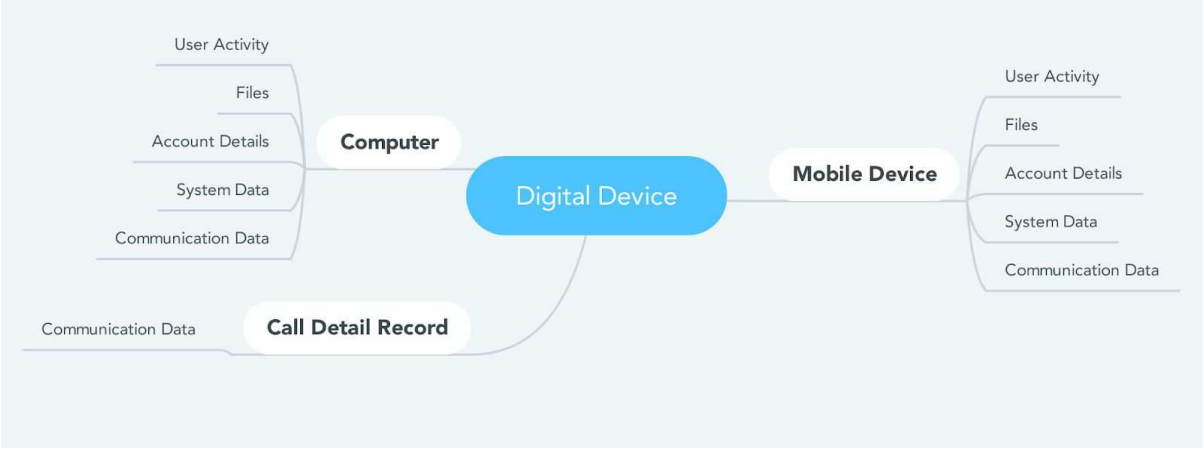

Figure 2. Digital Hosts

The data from computer based devices is as varied, if not more so. And due to the storage capacity of computers this can exponentially increase the amount of data be viewed. Computer based devices

also have the capacity to store significantly more data. The location of the information is varied across the devices looked at, for example user activity information on a Windows computer may come from the Windows Registry.

Current Tool Usage

There is a great reliance on tools within UK police forces. Given the amount of data and the number of exhibits that are due for an examination this must be the case. Currently UK police utilise the following tools as part of their analysis of digital exhibits. For mobile devices: XRY, UFED, Oxygen Forensics. For computers (and similar devices): IEF and X-Ways.

Each of these tools parse data from forensic images or physical devices. As part of this parsing they produce an output, the main output formats are listed in Table 1.

- Proprietary: this format is used by the forensic tool to store any data it has extracted and parsed, but is not in an accessible format for use by another tool for further analysis
- XML: a standard output that is not easily accessible to general users and will require further interpretation and analysis, but is generally accessible to other tools. The output is in a structured format that will allow relatively easy integration with other tools.
- XLS/XLSX: Microsoft Excel spreadsheets that utilise various worksheets and can be configured to output selected or all data from a device. This format can be problematic due to inconsistencies in the output and the lack of validation of the output. This has proven difficult to police due to badly formatted outputs preventing access to the data
- PDF: Portable Document Format, PDF, is a common format used by many applications and is readily accessible by all users. Generally, whilst this format is accessible, the data cannot be filtered due to its static nature, and therefore it is not a format that facilitates additional analysis.
- HTML: an output that is again accessible to all users, as a simple web browser. With additional coding the output can allow filtering and further analysis by the user. The output is in a structured format that will allow relatively easy integration with further tools.
- SQLite: whilst not a true output, this format is generally used for the storage of data following extraction from a digital device. The output is not easily accessible by a general user, but can be easily accessed through the development of additional tools using SQL queries.

Software	Output					
	Proprietary	XML	XLS/XLSX	PDF	HTML	SQLite
X-Ways		x	x	x	x	
XRY	x	x	x	x	x	
UFED	x	x	x	x	x	
Oxygen	x		x	x	x	
IEF			x	x	x	x

Table 1. Forensic tool outputs

The proprietary formats of XRY and UFED were initially reviewed as part of the project, these files have the file extension .xry and .ufed respectively. Proprietary formats are classed as close to “raw”

data; data that has not been formatted or had significant changes/interpretation. Proprietary formats are also generally created automatically by the tool, so no further user action is required. No information is provided by the manufacturers to allow access or interpretation of the data by third parties. Whilst some headway was made into the interpretation of the content of these files, given the time frames of the project it was decided that the proprietary formats of the XRY and UFED would not be the best use of resources. Police currently conduct some further analysis using XLS/XLS outputs. Anecdotally the force have had numerous problems with this format, due to an inconsistency of output, corruption/lack of checking of the output. The decision was then made to attempt to utilise the XML format.

XML Output Review

XML (Extensible Markup Language) is a text based language that utilises tags throughout to identify hierarchy and to organise data. Unlike HTML, XML does not specify how the data should be displayed, other tools/processes are needed to display the data stored in XML. The XML format was developed by the World Wide Web Consortium (W3C) and is available as an open standard, meaning that tools can be developed to simply read and interpret data stored in an XML format.

In terms of the extractions from XRY and UFED, the information available from XML files is wholly reliant upon what is available from the device itself. Neither XRY or UFED actually used a standardised XML format. Whilst the structure of the output is XML, it is not in a standard XML format and contains extra information and formatting. This means that it is not possible to simply allow interrogation of the data by standard libraries within various programming languages. This lack of standard output prevented standard access to the data; something anecdotally suggested by police during their early stages of looking at the outputs.

Analysis was made of the various outputs of the software and the structure of the XML to determine the structure and identify the key pieces of information needed for the project. Due to this having to be coded to be extracted not all aspects of the data were extracted, only those needed. This means that the parsers developed as part of the project are not generic parsers for all aspects of mobile phone outputs. The developed parsers extract the information depicted in Figure 3.

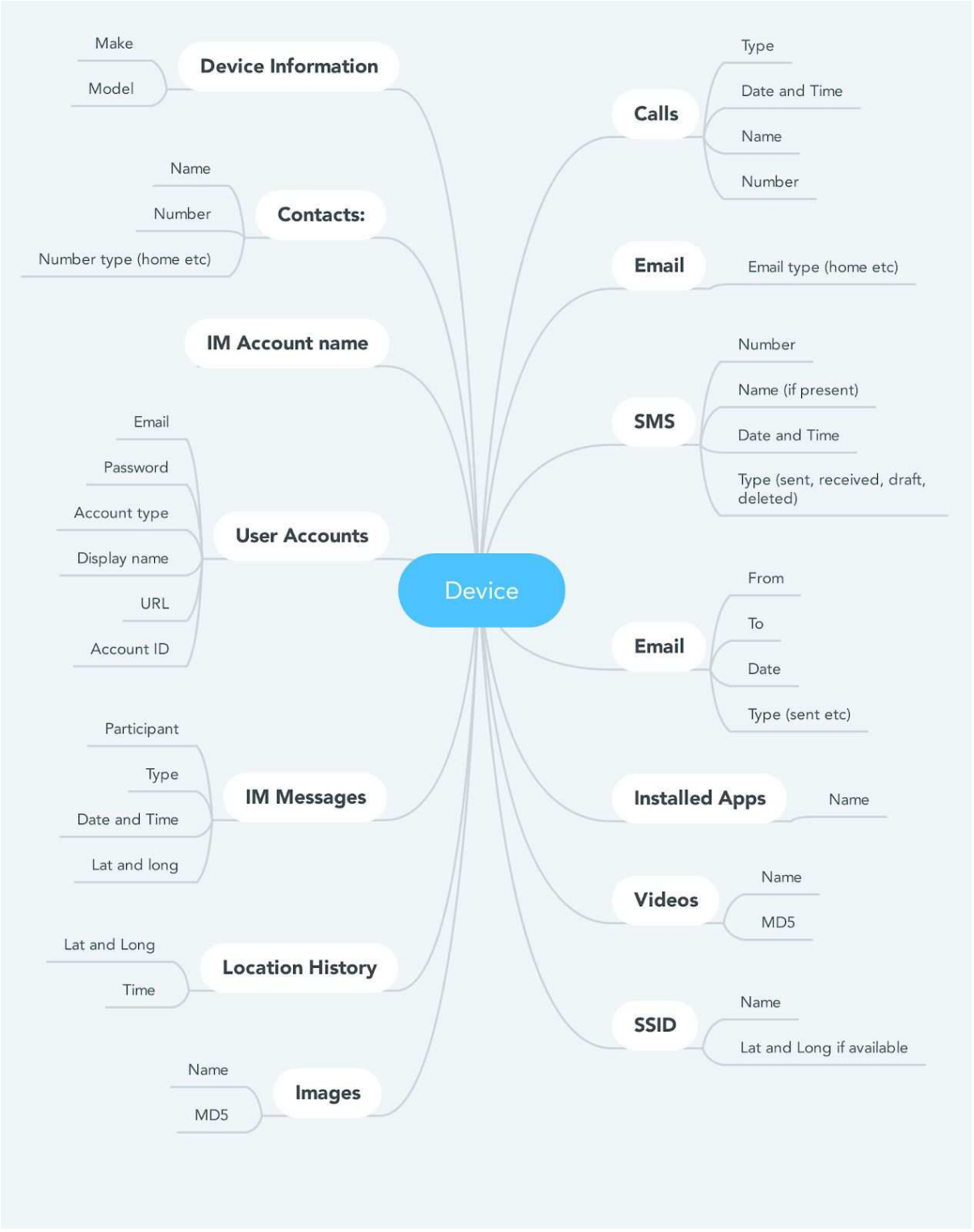

Figure 3. Extracted Information from Mobile Forensic Outputs

The way in which the key data is now identified and extracted has the added benefit of being used by the force to automatically populate reports, where appropriate, and make use of the data in alternative formats - although this process will require further investigation and development.

Computer Parsers

Initially the project looked into the integration with X-Ways, rather than manipulating the outputs of the tools. Integration with the tool would potentially allow analysts to work with and output data directly from within the application. X-Ways facilitates the use of an Application Programming Interface (API) referred to as X-Tension. X-Tension allows the following functions:

- "read from a disk/partition/volume/image
- retrieve abundant information about each file and directory in the volume snapshot
- read from any file
- add new objects to the volume snapshot, e.g. attach results of translations, decryption, decoding etc.
- bookmark/classify/categorize files by assigning them to report tables
- add free text comments to files
- run searches
- process, validate and delete search hits
- create and fill evidence file containers
- add events to the event list
- retrieve information about evidence objects
- add evidence objects to the currently loaded case"

(AG, 2017)

A review was made of X-Tensions with the view for utilising the tool to produce the required output. However, following investigation it was identified that little documentation or support exists for the development of functions within X-Tensions. This, given the project timeframe, and given the ongoing support and development by police made the use of such functions unfeasible.

Unlike most applications that utilise API's or modules, X-Tensions makes use of WIndows libraries and is required to be coded and built into DLL's that are loaded into the application when it is first launched.

As an alternative to X-Ways and to facilitate the uptake by other forces a review of current forensic tools was conducted. This identified numerous potential forensic frameworks that are well used, including: EnCase, X-Ways, FTK, and Autopsy. Of these only one tool is open source, allowing for a review, and where possible manipulation of the program to facilitate the generation of data in a wanted format. This tool is Autopsy (Carrier, 2017), which is a graphical frontend to a suite of command line forensic tools known as The Sleuth Kit (TSK).

Autopsy allows the development of modules that add additional functionality to the application, these can be written in Java or Python (Basis Technology, 2017). The use of Autopsy, or any forensic framework, should mean that the focus can be on the analytical analysis of data over concerns with the extraction of the required data from various sources.

By default Autopsy facilitates a number of key functions, these are defined by their usefulness to the workstream. When looking at alternative tools, it can be seen that they also provide similar functionality, generally with the use of additional modules or processing of the data, see Table 2.

Function	Software		
	EnCase	FTK	Autopsy
File System Support	FAT12/16/32/exFAT, NTFS,Ext2/Ext3/Ext4, Reiser, UFS (Solaris), ISO9660 (CD-ROM), DVD, HFS/HFS+, HP-UX (Digital Intelligence, 2017)	FAT12/16/32/exFAT, NTFS,Ext2/Ext3/Ext4, VXFS, ISO9660 (CD-ROM), DVD, HFS/HFS+, HFSX (Carbone, 2014)	NTFS, FAT12/FAT16/FAT32/ ExFAT, HFS+, ISO9660 (CD-ROM), Ext2/Ext3/Ext4, Yaffs2 (Carrier, 2017)
EXIF Extraction	x	x	x
Web History Analysis	x	x	x
Registry Analysis	x	x	x
LNK File Analysis	x	x	x
Email Analysis	x	x	x

Table 2. Tool Comparison

The fact that each of the tools are able to process similar data, and provide similar parsing of the data, puts it on par with the more expensive commercial environments. One thing not covered by much of the literature is a review of the reliability of each of these tools, some testing has been completed by the US Department of Homeland Security (US Department of Homeland Security, 2017), but much of this is dated and does not necessarily relate to the latest version of the software. The process of validation of a tool is partly undertaken as part of the ISO 17025 accreditation of forensic labs throughout the UK. There is nothing to suggest that Autopsy provides fewer tools or abilities, and in terms of allowing quick and easy access to the data, appears to be a much simplified interface and accessibility compared to EnCase and FTK.

When a forensic image image is added into Autopsy as part of a case, a user can automatically run ingest modules, the options available can be seen in Figure 4. A user can select to run some or all of these ingest modules. The ingest modules prepare the data for further analysis within the application.

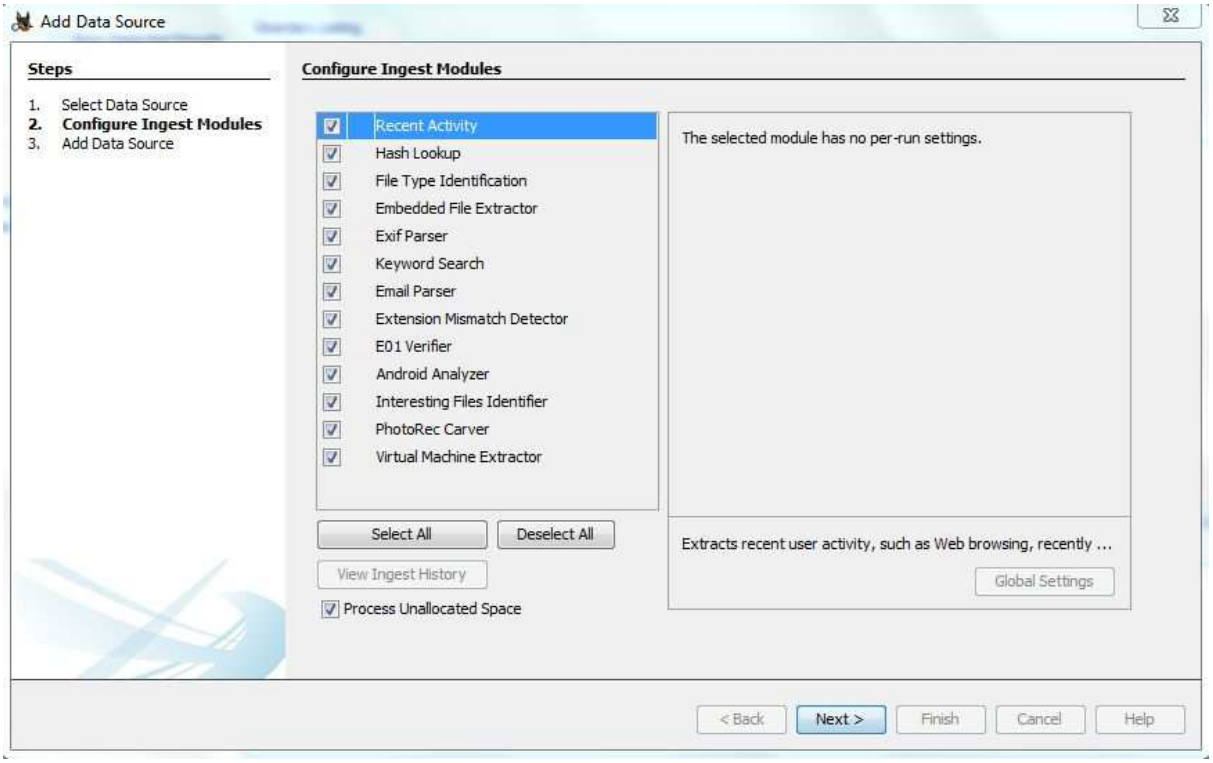

Figure 4. Autopsy Ingest Modules

Directly following the running of the ingest modules a user is able to produce reports. This includes (see figure 5): a listing of all files; Geo-coordinates; results of the ingress modules (in HTML and Excel formats).

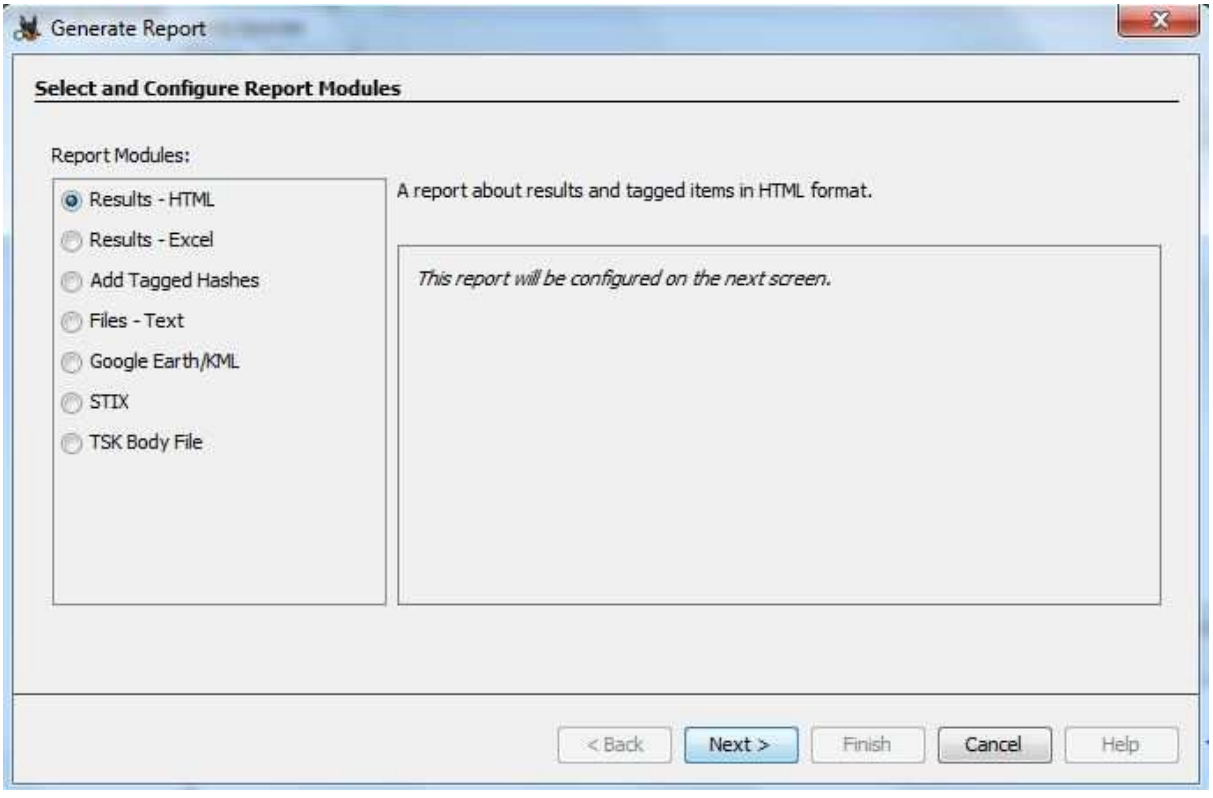

Figure 5. Autopsy Outputs

The file listing output produces a tab separated text file (TSV) detailing all of the files within the evidence, this contains the information listed in Table 2.

	Output			
	Name	*File Created*	*Hash Value*	*Full Path*
Description	The filename of the file	The date the file system reports the file was created	A unique digital fingerprint (hash value) of the file.	The full filesystem path to the location of the file on the exhibit

Table 2. Autopsy File listing

Autopsy allows the creation of output reports from the data parsing modules. The options are presented in Figure 6. A user can choose to output some or all of the modules. Once output modules are run Autopsy produces HTML reports in Figure 7.

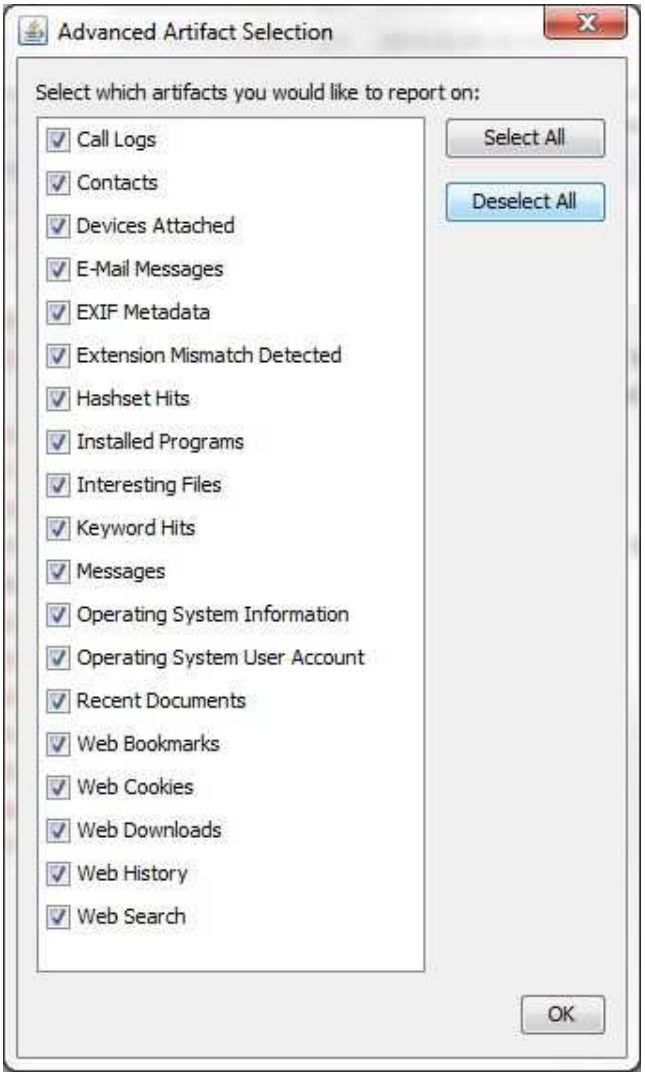

Figure 6. Module Output Reports

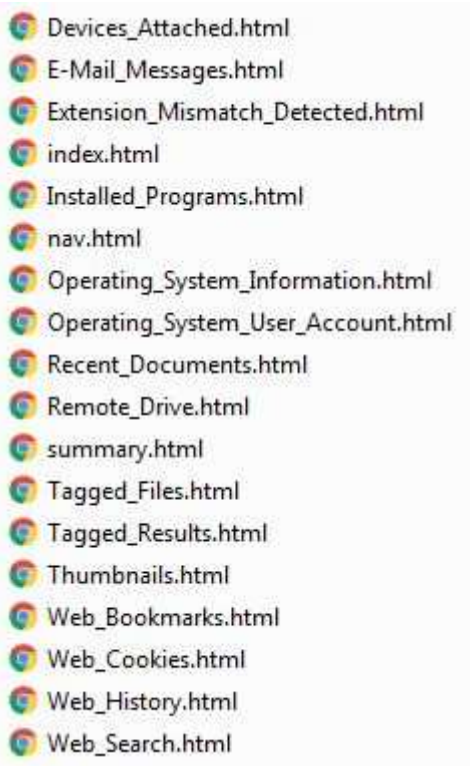

Figure 7. Autopsy HTML Reports

Whilst Autopsy does provide the ability to develop modules for the automated analysis of data, including the outputting of results through reporting modules, it was decided that the workstream would focus on utilising the current report outputs from Autopsy. This decision was based upon the timeframe and simplicity in which to create new parsers for HTML reports over additional modules for the reporting of new/different data within Autopsy.

Each of the HTML reports are made up of tables, these tables display the data in an organised and easily accessible output, and can be accessed by simply using a web browser unlike XML which is unable to display the information. A sample structure of the output from the "Recent Documents" module in Autopsy is displayed in Figure 8.

Figure 8. Recent Documents HTML Autopsy Output

Looking behind the graphical output, the HTML output is made up of a series of tags which defines a structure to the data, an example output of the Recent Documents is displayed in Figure 9. The <thead> section defines the structure of the table, with the column headings Path, Date/Time, Path ID, Source File and Tags.

```
<html>
<head>
    <title>Recent Documents</title>
    <link rel="stylesheet" type="text/css" href="index.css" />
<meta http-equiv="Content-Type" content="text/html; charset=utf-8">
</head>
<body>
<div id="header">Recent Documents</div>
<div id="content">
<table>
<thead>
    <tr>
        <th>Path</th>
        <th>Date/Time</th>
        <th>Path ID</th>
        <th>Source File</th>
        <th>Tags</th>
    </tr>
</thead>
```

Figure 9. HTML source code from Recent Documents

Following the table header the contents of each row is made up of a series of data within <tr> tags that defines each new row, and <td> tags that define the content of each cell, see Figure 10.

```
<tr>
    <td>D:\Documents and Settings\Frodo Baggins\My Documents\My Pictures</td>
    <td>2005-01-01 18:48:44 GMT</td>
    <td>3075</td>
    <td>/img_precious.E01/vol_vol2/Documents and Settings/Frodo Baggins/Recent/My Pictures.lnk</td>
    <td></td>
</tr>
```

Figure 10. Cell source code from Recent Documents

To facilitate the extraction of data from the reports produced by Autopsy, parsers were generated that parse each of the HTML reports and file listing outputs. The information that can be parsed from the reports are displayed in Figure 11.

Figure 11. Parsed information from Autopsy

Ontology Scope

The scope of the ontology can be clearly defined as: *For use by examiners within the Digital Forensic Unit to aide in the automated analysis of digital exhibits, with the ability to cross link between historic cases and exhibits.*

The scope can be further defined, and essentially a litmus test created, through the use of competency questions for the ontology. From the beginning the simple view of the project was to automate some of the analysis of the digital exhibits, in particular highlighting the relationships between exhibits and cases. Therefore in a primitive view, competency questions could be posed as:

- Has person X contacted person Y?
- Does file X exist on another exhibit/case?
- What other contact information does person X use?

Looking at the data parsed from the mobile forensic tools and computers, groupings of data can be created which form the classes of the ontology. A class is a concept within an ontology, or a type of thing.

Looking at Figures 3 and 11 the following broad classes were created, the hierarchy of such is depicted in Figure 12. The organisation of classes, using classes that are subclasses of other classes, allows the development/depiction of a tree-like structure that clearly shows how classes relate to one another.

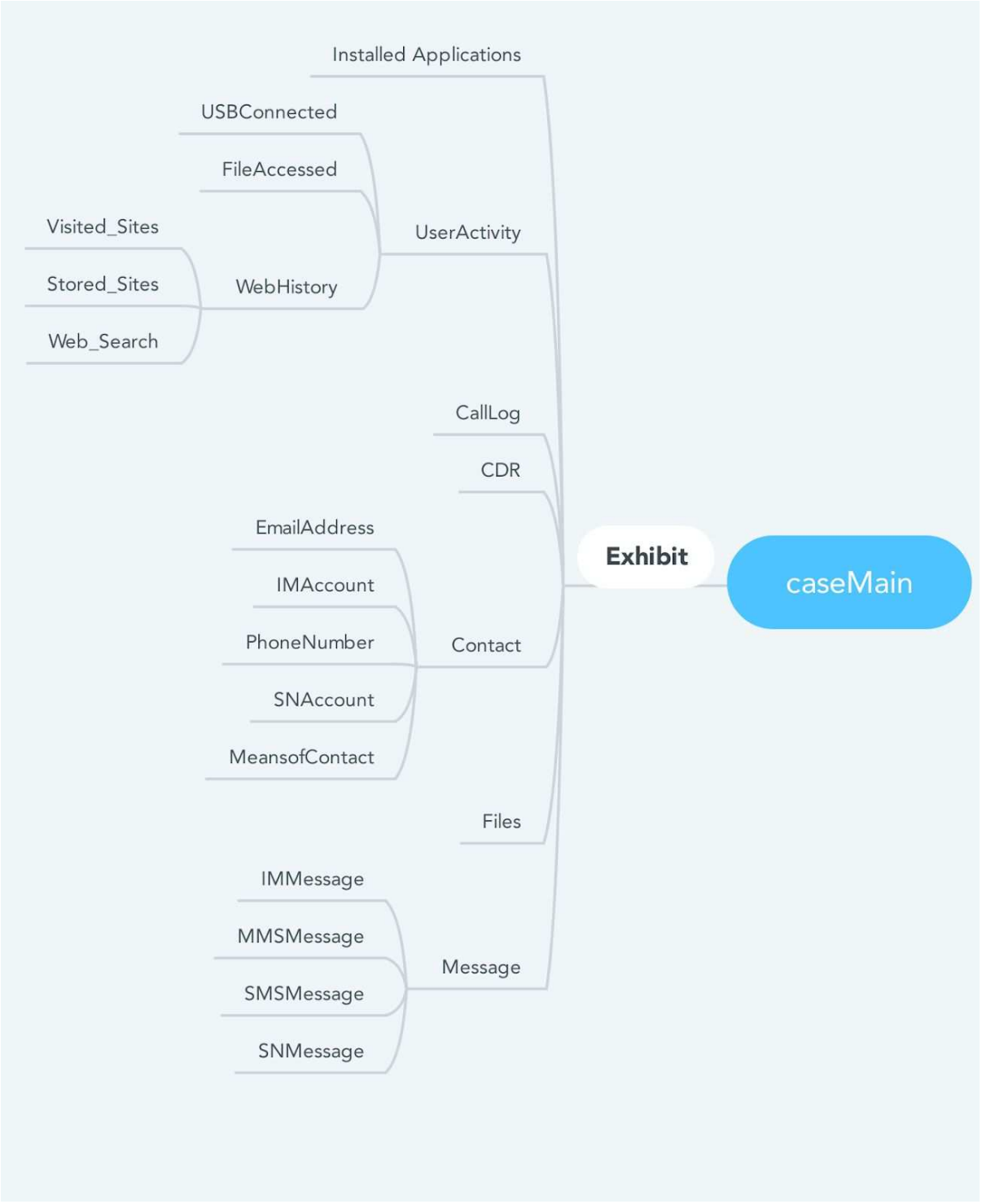

Figure 12. Ontology Class Hierarchy

One aspect that could not be accessed during the project were sample Call Details Records (CDR's). Without these records it was not possible to test this function or fully confirm the structure of these records. The records are therefore not included as part of this project.

Each of the classes contains properties that state the kind of data it will hold, for example the class files contain the properties filename; filePath, ipAddress and MD5Hash. The way in which to describe relationships between these classes and properties is implemented using predicates. There are two main types of properties within an OWL ontology:

- Object Properties: link classes to classes
- Datatype Properties: link classes to specific literals

The Ontology was encoded using Apache Jena. Apache Jena is an open source Semantic Web framework for Java. The use of this framework simplifies the creation of ontological data and is widely accepted as a suitable tool for the development of ontologies. Apache Jena has available libraries and excellent support for the development of ontologies. This level of support provides a level of support for police when they continue to develop and use this project.

Data Storage

The creation of an ontology creates a simple single file output for a single exhibit for a single case. The output is similar in structure to XML format and utilises tags to structure the data. This output file can be queried allowing questions to be asked of the data, but in itself this file has limited use and does not fully complete the project. A single output does not allow the linkage of cases or exhibits.

The facility to allow the linkage of cases and exhibits comes from the use of centralised storage systems, such as databases. There are a number of options that allow the storage and querying of ontologies. Specifically Apache Jena TDB is chosen for this project because of its simplified nature for the management and maintenance of the data. TDB compliments the development of the system using Apache Jena, and allows the full range of API's to be utilised. Unlike a Relational Database Management System (RDBMS) TDB can be used as a high performance RDF store on a single machine, and with the current configuration allows access from multiple sources. Access to the data stored within the TDB can be achieved through simplified command line scripts and via the Jena API. TDB has the added benefit that when performing transactions through TDB the data is protected against corruption, unexpected process terminations and system crashes.

There are more advanced storage options that will allow multiple access tokens from users and applications, such as Fuseki. The limitation of the TDB is that if more than a single user or application access at once it can cause corruption to the data. Whilst this is a potential future limitation, process and procedure will reduce the risk at this proof of concept stage. The use of a single TDB will also fit within the lab environment, not requiring significant infrastructure changes or implementation of high end servers.

Data Queries

Once all the data is stored within the TDB dataset it can then be queried. To facilitate the querying of the data SPARQL can be used. SPARQL provides a means of asking the dataset for information that matches the queries we use from the subset of data it has available. SPARQL queries are very similar to the SQL queries common to database administrators. SPARQL is W3C standardised and therefore well embedded in the community with a wide range of support and uptake. These queries allow for 'wildcards' to be entered that will then allow SPARQL to search the dataset.

SPARQL has four different query forms:

1. SELECT: Instructs SPARQL to return variables from within the dataset.
2. CONSTRUCT: Allows the conversion of data into other ontologies
3. ASK: Boolean result to inform if a particular graph exists. This can be used to test a query will work before wasting resources in allowing a query to run that will fail.
4. DESCRIBE: Used when the structure of the data source is unknown, allowing for less

concrete queries to be run and foundational information to be returned.

It is possible to use simple or complex SPARQL statements. We can simply return the value from within a triple statement, or we can build up more complex statements using variables within the query to pass onto another query.

As the data parsed from XRY, UFED and Autopsy is now stored within a Triples DB (TDB) it is possible to query all of the data from all cases and exhibits. For example, to return all case numbers and exhibit numbers where the telephone is set to 01138124440 the following query is run:

```
SELECT ?caseNumber ?exhibitNumber
WHERE {
        ?caseNumber contains ?exhibitNumber
        ?exhibitNumber hasContact ?PhoneNumber
        ?PhoneNumber hasNumber 01138124440
}
```

The system does not yet have the functionality of to develop new queries, queries are hard coded into the system - this is a future improvement to the system. However, given the data available it is possible to generate new queries to provide any information, for example:

- All exhibits with a certain SSID
- All contacts known by an individual
- All names associated with a number

The strength of an ontology comes from the ability to infer from the data, allowing for previously unknown relationships to be identified, for example, if suspectA has the BluetoothMAC address of suspectB within their bluetooth connection log, then it can be inferred that they have been in close proximity.

System Usage

Usage of the system has shown increased accessibility of data across devices, this has the potential to work across cases also: although this aspect has limited testing, and will only be detectable in a longitudinal study over a period of time.

The system has the benefit of making the data more accessible to both technical and less technical analysts, it has the negative aspect of taking a period of time to run and extract the data. With a reliance on the extraction tools of XRY, UFED and Autopsy the system is not yet in a position to fit comfortably within the DFU's general process. The DFU would need to change the way in which they process exhibits

The proposed system was assessed by performing queries of known data. The performance of the system was evaluated against the typical manual analysis of the data. The system performed well, with data being retrieved from the system providing information that would assist an investigation in a timely manner. The key aspects to the system were the interconnectedness of exhibits, which using previous methods of analysis would not normally be achieved.

The use and execution of the system may be improved with the implementation of higher end servers and higher specification machines. Much of the time lag in the initial stages of the system and querying the data come from the processing of the data into triples. Reduction in this time would see the system more suited for its role.

Conclusion

Due to the excessive nature of data submitted for analysis, it is no longer practical for DFU's to sit and examine every bit of data found on an exhibit. The use of an ontological approach as designed within this project has been shown to allow the identification of core forensic artefacts that collaborate between exhibits and cases. Whilst the current implementation does not perform advanced analysis of the exhibits it would be useful as part of the triage process in which it may allow the identification of those exhibits and links between those exhibits that contain potentially notable data. As with much of the forensic analysis performed by DFU's is now focused on efficient and focused investigations this project will allow such a thing to happen.

The first task of the project was to perform a literature review of ontologies and how they have been implemented within other areas of forensics, and cognate areas. Following the literature review, the needs assessment and discussions with the DFU specific requirements and evidential areas of interest when looking at digital exhibits were identified. Research and development was conducted into the relationship of this data and how interrelationships exist and can be exploited. This led to the development of a taxonomy (ontology) to store the data extracted from digital evidential devices.

To facilitate the extraction of data from different digital devices research was conducted into the main tools used by the DFU, and in particular the outputs of these. It was identified that the outputs of these were not standardised and therefore could not be used directly. The output of these tools were reverse engineered and software parsers built that parses the data from the outputs into the required format. Extractors were created for:

- file-system information, plus additional basic forensic information
- XRY and UFED XML parsers to extract core mobile phone data
- IEF parser not completed due to changes in the software (new software released which has changed the way it functions)
- CDR records were added in the early stages of the project, but due to the inability of police to release sample CDR records this is not tested

Further research and development was conducted to identify central storage options for the data parsed from the digital evidence. Apache Jena TDB data storage was chosen as the most effective, allowing inferences to be drawn of the data and the storage of data in an efficient manner. Output from the stored data, based upon queries has allowed it to be exported into a "report" - although simplistic in nature. Identifying key exhibits, cases and people.

The workstream has achieved its aims of creating a common data format that can combine various forms of evidence via ontologies and semantic queries against this centralised database. This has the potential to significantly improve digital forensic investigations, and this proof-of-concept can serve as a basis for further development to become a fully functioning tool for use by forensic examiners. It can currently identify key aspects of investigation across multiple exhibits and cases, allowing an examiner to focus their investigation; however, it could go further to build upon these core concepts to automate more of the investigation; drawing inferences from the data to add further depth to investigations.

The key knowledge generated is identifying the suitability of the use of ontologies within forensics to produce a standardised format for storing data extracted from digital evidence. It has been shown

that parsers can be created to obtain data in accessible formats from some of the key forensic tools used by the force. These also have the extra potential impact of additional future coding to allow automated reports to be generated as part of the police's Streamlined Forensic Reporting process.

The tool has future scope to develop more detailed outputs and queries. Currently the queries answer the initial questions of identifying cases and exhibits that may be linked, but queries can be further developed in conjunction with examiners.

Bibliography

AG, X. (2017) X-Ways Forensics X-Tensions API Documentation [Internet]. Available from: <https://www.x-ways.net/forensics/x-tensions/api.html> [Accessed 29 March 2017].

Basis Technology, (2017) Autopsy: Autopsy Forensic Browser Developer's Guide and API Reference [Internet]. Available from: <http://www.sleuthkit.org/autopsy/docs/api-docs/4.3/> [Accessed 29 March 2017].

Carbone, F. (2014) Computer forensics with FTK. 1st ed. Birmingham, UK, Packt Pub.

Carrier, B. (2017) Autopsy [Internet]. Available from: <https://www.sleuthkit.org/autopsy/> [Accessed 29 March 2017].

Digital Intelligence, (2017) Guidance Software Encase Forensic [Internet]. Available from: <https://www.digitalintelligence.com/software/guidancesoftware/encase/> [Accessed 29 March 2017].

US Department of Homeland Security, (2017) NIST CFTT Reports | Homeland Security [Internet]. Available from: <https://www.dhs.gov/science-and-technology/nist-cftt-reports#> [Accessed 29 March 2017].