Automating the Harmonisation of Heterogeneous Data in Digital Forensics

Hussam Mohammed[1,2], Nathan Clarke[1,3], Fudong Li[1,4]
[1]School of Computing, Electronics and Mathematics, University of Plymouth, Plymouth, UK
[2]University of Anbar, Ramadi, Anbar, Iraq
[3]Security Research Institute, Edith Cowan University, Western Australia
[4]School of Computing, University of Portsmouth, Portsmouth, UK
hussam.mohammed@plymouth.ac.uk
n.clarke@plymouth.ac.uk
fudong.li@port.ac.uk

Abstract: Digital forensics has become an increasingly important tool in the fight against cyber and computer-assisted crime. However, with an increasing range of technologies at people's disposal, investigators find themselves having to process and analyse many systems (e.g. PC, laptop, tablet, Smartphone) in a single case. Unfortunately, current tools operate within an isolated manner, investigating systems and applications on an individual basis. The heterogeneity of the evidence places time constraints and additional cognitive loads upon the investigator. Examplels of heterogeneity include applications such as messaging (e.g. iMessenger, Viber, Snapchat and Whatsapp), web browsers (e.g. Firefox and Chrome) and file systems (e.g. NTFS, FAT, and HFS). Being able to analyse and investigate evidence from across devices and applications based upon categories would enable investigators to query all data at once. This paper proposes a novel algorithm to the merging of datasets through a '*characterisation and harmonisation'* process. The characterisation process analyses the nature of the metadata and the harmonisation process merges the data. A series of experiments using real-life forensic datasets are conducted to evaluate the algorithm across five different categories of datasets (i.e. messaging, graphical files, file system, Internet history, and emails), each containing data from different applications across difference devices (a total of 22 disparate datasets). The results showed that the algorithm is able to merge all fields successfully, with the exception of some binary-based data found within the messaging datasets (contained within Viber and SMS). The error occurred due to a lack of information for the characterisation process to make a useful determination. However, upon the further analysis it was found the error had a minimal impact on subsequent merged data.

Keywords: Metadata, Digital Forensics, Heterogeneous, Characterisation, Harmonisation.

1. Introduction

The rapid development of technology over the last decade has brought various challenges to digital forensics. This development, including the variety of devices, operating systems, files and applications, clearly increases the complexity, diversity and correlation issues within the forensic analysis (Garfinkel, 2006). Conducting a forensic analysis of a case containing multi-resources and applications can be difficult due to the heterogeneity of the evidence across these devices. In general, the investigator normally takes each device and examines it individually using one of existing forensic tools to understand the nature and relationship of the artefacts. Unfortunately, these tools were designed to work on a single forensic image with specific data types (e.g. a workstation or a smartphone) (Mohammed et al, 2016).

With the significant increase in computing, individuals have increasingly become to own several devices (e.g. PC, laptop, tablet, Smartphone) with each using different applications across various platforms (Bennett, 2012). Additionally, companies producing electronic devices need to choose an operating system (OS) either open source or commercial for their core technology (Almunawar, 2018). Consequently, the files structure will be formatted according to the operating system and result in a variety of files across various OSs such as (NTFS, FAT, HFS, and Ext4) (Tanenbaum, 2009). Several applications can also run on one platform and achieve similar purposes such as web browsers (Google Chrome and Mozilla Firefox, and Apple's Safari), and messaging (SMS, Viber, WhatsApp). However, being able to examine and analyse data from across many systems and applications based on a data category at once is currently impossible.

Data categories, including files, databases, documents, pictures, media files, web browsers, etc., hold valuable information that can be used to answer some of the basic questions of a forensic investigation. Examples of the questions include, who did something to a file, when they did it and where it was carried out. Although a wide range of forensic tools and techniques exist both commercially and via open source (including Encase, AccessData FTK, and Autopsy), they only extract and analyse metadata for certain types of systems and applications (Ayers, 2009).

Recently, several researchers have tried to use metadata within the digital forensic domain to reconstruct the past events. The metadata describes the attributes of any files or applications in most digital resources (Guptill, 1999); it provides rich information about files that can lead to facilitate files processing using metadata instead of files themselves (Raghavan, 2014). Digital forensic cases can include several categories of similar metadata within a single forensic image or across multiple resources resulting in repeating the forensic process many times and increasing the workload of the investigator. Consequently, the automated correlation between the evidential artefacts from various sources is currently impossible. Therefore, in this paper, an automated approach for analysing and merging datasets by applying a novel algorithm of characterisation and harmonisation is proposed. This approach seeks to provide a fusion of similar metadata categories across multiple and heterogeneous resources within a single case. Consequently, it leads to overcome the heterogeneity issues and make the examination and analysis easier.

The remainder of the paper is structured as follows: Section 2 presents a literature review of the existing research which uses metadata in forensic investigations within single and heterogeneous resources. Section 3 describes the developed approach for metadata characterisation and harmonisation. Section 4 illustrates the entire architecture of proposed algorithms. Section 5 shows a comprehensive evaluation of the proposed via experimental results. The conclusion and future works are highlighted in Section 6.

2. Background Literature

To the best of authors' knowledge, there is no study trying to merge the datasets from across devices and applications based upon the metadata categories within the digital forensic domain. However, some researchers consider metadata as an evidentiary basis for the forensic process as it contains a rich information about electronic crimes. Therefore, a number of studies have utilised the metadata to achieve a particular purpose such as data reduction, correlation, evidential artefacts identification and many more. Regarding to the data reduction, Rowe and Garfinkel (2011) developed a tool (i.e. Dirim) to automatically determine anomalous or suspicious files in a large corpus by analysing the directory metadata of files (e.g. the filename, extensions, paths and size) via a comparison of predefined semantic groups and comparison between file clusters. Their experiment was conducted on a corpus consisting of 1,467 drive images with 8,673,012 files. The Dirim approach found 6,983 suspicious files based on their extensions and 3,962 suspicious files according to their paths. However, the main challenge with this approach is its inability to find hidden data in a file because the hidden data does not appear within the metadata of that file. It also analyses the data in each drive individually which leads to repeat the process multiple times.

Another effort was achieved by Dash and Campus (2014) to propose an approach to eliminate unrelated files for faster processing of large forensics data during the investigation by using five methods. These methods are hash values of files, frequent paths, frequent size, clustered creation, and uninteresting extensions. They tested the approach with different volumes of data that collected from various operating systems. Their experiment comprised of two steps: the first consisted of extracting frequent hashes, frequent paths, and frequent sizes to eliminate uninteresting files by matching them against NSRL-RDS database and hashsets.com hashsets; the second step was to cluster the files based on the creation time and unknown extensions for further elimination. The results of the experiment showed that an additional 2.37% and 3.4% of unrelated files were eliminated from Windows and Linux operating systems respectively. However, their approach can only be applied on file systems and applications will be excluded.

With the concept of the heterogeneity in resources and the correlation between artefacts, Case et al. (2008) proposed a Forensics Automated Correlation Engine (FACE), which is used to discover evidential artefacts automatically and identify the correlation among them. The FACE provides automated parsing over five main objects, namely memory image, network traces, disk images, log files, and user accounting and configuration files. FACE was evaluated with a hypothetical scenario, and the application was successful as the authors claimed. However, this approach can be applied to a limited number of specific resources and has not been tested with multiple resources which contain similar datasets. Raghavan et al (2009) also proposed a four-layer Forensic Integration Architecture (FIA) to integrate evidence from multiple sources. The first layer (i.e. the evidence storage and access layer) provides a binary abstraction of all data acquired during the investigation; while the second layer (i.e. the representation and interpretation layer) has the capability to support various operating systems, system logs and mobile devices. The third layer (i.e. a meta-information layer) provides interface applications to facilitate metadata extraction from files. The fourth layer (i.e. the evidence composition and visualisation layer) is responsible for integrating and correlating information from multiple sources, and these combined sources can serve as comprehensive evidentiary information to be presented to a detective. As the FIA architecture was merely conceptualised via a car theft case study, further investigation would be required for the evaluation of its practicality. Additionally, there is no explanation about how the system will work if the resources contain similar evidential categories.

In attempting to find the evidential artefacts in an automated way, Al Fahdi et al. (2016) proposed an automated approach for identifying the evidence and speeding up the analysis process for computer forensics. Their approach mainly consists of three general steps: metadata extraction, clustering and automated evidence identification. Real forensic datasets have been utilised to apply their approach, and four metadata categories instead of files themselves have been chosen and extracted individually (i.e. File system, Email, EXIF and Internet history). They then used unsupervised pattern recognition to cluster evidential artefacts to aid the investigators to focus on the evidential files thereby saving their time and efforts. The Self-Organising Map (SOM) was utilized for automatically grouping the input data without any supervision. The investigator determined the number of clusters before the process starts. Afterward, the automated evidence profiler (AEP) algorithm was applied to analyse and identity the related artefacts across all metadata SOMs. The AEP contain two steps: first is to identify the first cluster based on prior work achieved in profiling criminal behaviour; the second step is to identify subsequent clusters using the timeline analysis of each file in the first cluster. Their experiment was conducted by using four forensic cases, where each case includes a single forensics image. The experiment based on clustering has shown that 93.5% of interesting artefacts were grouped in the top five clusters. While the AEP algorithm has presented acceptable results and shown that the algorithm can reduce the investigator's time to analyse the cases and present the relevant evidence in a report. However, their approach was only applied to single images with a limited number of metadata categories. Moreover, the AEP algorithm does not work with all cases because it depends on some prior work completed in profiling criminal behaviour to identify the first cluster. There might be new criminal behaviour cases which are not analysed yet.

As demonstrated above, existing studies have attempted to use metadata for forensic purposes; however, they either applied their approaches on a single forensic image or on several forensics images which are different in nature. For instance, these cases consist of a hard disk, network packets, a memory dump, and many others which they do not contain same evidential resources or datasets. In contrast, more forensics cases that include the same evidential artefacts coming from different resources become more common. Therefore, there is a need to merge datasets based on metadata categories to process them as a single image thereby saving the investigator's time and effort.

3. An Automated Approach for Metadata Characterisation and Harmonisation

The proposed approach seeks to provide an automated framework to merge similar datasets by characterising similar metadata categories and then harmonising them in a single dataset. This approach overcomes the heterogeneity issues and makes the examination and analysis easier by analysing and investigating the evidential

artefacts across devices and applications based upon category to query data once. The proposed approach is illustrated in figure 1.

This approach utilises the metadata categories as a base to merge datasets; also, datasets that contain non-metadata fields should be eliminated. For example, Skype and SMS applications contain fields describing the actual content of messages. Therefore, the variability of string can be used to identify meta from non-metadata fields because most metadata of a same field has a specific structure and format; and most non-metadata fields are in the string format. For instance, the dimension of an image is presented as (width x height) (e.g. 300x200, 2000x1500), and this pattern of string can be represented as (NxN) which means (Number, Letter x, Number). Additionally, the file name in most operating systems can be represented (Name.exetension) which means (String, Full Stop, Short String). Consequently, the string variability has the ability to analyse the string to produce a pattern that aids to find the similar metadata fields across multiple categories.

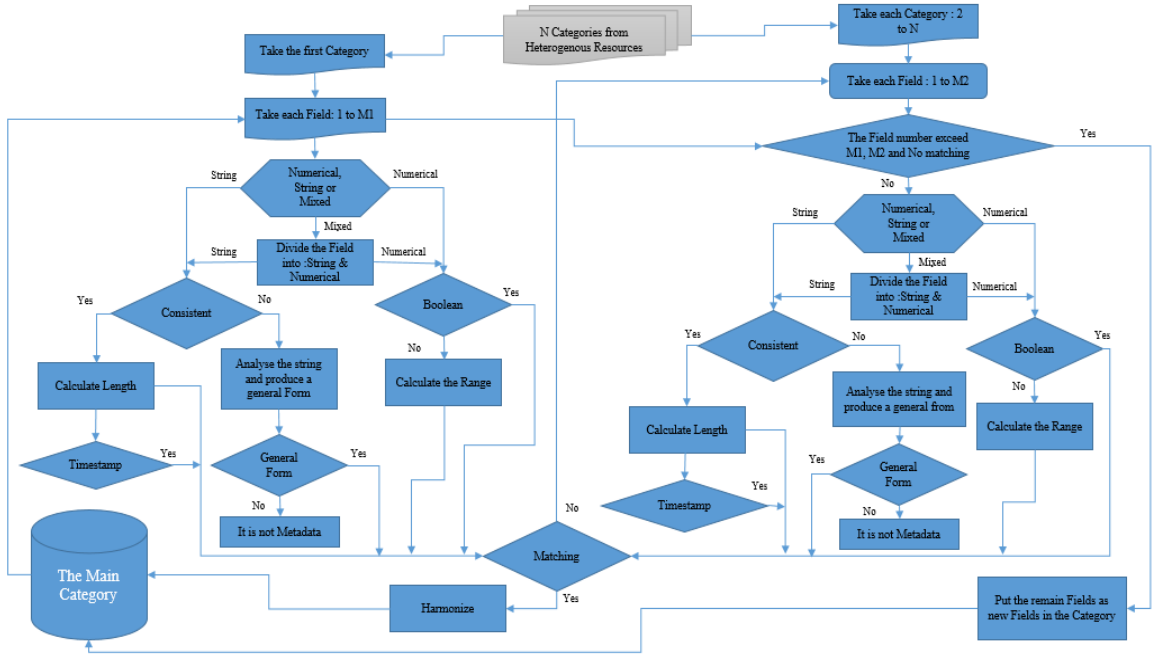

Figure 1: Metadata Characterisation and Harmonisation Process

The forensic cases can include several categories of similar metadata within a single forensic image or across multiple resources. This can lead to repeating the forensic process many times and increase the encumbrance placed upon the investigator. As a result, the automated approach for metadata characterisation and harmonisation splits the problem of merging the datasets into following aspects:

- How to characterise the metadata categories.
- How to merge and harmonise similar metadata categories.

The solution to the first problem can be achieved by using a rule-based system with a high level of fundamental conditions and rules. Rule-based systems are a method used to manipulate the knowledge to interpret information in a useful manner (Aronson et al, 2005). There is a limited number of the fundamental conditions utilised such as string, consistency, numerical, Boolean, and timestamp. The characterisation algorithm uses these rules and conditions which contain all the appropriate knowledge for matching similar categories. Regarding the string condition, the string variability algorithm will be utilised to produce a specific pattern which aids to check and match a similar field of strings across various categories. The consistency condition means that all the string values within the field should have a fixed length of string with the same pattern. While the numerical condition can be identified by measuring the range of the field within the category to match with another field in the compared category. Additionally, most files do have two sizes: physical and logical size with

a slight difference between them. The algorithm can identify the physical and logical size across various categories. The Boolean data type is a field with only two possible values: true or false. The timestamp is considered as a fundamental condition because it exists within most files and applications. This algorithm can characterise most of the timestamp formats across various categories. The final output of the characterisation process is a record that contains all similar metadata categories as shown in figure 2.

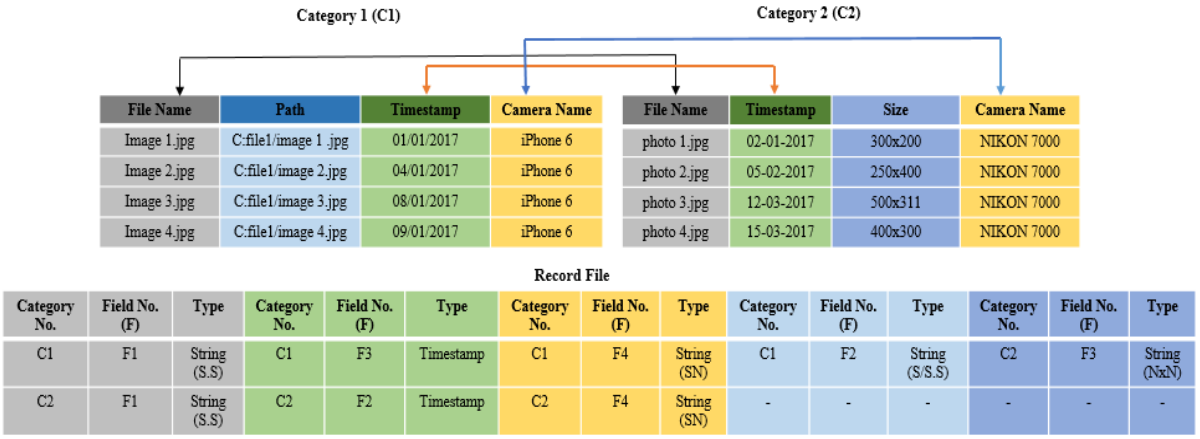

Figure 2: Characterisation Process

The second problem can be solved by applying the harmonisation algorithm which utilises to merge the similar categories based on the characterisation record. It can adjust the differences and inconsistencies among different measurements, methods, procedures, schedules, specifications, or systems to make them uniform or mutually compatible. Many fields within the metadata categories are stored in various forms across heterogeneous systems (i.e. timestamp, phone number, and file size). For example, the timestamp can be stored in several forms such as ('yyyy-MM-dd', 2014-04-19), ('dd/MM/yyyy', 19/04/2014), ('dd.MM.yyyy', 19.04.2014) ('yyyy-MM-dd''T''HH:mmXXX', 2014-04-19T21:41-04:00) or can be formed as a Unix timestamp which is just number with 10 digit or 13 digit. Likewise, phone numbers can be represented in different ways (i.e. they can be stored with country codes or area codes). Additionally, the country code can be placed in a varchar type (e.g. +91-9654637894). The file size can also be saved in variety units of measurement (i.e. it is measured from the lowest to the highest in bits, bytes, kilobytes, megabytes, gigabytes. Consequently, the core of harmonisation process to merge the similar categories in a systematically way and make them uniform as illustrated in table 1.

Table 1: Harmonisation Process

File Name	Timestamp	Camera Name	Path	Size
Image 1.jpg	01 Jan 2017	iPhone 6	C:file1/image 1 .jpg	-
Image 2.jpg	04 Jan 2017	iPhone 6	C:file1/image 2.jpg	-
Image 3.jpg	08 Jan 2017	iPhone 6	C:file1/image 3.jpg	-
Image 4.jpg	09 Jan 2017	iPhone 6	C:file1/image 4.jpg	-
photo 1.jpg	02 Jan 2017	NIKON 7000	-	300x200
photo 2.jpg	05 Feb 2017	NIKON 7000	-	250x400
photo 3.jpg	12 March 2017	NIKON 7000	-	500x311
photo 4.jpg	15 March 2017	NIKON 7000	-	400x300

4. System Architecture for Merging Multi-Images in Digital Forensics

The proposed architecture attempts to bridge the gap between several evidential resources included in a single case. It aims to decrease the burden on the investigator by merging similar datasets from multi-resources and producing a single forensic image thereby dealing with all data at once. To achieve this, preliminary steps should be undertaken to prepare the datasets before merging them. These steps include resources acquisition, data carving, and hashing (pre-processing), and metadata extraction. Therefore, all available suspect resources within

a single case should be acquired in a forensically sound manner to produce forensic images becoming authentic, reliably obtained, and admissible. The pre-process step can recover and extract files from the unallocated file system space (i.e. data carving), it then finds the hash values of all files for identification, verification, and authentication purposes. Having established that the metadata can help to recognise patterns, establish timelines, and point to gaps in the datasets, it can aid to correlate the evidential artefact in the digital investigation. Therefore, automated process of metadata extraction undertakes to obtain the suitable information (metadata) for the digital forensic process. This information can be extracted or created from any file or application such as file systems, network packets, databases and many more. However, a number of metadata categories might contain fields which are not metadata. Thus, the meta and mon-metadata identification process eliminates these fields, but at same time it considers an optional step as it can only be applied to specific categories. Afterward, the characterisation process identifies and analyse the nature and the types of datasets in order to merge them using the harmonisation process. The entire system is illustrated in figure 3.

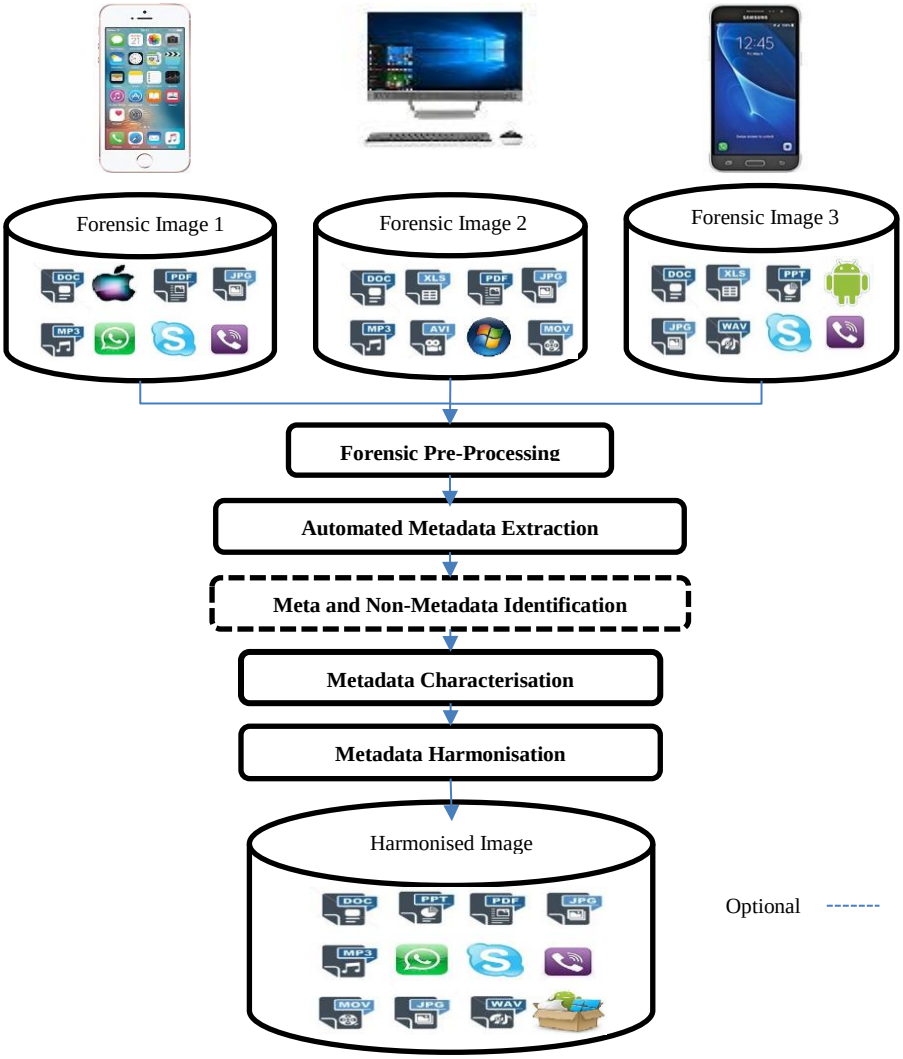

Figure 3: Overview of Proposed Process

5. Experimental Evaluation

5.1 Methodology

The purpose of the experiment is to evaluate and validate that the characterisation and harmonisation. The following aims are defined:

- To differentiate between metadata and non-metadata,

- To identify the metadata categories which are equivalent,
- To merge similar categories.

There is a need to access real forensic data to make the entire experiment more reliable. As such the experiments have been conducted using five images of real forensic data from multiple resources such as smartphones, computers, and external hard drives. The experiment considered these evidential resources as a single case to validate the proposed framework. In addition, to support the limited number of real cases, an artificial forensic image has also been used to validate the reliability and effectiveness of the algorithms. Therefore, six forensics images were provided. The public image (image 1) was generated by the National Institute of Standards and Technology (NIST) (NIST, 2015). This image is an artificial image describing the computer of a suspected person who tried leak sensitive information related to the newest technology in his company. The remained images were obtained from Iraq, and contain information of various crimes committed by convicted criminals. During the metadata extraction phase, various metadata were generated and extracted from these resources as illustrated in table 1 such as file systems and applications.

Table 2: Overview of Experimental Datasets

Id	Type	OS	Evidence Type				
			Messaging	Pics.	File List	Internet	Emails
1	Personal Computer	Microsoft Windows	-	EXIF	NTFS	Chrome, Mozella	Outlook
2	Smart phone	Android	SMS	EXIF	Ext4	Samsung Internet Browser	-
3	External Hard Drive	-	-	EXIF	FAT32	-	-
4	Personal Computer	Microsoft Windows	-	EXIF	NTSF	Chrome	-
5	Personal Computer	Microsoft Windows	-	EXIF	NTSF	Chrome	Outlook
6	Smart Phone	Android	Viber, SMS	EXIF	Ext4	-	-

The metadata of these images was exported into individual Comma Separated Value (CSV) files. A number of CSV files contains missing metadata features within the same category because they have been extracted from heterogeneous resources. For instance, the EXIF metadata, which is extracted from smartphone datasets, has completed metadata features such as filename, timestamp, camera manufacturer and model, size of image file, size of the image (width x height), IOS, latitude, longitude, and GPS timestamp. The EXIF metadata within computer datasets, however, contains missing features such as IOS, latitude, longitude, and GPS timestamp. Similarly, the internet browsing metadata is differentiated across the forensic images based on platforms and applications. In computer images, there are two browsers (Firefox and Chrome) which they have features such as URL, visit count, visit timestamp, referrer URL, title, and profile. Whereas the smartphone browsers only have (URL, visit count, visit timestamp). The smartphones images contain SMS and Viber application, and both of them serve to send and receive messages. Many features between SMS and Viber are similar such as account number, sending timestamp, delivery timestamp, message body, status, seen, and recipient number; as well as they contain binary-based data such as opened, deleted, seen, etc. Regarding the file system, heterogeneous operating systems (OS) are included across these images, but most of these OS hold common features as file name, timestamp, size, etc. Likewise, the emails of two images include mutual features in addition to email body that represents as a non-metadata characteristic.

5.2 Results

All the metadata categories within the six images (a total of 22 disparate datasets) were provided to the system in a single instance. As illustrated in table 1, there are three categories (email, Viber, and SMS) containing non-metadata fields. Therefore, the meta from non-metadata identification based upon email, Viber, SMS categories was achieved successfully, and all non-metadata fields were automatically eliminated.

In order to identify the categories, the characterisation process was utilised to generate a record file. This record contains the categories that are similar as represented in section 3. To make it clear, the algorithm takes a dataset and checks it with all datasets in sequence. Then, it will count the number of identical fields (I) within

the compared datasets against different fields (D). There is a threshold used to decide whether the two datasets are similar or not. This threshold has been modified five times to obtain the ultimate threshold as shown in table 2. The experiment results prove that when the threshold of I is greater than or equal to D the best results can be obtained. Consequently, the algorithm creates the record which contains the similar files.

Table 3: Experimental results for 22 CSV files

	Threshold	True Positive (Merged Correctly)		False Positive (Merged Incorrectly)	
		Files Num.	%	Files Num.	%
1	I < D	0	0	22	100
2	I <= D	4	18.18	18	81.82
3	I == D	8	36.36	14	63.64
4	I >= D	22	100	0	0
5	I > D	14	63.64	8	36.36

Table 2 shows the impact upon the performance of characterisation algorithm across different thresholds. The worst results have been obtained when using the threshold of I less than D, where the algorithm matched the files which are completely different. When the threshold of I less than or equal D matched only four files properly with 18.18 % of the true positive. By using the equality threshold, the results were enhanced a little with only eight files matched out of 22 files, and this is still unacceptable. While the threshold of I greater than D showed a good rate of matching compared with aforementioned thresholds with 63.64 of the true positive. Ultimately, the threshold of I greater than or equal to D gave the best results with 100% of the true positive. Noticeably, this threshold might be changeable according to the nature of the study cases and their metadata categories.

To merge the similar categories, the harmonisation algorithm took the record file and the CSV files. The algorithm was able to merge and produce new five CSV files representing the main five categories. The main five categories were SMS and Viber together, EXIF, emails, file list, and Internet browsing metadata. In addition, the performance and the accuracy of this algorithm completely depend on the record which is generated by the characterisation algorithm. Accordingly, it merges and harmonises the similar categories together in one file. Although the results of this algorithm are encouraging, there are some errors detected due to the only the binary-based data that exists within the Viber and SMS categories. Only two fields of binary data within each category were wrongly merged. These were the seen field merged with the deleted field, and the read field merged with a hidden field. However, the binary data represents with only two values: 0 or 1 and does not contain valuable information compared with other fields of SMS and Viber categories.

6. Conclusions & Future Work

The evidentiary nature of digital forensics has changed over the years and cases increasingly contain Multiple devices and applications. Existing digital forensic tools are struggling to keep pace in achieving modern forensic investigations such as examining and analysing many systems and applications at once. Therefore, this paper has proposed and demonstrated an automated approach for metadata characterisation and harmonisation to overcome the heterogeneity issues. In the experimental study, the live forensic data has been utilised to evaluate the novel process. The results have shown that the characterisation and harmonisation process can be appropriated to merge and create a common standard across different formats for a similar metadata category. Although the harmonisation algorithm has not been able to merge all binary data fields, the binary data has the minimal valuable information within the investigation process. Future research will focus upon developing the harmonisation process to make it more accurate by using an intelligent procedure to merge the similar fields. A further evaluation also requires to be undertaken upon wide range of technologies and applications to make the characterisation and harmonisations algorithms more generalise in practice.

References

Al Fahdi, M., Clarke, N. L., Li, F. & Furnell, S. M. (2016) 'A suspect-oriented intelligent and automated computer forensic analysis'. *digital investigation*, 18 pp 65-76.

Almunawar, M. N., Anshari, M. & Susanto, H. (2018) 'Adopting Open Source Software in Smartphone Manufacturers' Open Innovation Strategy'. *Encyclopedia of Information Science and Technology, Fourth Edition.* IGI Global, pp 7369-7381.

Aronson, J. E., Liang, T.-P. & Turban, E. (2005) *Decision support systems and intelligent systems.* Pearson Prentice-Hall.

Ayers, D. (2009) 'A second generation computer forensic analysis system'. *digital investigation*, 6 pp S34-S42.

Bennett, D. (2012) 'The challenges facing computer forensics investigators in obtaining information from mobile devices for use in criminal investigations'. *Information Security Journal: A Global Perspective*, 21 (3). pp 159-168.

Case, A., Cristina, A., Marziale, L., Richard, G. G. & Roussev, V. (2008) 'FACE: Automated digital evidence discovery and correlation'. *digital investigation*, 5 pp S65-S75.

Dash, P. & Campus, C. (2014) 'Fast Processing of Large (Big) Forensics Data'.[in. (Accessed:Dash, P. & Campus, C.

Garfinkel, S. L. (2006) 'Forensic feature extraction and cross-drive analysis'. *digital investigation*, 3 pp 71-81.

Guptill, S. C. (1999) 'Metadata and data catalogues'. *Geographical information systems*, 2 pp 677-692.

Mohammed, H., Clarke, N. & Li, F. (2016) 'An Automated Approach for Digital Forensic Analysis of Heterogeneous Big Data'. *The Journal of Digital Forensics, Security and Law: JDFSL*, 11 (2). pp 137.
NIST. The CFReDS project. 2015. https://www.cfreds.nist.gov/data_leakage_case/data-leakage-case.html
Raghavan, S. (2014) *A framework for identifying associations in digital evidence using metadata*. Queensland University of Technology.

Rowe, N. C. & Garfinkel, S. L. (2011) 'Finding anomalous and suspicious files from directory metadata on a large corpus', *International Conference on Digital Forensics and Cyber Crime*. Springer, pp. 115-130.

Tanenbaum, A. S. (2009) *Modern operating system.* Pearson Education, Inc.