E-mail Forensic Authorship Attribution

by

Mr Himal Lalla

200506428

Dissertation

submitted in fulfilment of the requirements for the degree

Master of Commerce

in

Information Systems

in the

Faculty of Management and Commerce

of the

University of Fort Hare

December 2010

Supervisor: Prof. Stephen Flowerday

Abstract

E-mails have become the standard for business as well as personal communication. The inherent security risks within e-mail communication present the problem of anonymity. If an author of an e-mail is not known, the digital forensic investigator needs to determine the authorship of the e-mail using a process that has not been standardised in the e-mail forensic field. This research project examines many problems associated with e-mail communication and the digital forensic domain; more specifically e-mail forensic investigations, and the recovery of legally admissible evidence to be presented in a court of law.

The Research Methodology utilised a comprehensive literature review in combination with Design Science which results in the development of an artefact through intensive research. The Proposed E-Mail Forensic Methodology is based on the most current digital forensic investigation process and further validation of the process was established via expert reviews. The opinions of the digital forensic experts were an integral portion of the validation process which adds to the credibility of the study. This was performed through the aid of the Delphi technique.

This Proposed E-Mail Forensic Methodology adopts a standardised investigation process applied to an e-mail investigation and takes into account the South African perspective by incorporating various checks with the laws and legislation. By following the Proposed E-mail Forensic Methodology, e-mail forensic investigators can produce evidence that is legally admissible in a court of law.

Declaration

I Mr Himal Lalla, hereby declare that:

- The work in this dissertation is my own work.
- All sources used or referred to have been documented and recognised.
- This dissertation has not previously been submitted in full or partial fulfilment of the requirements for an equivalent or higher qualification at any other recognised educational institution.

Mr Himal Lalla

Acknowledgements

I would like to acknowledge various individuals and organisations for all the support shown throughout this project. Without their support and understanding, this research project would never have materialised.

I would like to thank my supervisor, Professor Stephen Flowerday, for the advice, encouragement, guidance and words of wisdom whenever I required them.

I would like to extend my gratitude to the University of Fort Hare, Department of Information Systems for the support it has provided through the duration of my academic career. I would like to extend special mention to my fellow students for their advice and support.

I would like to acknowledge the University of Fort Hare and its associated research body, The Govan Mbeki Research and Development Centre, for the financial support that they provided to me throughout this research project. I would also wish to thank the university library and its staff for their support and help when needed.

I would like to thank each of the expert reviewers who took the time out of their busy schedules to participate in this research project and provide valuable feedback.

Finally, I would like to thank my family, my girlfriend and my friends for the understanding and support shown to me through this time consuming endeavour.

Himal Lalla

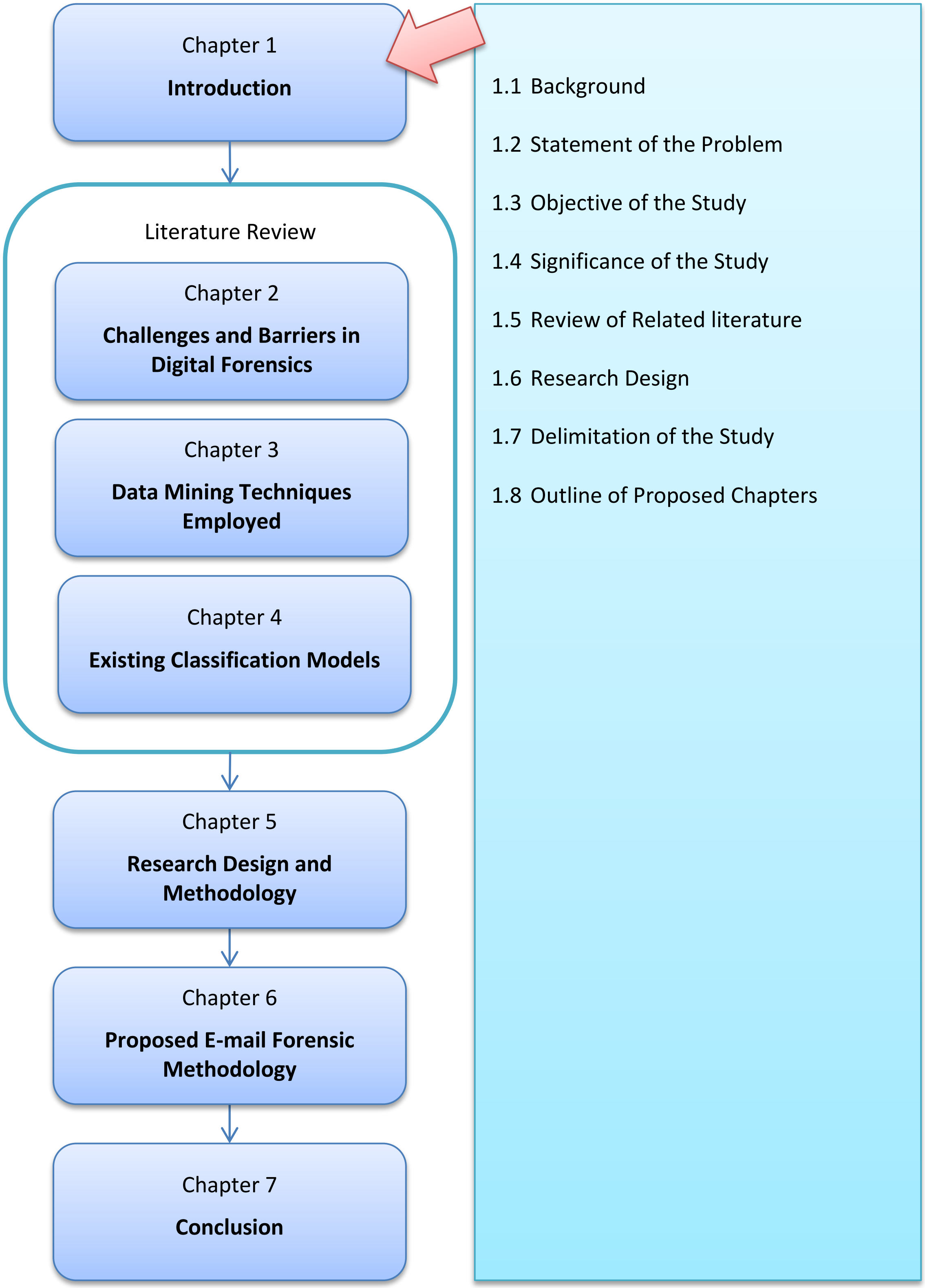
Table of Contents

List of Figures

List of Tables

1. INTRODUCTION

CHAPTER 1



Chapter 1

Introduction

Literature Review

Chapter 2

Challenges and Barriers in Digital Forensics

Chapter 3

Data Mining Techniques Employed

Chapter 4

Existing Classification Models

Chapter 5

Research Design and Methodology

Chapter 6

Proposed E-mail Forensic Methodology

Chapter 7

Conclusion

1.1 Background

1.2 Statement of the Problem

1.3 Objective of the Study

1.4 Significance of the Study

1.5 Review of Related literature

1.6 Research Design

1.7 Delimitation of the Study

1.8 Outline of Proposed Chapters

1.1 Background

We are living in the information age (Cardwell, et al., 2007) and important documents and information are stored on digital media and computer systems that have been incorporated into many institutions to improve efficiency and productivity (Reith, Carr, and Gunsch, 2002). Many companies and institutions use e-mail as a method to communicate and conduct business, and this has led to an increase in e-mail traffic volume due to the advent of the World Wide Web (WWW) (Iqbal, Hadjidj, Fung, and Debbabi, 2008). This large amount of e-mail traffic has opened the door to abuse by criminals and terrorists because primarily they can remain anonymous (Lim, 2008).

The anonymity factor of e-mail has made it difficult for digital forensic investigators to identify the authorship of an e-mail, and to compound this problem further, digital forensic investigators do not have a standardised procedure to follow (Ieong, 2006). This study examines the question of authorship attribution of e-mails and the need for digital forensic investigators to provide credible evidence. According to Lim (2008), the primary reasons for e-mail misuse is due to: the availability of e-mail; it is relatively inexpensive; it allows the same message to be sent to many users; and it allows users to remain anonymous i.e. when creating new accounts. Iqbal, Hadjidj, Fung and Debbabi (2008) suggest that the abuse of e-mail through anonymity is as a result of server routing which hides the e-mail origin. Therefore, the problem arises as to authorship of the e-mail.

To address this issue digital forensic investigators have to follow a number of steps that are part of a process and in the opinion of Eloff, Kohn, and Olivier (2006), the number of forensic models has added to the complexity of the field. Therefore, this has led to a call for standardisation in the field of digital forensics and in Leigland and Krings' (2004) view, this hinders the investigation process. Notwithstanding, Ieong (2006) states that a few procedures from different authors are known to be the 'standard' procedures in digital forensic investigations. However, he also notes that there are a number of discrepencies.

In the opinion of Casey (2004), the lack of rules results in incomplete evidence collection and errors in evidence interpretation. To add further to this complexity, the legal foundation which is evolving, restricts digital forensics (Ryan and Shpantzer, 2005). Therefore there is a need, firstly to standardise the process and secondly to comply with the law when collecting evidence in order for it to be legally admissible in a court of law.

1.2 Statement of the Problem

With the advantages that e-mail has brought to the work place and to individuals, it has also introduced the threat of the genuine user being compromised (Gupta, Mazumdar, and Rao, 2004). In order for digital forensic investigators to determine if a genuine user has been compromised, many tools and techniques need to be utilised to determine the authorship of an e-mail.

Digital forensic investigators need to be able to provide credible proof in order to prosecute a suspect in a court of law; however, the burden of authorship attribution lies with the digital forensic investigator. The problem is that if an author of an e-mail is not known, the digital forensic investigator needs to determine the authorship of the e-mail using a process that has not been standardised in the e-mail forensic field. Thus, this complicates the regulatory compliance issues making it difficult to provide legally admissible evidence in a court of law.

In order to address the main research problem the following sub-questions will be investigated and these form the objectives of the study.

1.3 Objectives of the Study

The main objective of this research project is to produce a methodology that will aid a digital forensic investigator in determining authorship of an e-mail and produce legally admissible evidence in a court of law in the process.

1.3.1 What are the challenges faced by digital forensic investigators in conforming to the law with respect to presenting legally admissible evidence?

The purpose of this sub-question is to establish the challenges that digital forensic investigators face in order to provide legally admissible evidence in a court of law.

1.3.2 How can data mining techniques aid in the attribution of authorship of e-mails?

This sub-question addresses the techniques used in digital evidence recovery, and how the digital forensic investigator arrives at determining authorship of an e-mail.

1.3.3 What are the classification models used and how do they assist in the verification of evidence?

It is necessary to understand how the digital forensic investigator arrives at presenting the collected evidence and how the models contribute to the integrity of the evidence.

1.4 Significance of the Study

This study is important as it addresses the need to provide credible evidence that will be used in a court of law to convict a suspected criminal. This needs to be done by verifying the authorship of e-mail. However, this will not provide credible evidence in all criminal cases because sometimes investigators find no evidence that can help a case. As some criminals will not use or know how to use e-mail, e-mail evidence might not be relevant in all attempts to convict suspect criminals. Furthermore e-mail related evidence is not only applicable in criminal case but also in civil cases. Additionally, there is no best practice or generalised standards in the field of digital evidence investigations; hence this study is necessary.

1.5 Review of Related Literature

This research project will refer to the Diffusion of Innovations Theory. This theory postulates that the innovation adoption process is one of information gathering and uncertainty reduction (Agarwal, Ahuja, Carter and Gans, 1998). According to Rogers, Singhal and Quinlan (2007) diffusion research is distinctive due to the communication messages that individuals perceive as “new.” Therefore, this is the reason for the high uncertainty in information gathering that individuals experience.

Rogers utilised Vernon’s Product Life Cycle model, which has an S-shape pattern, to assert that the diffusion process has an S-shaped curve (Wonglimpiyarat and Yuberk, 2005). This curve gives rise to a bell shaped distribution of adopters that Rogers employs to differentiate between five categories of adopters ranging from “innovators” to “laggards” based on their time taken to adopt the innovation (Agarwal, et al. 1998). Thus with the call for standardisation in the field of digital forensics, investigators will fall into the different categories of adopters. Creating a new process by which e-mail forensic investigations should be conducted reduces the high degree of uncertainty inherent in all e-mail forensic investigations. The following sections aim to address the main objective of the study.

1.5.1 Challenges Faced by Digital Forensic Investigators

According to Taylor, Haggerty and Gresty (2009) e-mail investigations were mainly undertaken by law enforcement agencies; however in the United Kingdom a wide variety of organisations have now begun to do so. These organisations have appointed teams including a forensic expert in order to overcome the skills shortage (Taylor et al. 2009). This poses a challenge for the digital forensic investigator as the environment in which the recovery of evidence is performed is not ideal.

This issue is further compounded, as many organisations underestimate the admissibility and reliability requirements of digital evidence placed on them by the legal system (Kent and Ghavalas, 2005). Thus, the policies and procedures in place are not adequate enough to provide legally admissible evidence in a court of law.

In order for evidence to be legally admissible, it needs to be compliant with the applicable legislation. South African law has its origins in Roman law founded centuries ago, thereby making it difficult to cope with the advances in technology. This has constrained traditional methods of investigation and prosecution of crimes (Maat, 2009). Therefore, there are loopholes in the law for criminals to exploit with respect to the Internet.

In August 2002, the Electronic Communications and Transactions Act 25 (ECT Act) became law. This law was developed to govern e-commerce in South Africa, and it applies to any form of electronic communication i.e. e-mail, Internet, SMS, etc. (Michalsons, 2005a). One of the primary issues that the ECT Act seeks to address is the illegal activities of cyber criminals.

The ECT Act includes the creation of new “Cyber Offences” creating certain provisions for cyber inspectors (Michalsons, 2005a). The crux of the ECT Act with regard to e-mails, is that the ECT Act permits electronic documents and e-mails as evidence; however there is a requirement to show authenticity and integrity of the information (Michalsons, 2005b).

Another critical challenge that digital forensic investigators face is that of the forensic tools available at their disposal. These tools have a short life span and as a result they are not able to keep up with current investigations (Ayers, 2009). According to Ayers (2009) tools such as Encase and FTK products have been around for over a decade and their limitations, such as processing speed and software errors, have been discovered during this time.

In order for this challenge to be overcome, new tools will need to be developed to cope with the demands of the latest investigations. Arthur and Venter, (2004) are of the opinion that while the current tools have the ability to discover for instance, all important system files, they are in need of attention because they are not guarenteed to recover unreferenced files. Thus new tools are needed in the field in order to maintain the level of reliability and admissibility required in a court of law.

1.5.2 Data Mining Techniques

Digital forensic investigators have a myriad of techniques and tools at their disposal to perform the analysis of a large amount of digital information stored on digital media. E-mail has become the new form of communication for millions of people and the anonymous nature of e-mail has made authorship identification a problem. Hadjidj, Debbabi, Lounis, Iqbal, Szporer, and Benredjem (2009) suggest that while there is no proactive mechanism to protect e-mail, there are however techniques that can be used to determine authorship of e-mails, and one such technique is literary stylometry which is the determination of authorship from writing styles (Corney, Anderson, Mohay and De Val, 2001).

Corney et al. (2001) have used stylometry in conjunction with a learning machine technique called Support Vector Machine, which is based on the structural risk minimisation principle, to determine authorship. However, this approach will not yield admissible evidence in court cases. Notwithstanding this, Zheng, Li, Chen and Huang (2006) use stylometry in a framework, which they have developed to address online messages specifically, to identify the author of such messages. The process can be divided into four steps, namely: message collection; feature extraction; model generation and finally author identification. Hence the use of stylometry in conjunction with other techniques can be useful in the identification of an unknown e-mail author.

Hadjidj et al. (2009) have developed a framework based on a combination of established techniques including statistical analysis, text mining and stylometry together with social networking techniques. The authorship attribution occurs in two steps; the first e-mail grouping is conducted using content based and stylometry based clustering of the data; the second step is the classification phase which entails feature extraction from the body of the e-mail followed by model generation and model application (Hadjidj et al., 2009).

Iqbal et al. (2008) have proposed a novel method of data mining called AutoMiner. AutoMiner is an algorithm which can determine authorship of an e-mail by extracting frequent patterns from the e-mail and comparing it to a write-print of a suspected individual. Iqbal et al. (2008) point out that they do not claim that the write-print can uniquely identify an individual; instead they believe it is accurate enough to identify an individual from a list of suspects. The problem that digital forensic investigators are facing is that with all the information that a computer can store for one e-mail message, they are merely looking for some trace of evidence to indicate authorship, but without techniques such as Autominer, the digital forensic investigators have a difficult job finding that evidence.

1.5.3 Classification Models

The explosion of growth that technology and in particular the computing world has experienced, has resulted in highly sophisticated equipment. This has in essence intensified the criminals' potential to perform criminal activity (Reith, Carr and Gunsch, 2002). In light of this, law enforcement agencies have been busy trying to keep up with the criminal element that is persistent in abusing technology. In order for digital forensic investigators to perform their job, there are a number of steps that need to be well thought out and dealt with (Eloff et al. 2006). These steps are encompassed in digital forensic models.

Digital forensic investigators follow a generalised methodology when conducting an investigation to ensure credibility and integrity of the digital devices (Arthur and Venter, 2004). The methodology followed is a stepwise process and is as follows: protect; discover; recover; reveal; access; analyse; print and provide consultation (Arthur and Venter, 2004). This method is a sequential process and Arthur and Venter believe that while this method is a strict process, it ensures the integrity of the evidence. All digital forensic investigators use a variation of this method, although the overall method is similar.

Cardwell et al. (2007) have divided digital forensics into three categories, namely litigation support, digital media analysis and network investigations. The first step, litigation support is the process of identification, collection, organisation and presentation of digital media while the second and third processes deal with the specific types of digital media (Cardwell et al., 2007). Thus in the first step of this model, the processes 'identification' and 'collection' are similar to Arthur and Venter's (2004) 'discover' and 'recover' processes.

While Cardwell et al. (2007) deal with the methodologies in a practical way i.e. by detailing the steps in the different categories, other methodologies include similar principles. One such methodology is the United States Department of Justice Forensics (USDOJ) process model.

This model consists of four phases namely collection, examination, analysis and reporting (Cardwell et al., 2007). The model is an abstract model not specific to any technology or methodology and therefore is a generalised process, focusing mainly on the core aspects (Reith et al., 2002). Hence this model will be more applicable in digital investigations as it can be adapted to the technology under examination.

Kruse and Heiser's methodology includes three components that ensure the integrity of the evidence while investigating; these components are: acquiring the evidence; authenticating the evidence; and analysing the data (Eloff et al. 2006). There are a number of frameworks and methodologies that cover the digital forensic investigation differently; the (USDOJ) and Kruse and Heiser's methodology are the most commonly referred to in literature but this adds to the complexity of the digital forensic process (Eloff et al. 2006). Therefore the need for a standardised process has become more evident.

Ieong (2006) names Lee; Casey; Reith, Carr and Gunsch as the most frequently quoted authors and states that their procedures are known to be the 'standard' procedures used during investigations. Ieong (2006) presents a framework called FORZA (FORensics Zachman framework) that links all the common procedures as well as binds roles and reponsibilities of eight different individuals involved in the investigation process. The FORZA framework will be discussed in-depth in the literature review chapters.

Leigland and Krings (2004) point out that there are limitations that exist in the process models. There were four deficiencies found at a digital forensic research workshop in 2001, procedural, technical, social and legal (Leigland and Krings, 2004). The aim of Leigland and Krings' (2004) research is to formalise the forensic process by providing a tool to manage the procedures followed when dealing with a compromised computer system. Meyers and Rogers (2004) believe that it is the scientific community's responsibility to standardise procedures and to certify individuals with a formal educational process.

Whilst these models are being proposed and used in the digital forensic field, there is no best practice or standardisation of the procedures followed; thus many of the models are guidelines developed ad hoc for performing investigations (Leigland and Krings, 2004). This highlights the importance of the standardisation of procedures and techniques used.

1.6 Research Design

The study follows the Design Science research methodology, including research methods i.e. data collection and data analysis. This study includes empirical research as well as a literature review comprised of secondary data that will include theories, models and frameworks. All attempts are made to keep the content as current as possible and this will form the theoretical base of the study.

Design Science is technologically orientated and is essentially a problem solving process that leads to the development of an effective artefact, which is of four types: namely constructs; methods; models and implementations (March and Smith, 1995). Hence this research output is a methodology developed through intensive research.

1.6.1 Design Science

Design Science is a research paradigm that attempts to expand human potential by creating ground breaking artefacts. Hevner and March (2003) have proposed a conceptual research framework in order to assist understanding, execution and evaluation of Information Systems research. This framework is used to assess what is being produced from each paradigm against each other in the context of business needs.

Peffers et al. (2006) propose that a conceptual process and mental model, called the Design Science Research Process shown in Figure 1 - 1, be used for the production and presentation of Design Science research.

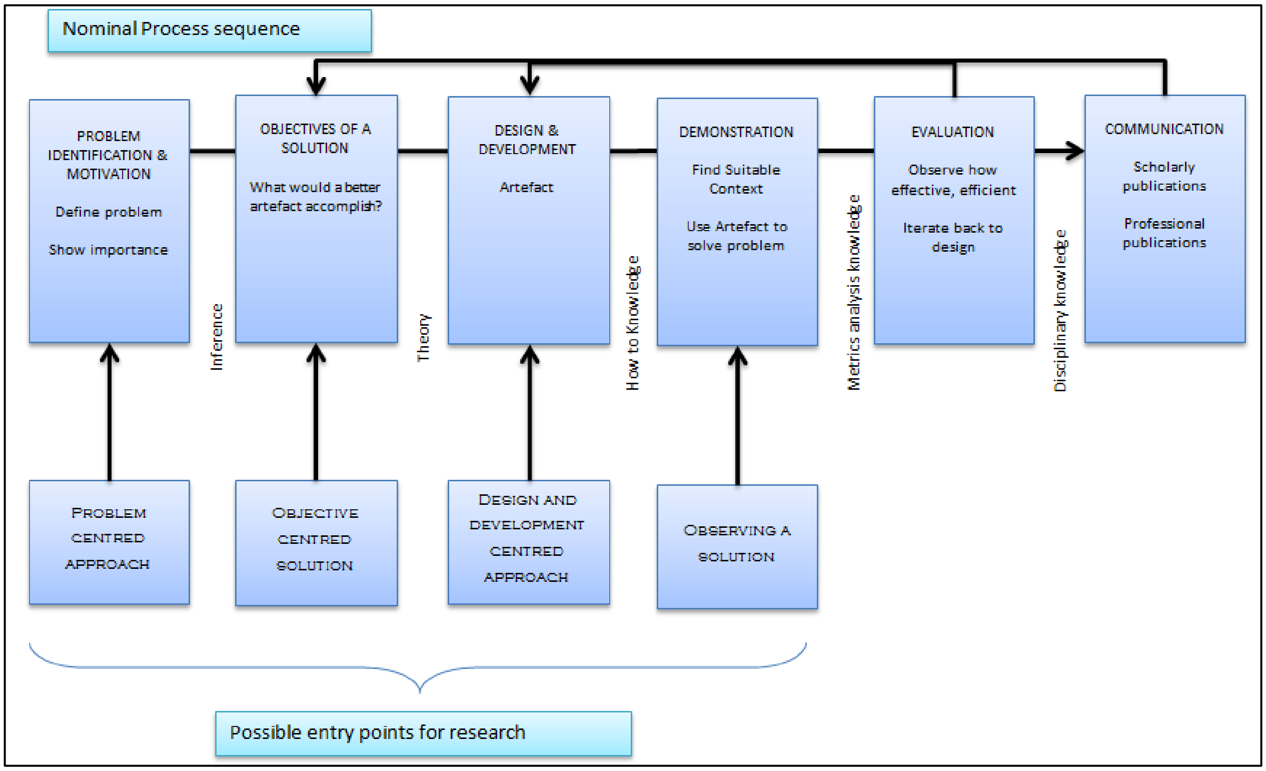


Figure 1 - 1 Design Science Research Process (Peffers et al. 2006)

1.6.2 Data Collection Methods

This is the method of collecting data, and this study utilised secondary data and primary data with experts in the field.

I. Primary Data Collection Methods

- Qualitative Data:

An informal survey was used to obtain feedback from digital forensic experts in this field in order to further refine the proposed model. These 10 experts were presented with the study's findings and they were asked to reflect and comment on the findings as a step to further refining the methodology.

1.6.3 Data Analysis Methods

Recommendations are made based on the research findings. The qualitative data from the experts was summarised and changes were made according to their feedback and as a further stage of refining the proposed solution. Their feedback either supported or opposed the proposed solution and this added to the integrity of the project.

1.7 Delimitation of the Study

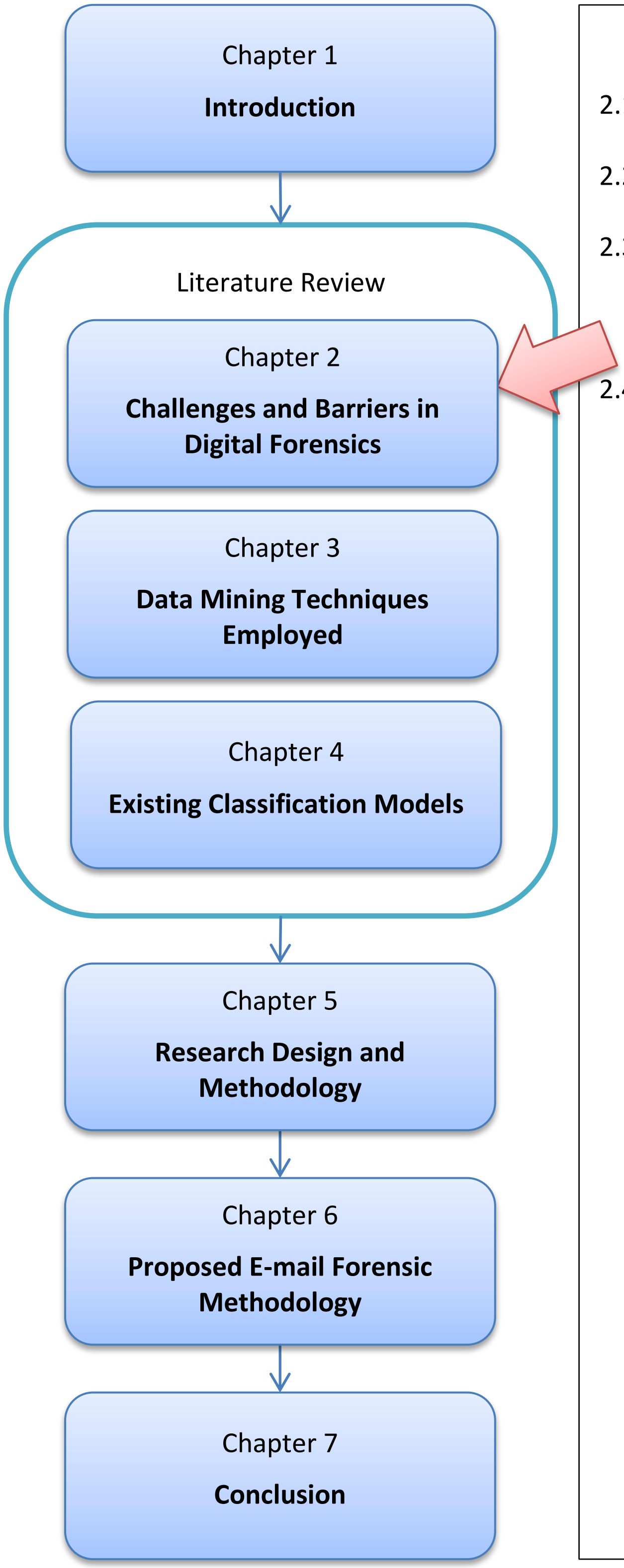
This study will be restricted to the recovery of digital evidence from e-mail and the verification of authorship of the e-mail. Existing literature formed the theoretical base from which a methodology of e-mail authorship verification was produced. The study was also limited to South African law; however, it drew from other countries' laws and frameworks.

1.8 Outline of Proposed Chapters

Chapter one is the Research Proposal and provides a brief introduction to the general area of research and clearly define the research problem investigated. The significance of the study is given as well as the research methodology that will be followed. Chapter two, three and four are the Literature Review chapters, a critical analysis of the existing literature, and addresses the main research objective and the sub-problems associated thereof. Chapter five presents the Research Design and Methodology including data collection and data analysis techniques. Chapter six is the Proposed E-Mail Forensic Methodology and is a combination of the strengths of the different methodologies examined. Included in Chapter six are the Research Findings and Recommendations and the analysis of the results. Chapter seven is the summative conclusion followed by suggestions for future research.

2. CHALLENGES AND BARRIERS IN DIGITAL FORENSICS

CHAPTER 2


Chapter 1
Introduction
Literature Review
Chapter 2
Challenges and Barriers in Digital Forensics
Chapter 3
Data Mining Techniques Employed
Chapter 4
Existing Classification Models
Chapter 5
Research Design and Methodology
Chapter 6
Proposed E-mail Forensic Methodology
Chapter 7
Conclusion

2.1. Introduction

2.2. Categorisation of Existing Challenges

2.3. Additional Challenges to the Proposed Categories

2.4. Conclusion

2.1 Introduction

In Chapter one an outline of the study was provided and a brief literature review was undertaken. The main objective of the research was identified; the research sub-questions assist in addressing the research objective. The first research question attempts to address the challenges faced in the digital forensic investigation process; the second research question attempts to narrow the data mining techniques of digital forensic investigators by focusing on e-mails; and the third research question examines the classification models that create the digital forensic methodology. In the literature review a brief description of the problem was uncovered. However the problem of e-mail authorship attribution runs deeper and hence a comprehensive literature review was performed. The literature examined was secondary data sources comprising of journal articles, academic papers, books, and various other articles, and the Internet resources.

Embedded in this secondary data are models, methodologies and frameworks that were examined and these contribute towards an e-mail forensic methodology in the later chapters. In the preliminary literature review various sub-problems were established. The first sub-question of the challenges facing digital forensic investigators was included in the preliminary literature review; however a thorough examination of the literature was required to establish the current challenges within the domain of digital forensics. There are a number of frameworks, models and methodologies but there is a lack of agreement as to the challenges facing the digital forensic investigator; this can be explained by the growing number of challenges and proposed solutions (Broucek and Turner, 2002).

In this Chapter the first sub-question is examined. The various challenges that digital forensic investigators are faced with are examined and updated, in order to give a clear view of the scope and magnitude of the challenges that need to be overcome in the discipline. The main purpose of this Chapter is to establish the foremost challenges that confront digital forensic investigators as well as determine the set of challenges that need to be addressed in the domain of digital forensics. These challenges are not new, by any means, however they are ever present and evolving. As discussed earlier the digital forensic challenges were first documented and categorised at the first Digital Forensic Research Workshop (DFRWS) in 2001 by Dr Eugene Spafford (Palmer, 2001). This was done in a drive to establish the digital forensics domain as a new discipline.

These challenges gave rise to the need for standardisation of a digital forensic investigation process and as a result a broad framework was developed in order to describe the entire computer forensic investigation (Leigland and Krings, 2004). The following section presents the various categorisations of the challenges that have been put forward by different researchers. From this a new categorisation of these challenges will be made in order to cover a broader spectrum of challenges as well as to include any new challenges encountered in the domain of digital forensics due to the progression of the field. This is followed by a discussion of the challenges that fall into those categories highlighted by researchers. Finally, a conclusion is given with respect to the challenges in the digital forensic domain.

2.2 Categorisation of Existing Challenges

In order for the field of digital forensics to grow and develop, the various challenges that prohibit growth must be overcome. Over the years many researchers have attempted to classify the challenges and some have successfully identified the crucial ones. However the digital field evolves and progresses at an alarming rate due to technological influences. A number of challenges have been identified in the earlier literature review; thus this section includes a breakdown of various views and provides a baseline on which the new categorisation of the challenges will be based. Two prominent researchers' views on the categories will be examined and this will be compared as well as criticised in order to establish a new consolidated categorisation.

2.2.1 Spaffords Categorisation

The challenges were documented at the first Digital Forensic Research Workshop in 2001 and Spafford categorised the challenges into four types, namely technical, procedural, social and legal (Palmer, 2001). Spafford emphasised these four categories as the full spectrum approach resulting from academic research in support of government concentration on technological results. Spafford states that the research must address these four categories in order to “heal” rather than “treat” the digital problems (Palmer, 2001). Therefore, this is the baseline for the categorisation of the challenges and a brief description of each field as set by Spafford is given.

The **Technical** challenge includes issues dealing with the rate of progression of technology and the constant struggle to stay abreast. Spafford notes that the two main issues; terabyte disks and the time to market, are the reasons that cause investigators difficulty in applying analytical tools (Palmer, 2001); however, according to a recent study performed by the market research firm International Data Corporation (IDC), hard disk drive shipments are expected to soar by 12 million units, representing 300 000 petabytes (approximately 10^{15} bytes) by 2014 (Mearian, 2010). Currently the standard disk drive is a terabyte (Jordaan, 2010). Therefore, the growth in technology increases the challenge for digital forensic investigators; as technology grows and creates more storage space for various users, the digital forensic investigators will have a larger amount of data to analyse.

The expected growth and increase in digital media offered represents a shorter time to market hard disk drives due to various factors i.e. reduced production costs and increase demand by enterprise class companies (Mearian, 2010). Therefore the time for a given product to reach the market is decreasing and hence this represents a twofold challenge in terms of the growth in technology and the reduction in the time to market. Decreasing costs contribute further to the problem of increased storage space on a medium as many enterprises will leap at the opportunity to increase the space requirement at a relatively cheaper price. Therefore the digital forensic discipline must cater for this growth in technology and adapt accordingly.

Spafford also states that the unknown level of trust in development tools and the widespread lack of experience and training in the digital forensic field contribute further to this difficulty (Palmer, 2001). Therefore digital forensic investigators are faced with the challenge of applying tools, where an unknown level of trust exists, to massive amounts of data stored on large volume hard disk drives. Hence, this first challenge is a compound challenge as technological problems need to be addressed holistically and not in isolation of each other.

The **Procedural** challenge as described by Spafford is that digital forensic investigators must collect everything that in the digital world leads to examination and scrutiny in support of investigations (Palmer, 2001). Additionally the protocols and procedures in the digital forensic domain are not standardised; moreover the researchers do not use standard terminology (Palmer, 2001). Hence with the technological challenge established, the increasing hard disk size means that more data can be stored and therefore the recovery of data becomes more challenging if there are no standardised protocols and procedures.

Spafford indicates that in the **Social** challenge, individual privacy and the collection and analysis needs of investigators collide continuously and the uncertainty about the accuracy and efficacy of the techniques used causes data to be stored for long periods of time and this therefore utilises resources that can be better used to solving real problems instead for storage (Palmer, 2001). This challenge is a sensitive one since individual rights and privacy laws are well established and therefore the digital forensic process must accommodate these.

Spafford explicitly states that even the most advanced technology created is debatable if it does not comply with the law, and this constitutes the Legal challenge (Palmer, 2001). Hence the **Legal** challenge is the hurdle that needs to be crossed in order for all other challenges to be met. Spafford's views on the challenges indicate the number of problems the discipline must overcome; however, although Spafford's challenges appear holistic, it is necessary to examine other views on these challenges. The following section addresses Rogers's categorisation of the challenges.

2.2.2 Rogers Categorisation

Rogers (2005) identifies five areas of challenges. These challenges are **Public Policy**, **Technology**, **Scientific**, **Resources** and **Legal** requirements; however, Rogers (2005) indicates that the four challenges lead to the legal challenge. This is depicted in Figure 2 - 1 and hence the Legal challenge, termed “legal requirements’ is at the centre of all the other challenges. Rogers (2005) depicts the challenges in this manner because in each challenge there are legal requirements that must be fulfilled, and thus highlighting the legal challenge. Investigators search and seizure efforts impinge on individual rights and freedoms, and this forms part of the privacy issue (Rogers, 2005). Rogers (2005) also notes that the efficacy of current laws must be examined and there needs to be a ‘harmonisation’ of international laws as privacy laws are different between countries. Hence, these legal and privacy issues comprise the **Public Policy** challenge.

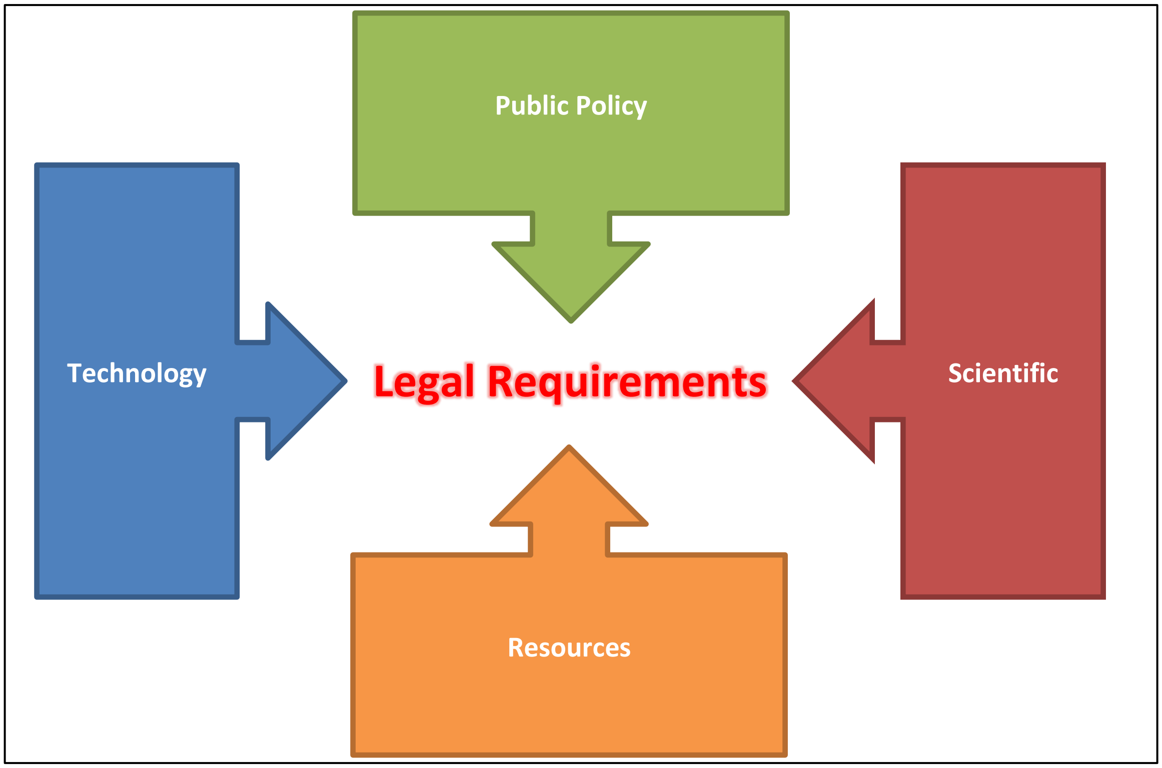


Figure 2 - 1 Rogers Categorisation of Challenges (Rogers, 2005)

The **Scientific** challenge is the theoretical foundation of the digital forensic discipline (Rogers, 2005). The contributing disciplines need to be broadened. There also needs to be a scientific governing body to provide guidance (Rogers, 2005). Therefore, the focus should be on building the knowledge base of the discipline in order to address the Scientific challenge.

The challenge of **Resources** comprises of a lack of trained law enforcement and the private sector therefore a skills gap exists (Rogers, 2005). This is due to a lack of training and education of individuals as well as a standard certification, which is required in other disciplines e.g. Certified Information System Auditor (CISA), Certified Public Accountant (CPA) and Chartered Accountant (CA), that is lacking in the digital forensics discipline. Rogers (2005) also highlights that lawyers and judges lack knowledge of the digital forensic discipline and this contributes to the gap in the law where legally admissible evidence is subjected to scrutiny in a court of law. Hence, up skilling of the digital forensic discipline with the aid of standard certification is needed as well as increasing the knowledge base by involving individuals from other disciplines i.e. Legal discipline. Consequently, the Resource challenge feeds into the Scientific challenge by recognising the need for interdisciplinary participation.

The **Technological** challenge includes issues such as non-certified tools which contribute to the lack of standardisation in the digital forensic discipline and Rogers states that this is “placing the cart before the horse” (Rogers, 2005). Hence there is a need to address the issue of tools currently utilised in the digital forensics field. Additional issues that form part of this challenge correspond with those highlighted by Spafford, i.e. the large storage devices available e.g. one terabyte (Rogers, 2005).

Another growing concern is the number of smaller storage devices that have come into the digital world e.g. Flash drives and iPods (Rogers, 2005). Rogers (2005) also lists the 50 plus operating systems being used as part of the Technological challenge as the tools used by digital forensic investigators need to accommodate different types of operating systems. Each tool developed does not necessarily accommodate all the operating systems, although most will, if not all, accommodate Windows operating system (Garfinkel, 2010). Hence the task of evidence recovery becomes more difficult as the number of operating systems increases and non-certified tools are used to perform the evidence recovery. Therefore, the complexity of the Technological challenge is ever increasing and this requires that the digital forensic discipline stay ahead of technological trends.

Rogers's fifth challenge is that of the **Legal** requirements. This forensic science discipline is all about legal admissibility; the main aim of digital forensics is to support an investigation and contribute towards conviction or exoneration of a suspect. Hence the discipline needs to be closely related to the law and satisfy the requirement of legality during an investigation. Rogers (2005) poses the question of what constitutes scientific evidence and using the Daubert Test to establish the criteria for scientific evidence. In order for the requirements to be met, the following question needs to be satisfied (Rogers, 2005):

- Whether the theory or technique has been reliably tested;
- Whether the theory or technique has been subject to peer review and publication;
- What is the known or potential rate of error of the method used; and
- Whether the theory or method has been generally accepted by the scientific community.

Chaski (2005) observes that the Daubert criteria has been accepted by federal and state courts in America although courts that do not use the Daubert criteria follow the Frye criteria. However, Chaski (2005) asserts that irrespective of which legal criteria is applied, from a scientific perspective, producing good 'normal science' in the Kuhnian sense will automatically meet legal requirements. Good normal science is empirical, replicable, utilises hypotheses for falsifiability and seeks knowledge criticism and review. Therefore the focus must be on the procedural aspects of the investigations and a standardised approach in both techniques used and appropriate tools utilised.

Nevertheless, the views of Spafford and Rogers appears to be similar. The challenges incorporate different areas of focus and in order to understand these a comparison must be performed between the different views that have been presented. The following section presents a comparison of both Rogers and Spafford's views on the categories.

2.2.3 Comparison and Criticisms of Current Categories

The categorisations presented are therefore unabridged because the challenges of the discipline are continually changing and growing ever more complex. One can gauge from

Spafford's Technical challenge and Rogers' Technology challenge that they represent the same challenge. However while Spafford addresses the issue of a lack of trust and experience of the tools in development, Rogers' issue of non-certified tools is more crucial because there is a need for standardisation in overall the digital forensic discipline.

The standardisation of tools is emphasised by Spafford although this has been included in the Spafford Procedural challenge. Spafford falls short in identifying the challenge of education within the digital forensic discipline whereas Rogers' Resource challenge identifies a lack of training and education standards in the digital forensic field. Hence the issue of standardisation cannot be separated from the Technological challenge; moreover this aspect cannot be met without addressing the educational aspect of the discipline.

Rogers' Scientific challenge addresses the issue of education as well as the theoretical foundation of the discipline; this is important as one cannot standardise tools and techniques in a discipline if there is no standardisation of education. Rogers also states that there should be standard certification as there is in the accounting and information technology fields i.e. CA and CISA. This would therefore increase the level of professionalism within the discipline but also create a process whereby all investigators attain a level of certification commensurate with their skills.

Rogers' Resource and Scientific challenges are closely related as they both involve education, an issue that needs to be addressed. Spafford falls short in identifying the need for education as a challenge, falling short of the standard of professionalism required in the digital forensic field.   Thus the challenge of education must be addressed in order to increase the credibility of the investigators in the discipline. The Social and Public policy challenges address the same issue of privacy and rights although Rogers again emphasises the law in terms of efficacy where as Spafford highlights the efficacy of the techniques used for recovery unfortunately leading to lengthy periods of storage. The awareness of the Law as a challenge is easily identifiable by researchers as this is the crux of all interrelated issues and challenges to which digital forensic investigators are exposed.

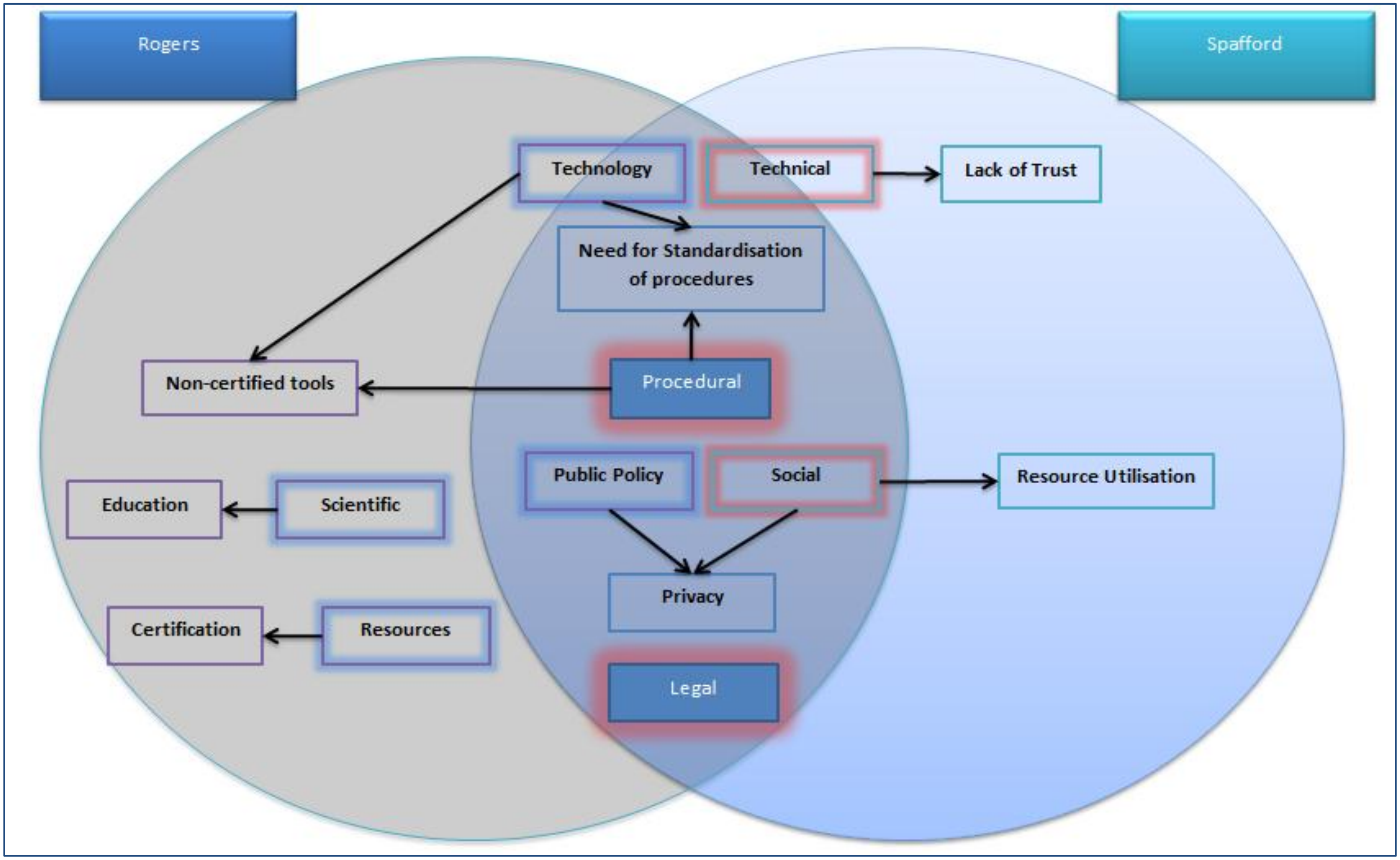


Figure 2 - 2 Similarity of Challenges (Own Compilation)

Rogers' approach of indicating that all the challenges lead to the legal requirement challenge is critical because every step during an investigation must follow the prescribed legal requirements in order to present evidence in a court of law. Hence Rogers' approach is more holistic than Spafford's. Figure 2 - 2 highlights the differences between the categories proposed by the authors. The circle on the left represents Rogers' view and the circle on the right represents Spafford's view; the section where the circles overlap shows the common ground between the views.

Both Rogers and Spafford identify the need for standardisation and these are incorporated in the specific challenges. This is the main aim of the discipline thus far. Addionally, the law is highlighted as a critical challenge and this is the crux of many investigations. The diagram also depicts the relationship between the various challenges and therefore it is difficult to separate each challenge into a specific category.

2.2.4 A Consolidated Categorisation

The challenges presented by Spafford and Rogers overlap each other to some extent and each author presents similiar views with respect to the individual challenges although Rogers approach to categorising the challenges is more holistic. Therefore from the challenges identified a wide categorisation can be made. Thus the categorisation of challenges are **Technological**; **Educational**; **Societal**; **Procedural** and **Legal**. The most important challenge though should be that of **Education**. This category must address the issues related to qualifications and the governing bodies overseeing those qualifications.

The Education challenge must also address the training of both investigators as well as developers of the tools, in an effort to foster trust in the tools utilised for evidence recovery. An additional issue is that of the contributing disciplines; digital forensics is a scientific discipline and therefore needs input from other disciplines such as the legal discipline. The contributing disciplines are wide and varied and is showcased by the legal discipline which plays an important role during digital forensic investigations. Therefore, there is a need for greater input from other disciplines and the digital forensic discipline needs to focus on building a stronger theoretical base from which acceptable standards can be derived. These standards must be created in order to further enhance the professionalism of the discipline.

The **Technological** challenge must address issues relating to the fast growth of technology. Digital forensic investigators need to stay ahead of new technologies and prepare for investigations that are more complex and difficult. The technological challenge must also address the issue of the tools used for digital evidence recovery; these tools need to be admissible in court. This is a difficult challenge and keeping up is a priority because an investigator cannot use out of date tools on new technologies e.g. newer operating systems may pose additional problems for the tool.

The **Societal** challenge must address the issue of privacy of individuals; therefore, this issue is closely related to the Legal challenge. There is a need for a governing body to be established to provide direction to the discipline. This will regulate the discipline accordingly and guidelines must be proposed so that indivdual and organisational privacy is respected. It is not feasible to regulate all investigations however if guidelines and procedures are in place; there is a recourse action that can be applied by both the discipline and by the law e.g. black list an investigator from performing investigations where serious criminal breaches occur.

The **Procedural** challenge must include issues relating to the recovery of evidence, the tools and techniques associated with that process as well as the legal aspects of the investigation. The tools need to have certification so as to allow standardisation, a criterion required by digital forensic investigators (Palmer, 2001). If tools are certified, digital forensic investigators should be encouraged to utilise the accepted tool before applying other non-certified tools. A process of standardisation also needs to take place with respect to the protocol and procedures followed. Investigators follow a generalised process; however each have variations in the process. Therefore, in order for standardisation to occur, the digital forensic discipline must address the tools used in the investigation process and the procedures followed during an investigation.

The **Legal** challenge is one that is encountered in all other challenges. However this challenge must be addressed as a separate challenge in an effort to ensure that the correct legal approach is taken during an investigation (Rogers, 2005). It is important that investigators are aware of the applicable legislation and laws in the jurisdiction that they reside in. These must be adhered to at all times. The legal challenge is deeply rooted in all processes of the digital forensic investigation process. Therefore the legal challenge must be addressed before, during and after an investigation.

Spafford and Rogers highlighted key challenges in the digital forensic field; however because of the technological nature of digital forensics, these challenges are ever present and newer challenges then present themselves as a result of technological advances. Therefore, the current categorisation of challenges needs to be revised in order to cater for new threats to the digital forensic field. As with many other issues, the currrent categories can be expanded to include the most influential challenges. The following section seeks to add to the current categories.

2.3 Additional Challenges to the Proposed Categories

The following section presents some of the additional challenges that digital forensic investigators encounter. Garfinkel (2010) argues that the “Golden Age” of digital forensics is ending and that without a clear strategy for research the digital forensics discipline will fall behind the market and furthermore, tools will become obsolete and hence law enforcement and digital forensic investigators will not be able to respond. The scope of challenges is wide and varying in difficulty. These challenges were categorised above as follows: Social and Environmental; Technical Expertise; Regulatory and Procedural.

The categories of challenges have been renamed in order to accommodate a broader spectrum. The challenges discussed below are in addition to the existing categories of challenges but not a complete list as stated previously; furthermore the very nature of digital forensic forces change and therefore these challenges must be revisited and reviewed.

Environmental challenges include the increasing number of users of computers amongst other variables. The pervasive nature of information and communication technology has led to an increase in electronic crime (Brody, Mulig and Kimball, 2007). As more users gain access to computers, more and more electronic crime will be perpetrated. Criminals have therefore become more resourceful in their attempts to lure users into a false sense of security and steal personal information i.e. phishing and pharming, a new crime that has surfaced since the Internet came into being (Brody, Mulig and Kimball, 2007). As criminals become more innovative, computer security must be enhanced; moreover users must remain a step ahead by educating themselves to take the necessary precautions.

The technical expertise challenge comprises of a skills shortage. E-mail investigations were mainly undertaken by law enforcement agencies; however, in the United Kingdom a wide variety of organisations have now begun to do so (Taylor, Haggerty and Gresty, 2009). These organisations have appointed teams which include a forensic expert in order to overcome the skills shortage; furthermore the teams are not composed of digital forensic individuals specifically and this poses a problem (Taylor, Haggerty and Gresty, 2009). The skills shortage is partly because computer related crime is still investigated (but not limited to) by law enforcement agencies and the key to closing the gap lies in building a comprehensive approach to forensic education (Yasinsac, Erbacher, Marks, Pollitt and Sommer, 2003).

Additionally, the education knowledge base has not yet been standardised to create standard certification. Moreover, Garfinkel (2010) identifies numerous researchers publishing and developing potentially useful tools; however despite this activity relatively few cases are being transitioned to end users. Therefore, the environment in which evidence recovery is performed is not ideal. Hence the lack of skills forces organisations to tackle digital forensic investigations even though the education knowledge base has not yet been standardised.

This issue is further compounded by many organisations who underestimate the admissibility and reliability requirements of digital evidence required by the legal system (Kent and Ghavalas, 2005). Thus, the policies and procedures in place are not adequate enough to provide legally admissible evidence in a court of law. This regulatory challenge forms the basis for all digital forensic investigations and hence is the second most important challenge. This regulatory challenge can be demonstrated in the following case whereby an Internet Service Provider (ISP) performed a routine repair to Dale T. Weir's (the accused) e-mail box and found an e-mail message containing an attachment thought to be child pornography (Smith, 1998). The ISP reported the discovery to the police and when asked, provided the e-mail message and attachment to the police. Using this information the police obtained a search warrant for the accused's residence and seized the computer which contained the original e-mail message and attachment. The defence argued that the ISP breached the accused's privacy because the police obtained a warrant to access the information and therefore the search was unreasonable and the evidence should be excluded (Smith, 1998).

However the presiding judge found the accused guilty although he summarised that there was a reasonable expectation of privacy of e-mails but stated that "*because of the manner in which the technology is managed and repaired that degree of privacy is less than that of first class mail.*" Hence it can be seen that this case demonstrates the legal and procedural challenge and thus the nature of an investigation with respect to e-mail technology is volatile and the law must be adhered to at all times. Therefore, a methodology that provides a guide for the investigator must correlate with the law.

In order for evidence to be legally admissible, it needs to be compliant with the applicable legislation. The ECT Act is one such legislation that must be complied with. This law was developed to govern e-commerce in South Africa, and it applies to any form of electronic communication i.e. e-mail, Internet, SMS, etc. (Michalsons, 2005a). The crux of the Act with regard to e-mails is to permit electronic documents and e-mails as evidence in a court of law; however, there is a requirement to show authenticity and integrity of the information (Michalsons, 2005b). The governance of electronic evidence collection, storage and presentation is lacking (Watney, 2009). Hence, this makes it difficult to produce legally admissible evidence as there is no standardised method of recovering digital evidence. Additionally the Regulation of Interception of Communications and Provision of

Communication-Related Information Act 70 of 2002 (RICA) further governs e-mail access. RICA is South Africa's monitoring law and allows companies to monitor and intercept e-mail in certain specified circumstances (Giles, 2009). Companies may monitor and intercept e-mail where:

- Employees have consented to it; or
- It takes place "in the course of the carrying on of any business" at the company and provided that the other requirements of section 6 of RICA have been met.

Another critical challenge that digital forensic investigators face is that the forensic tools available at their disposal. These tools have a short life span and as a result they are not able to keep up with current investigations (Ayers, 2009). Tools such as Encase and Forensic Toolkit products have been around for over a decade but have limitations, such as processing speed and software errors (Ayers, 2009). There are two fundamental issues with the design of tools and according to Garfinkel (2010) these are:

- tools were designed to help examiners find specific pieces of evidence, not to assist in investigations
- tools were created for solving crimes committed against people where the evidence resides on a computer and not to assist in solving typical crimes committed with computers or against computers.

Additionally Garfinkel (2010) identifies issues with the current tools and states that data cannot be analysed because of imcompatibility issues, encryption as well as lack of training. Therefore, it is even more difficult to satisfy the authenticity and integrity of information when the tools were created for a different purpose.

A further procedural challenge is that of the scale of the forensic investigations. The trend has been increasing, from 80 GB in Fiscal Year (FY) 2003 to 250 GB in FY 2006 (FBI, 2006). The latest statistic showed that 1.756 TB of data was processed for FY 2008 which was a 27%

increase from the previous year (FBI, 2008). The impact of this trend is that it adds to the already complex matter of acquisition and extraction of data sources adding to a list of technical problems (Case, Cristina, Marziale, Richard and Roussev, 2008). In order for this challenge to be overcome, new tools need to be developed to cope with the demands of the latest investigations.

The current tools have the ability to discover all important system files; however, they are in need of attention because they cannot guarantee to recover unreferenced files (Arthur and Venter, 2004). Thus, new tools are needed in the field in order to maintain the level of reliability and admissibility required in a court of law. Each challenge has been addressed by contrasting views and by determining the best possible list of challenges that have been identified.

Moreover, the challenges already identified are changing and this requires an effective plan of action in order to address the challenges that digital forensic investigators must overcome. The critical challenges have been listed and different views have been contrasted in order to create a consolidated set of challenges. Additionally Garfinkel (2010) predicts that there is am “impending crisis” in the digital forensics domain and all the challenges support this prediction. However Garfinkel (2010) states that there needs to be co-operation, standardisation, and shared development in order for the digital forensics research community to survive the coming crisis and address these challenges.

2.4 Conclusion

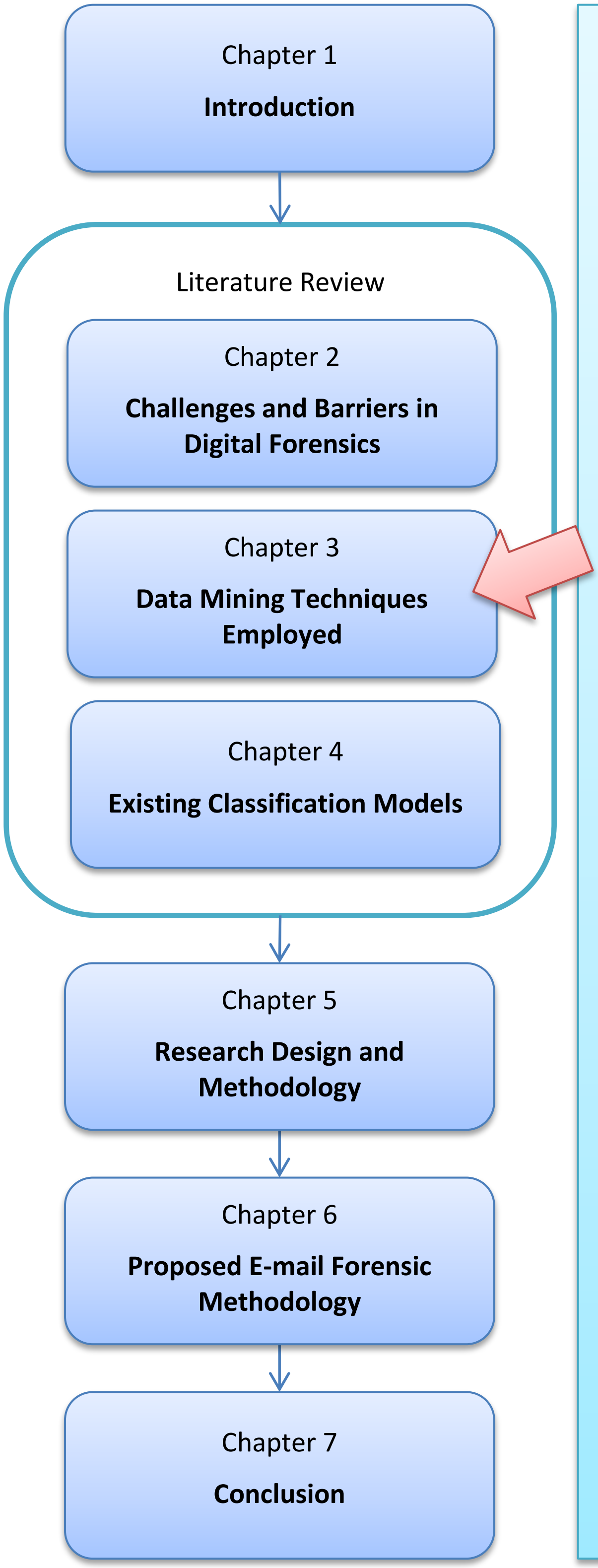
The key author’s views, which initially identified the challenges that the digital forensic profession face, were highlighted and discussed. The categorisation put forward by each was examined and the key issues within those challenges were discussed. A comparison was made between the key author’s views and similarities and dissimilarities were identified. From this comparison a consolidated categorisation was then presented which included all the challenges from different viewpoints, although, this categorisation aim was to broaden the scope of the challenges. Additional challenges were presented in order to highlight the increasing threat of new challenges as well as the expansion of existing challenges. The number of challenges is not a full list of and this demonstrates the urgency for them to be addressed. The categorisation cannot remain static and therefore it needs to be reviewed because of the changing nature of the digital forensic discipline.

The challenges discussed, can be summarised as Technological; Educational; Societal; Procedural and Legal, are not new, although the responsibility is now far greater than initially placed on the digital forensic investigator. Although the law forms the basis of every investigation, it is the responsibility of the investigator to maintain a carefully documented chain of custody that will allow the use of evidence in a court of law. The procedural challenge is a crucial one, as the tools used are the key to gathering the evidence used in a court of law. Therefore the law is the binding aspect within all the challenges and must be addressed first during an investigation; however, it was further established that education must be improved and the tools need to be developed in line with a standardised process.

In order to address these challenges, the very nature of digital forensics needs to change. Digital forensic investigators can no longer rely on traditional techniques and methods, but must adapt to the environment in order to be successful in their investigations. This will require the use of innovative methods and cutting edge tools and techniques while remaining within the ambit of the law. Furthermore, cooperation and standardisation in the digital forensic domain is needed in order to successfully address the challenges. The next Chapter will discuss some of the tools and techniques available to the digital forensic investigator.

3. DATA MINING TECHNIQUES EMPLOYED

CHAPTER 3


Chapter 1
Introduction
Literature Review
Chapter 2
Challenges and Barriers in Digital Forensics
Chapter 3
Data Mining Techniques Employed
Chapter 4
Existing Classification Models
Chapter 5
Research Design and Methodology
Chapter 6
Proposed E-mail Forensic Methodology
Chapter 7
Conclusion

3.1. Introduction

3.2. Authorship Attribution

3.3. Authorship Categorisation

3.4. Data Mining Tools that Aid in Author Identification

3.5. A Unique Method of Identifying an Author

3.6. Mining Write-Print

3.7. Conclusion

3.1 Introduction

Chapter two discussed the various challenges that the digital forensic discipline must overcome. Various viewpoints were examined and a consolidated categorisation of the challenges was established. Although the challenges took a high-level view it was not holistic; as there are specific challenges that must also be addressed within the investigation process. It is for this reason that the data mining techniques are examined as many methods exist to recover digital evidence. This Chapter addresses specific issues that endure within the procedural challenge.

Many industries and governments have become dependent on e-mail as a method of communication. The reason for the widespread use of e-mail is that it is an expedient and economical form of communication. However, the very nature of e-mails has led to abuse and misuse. Misuse includes unsolicited junk mail (more commonly known as spam), transmission of sensitive and private information, and mailing of threatening e-mails to name a few (De Val, 2000). The very nature of e-mails has allowed criminals to abuse and perform illegal activities via e-mail. In some instances, the sender of the e-mail will attempt to hide their true identity. Therefore, it is important to provide empirical evidence and identify the author of the original e-mail.

In order to identify the author of an e-mail, digital forensic investigators employ various techniques that aid in authorship identification. One such technique is that of Stylometry, and a method of examining a ‘textual fingerprint’ is emphasised. This textual fingerprint points towards a unique characteristic that can identify an author. These techniques are at times used in conjunction with tools that allow digital forensic investigators to identify the author of an e-mail. Various tools and techniques must be utilised during the investigation process in order to recover as much evidence as possible while remaining within the ambit of the law.

Therefore this Chapter is an in-depth examination of some of the data mining techniques employed by digital forensic investigators as well as authorship identification techniques. A brief description of authorship attribution is given and advantages and disadvantages are examined. Various views are also examined in order to determine the best approach that a digital forensic investigator should take. A number of tools and data mining techniques are examined and a comparison is performed in order to establish the best technique. Authorship identification is examined in the e-mail context and the tools used are identified. The authorship categorisation is determined by examining the subset of criteria that must be satisfied in order for authorship attribution to occur. A comparison of the various tools that are used is performed and their strengths and weaknesses are identified. Unique methods of authorship attribution are also examined and a conclusion is given.

3.2 Authorship attribution

3.2.1 Problem of Authorship Attribution

Authorship analysis has three underlying problems, namely authorship attribution, author categorisation and plagiarism detection shown in Figure 3 - 1 (De Val, Anderson, Corney and Mohay, 2001).

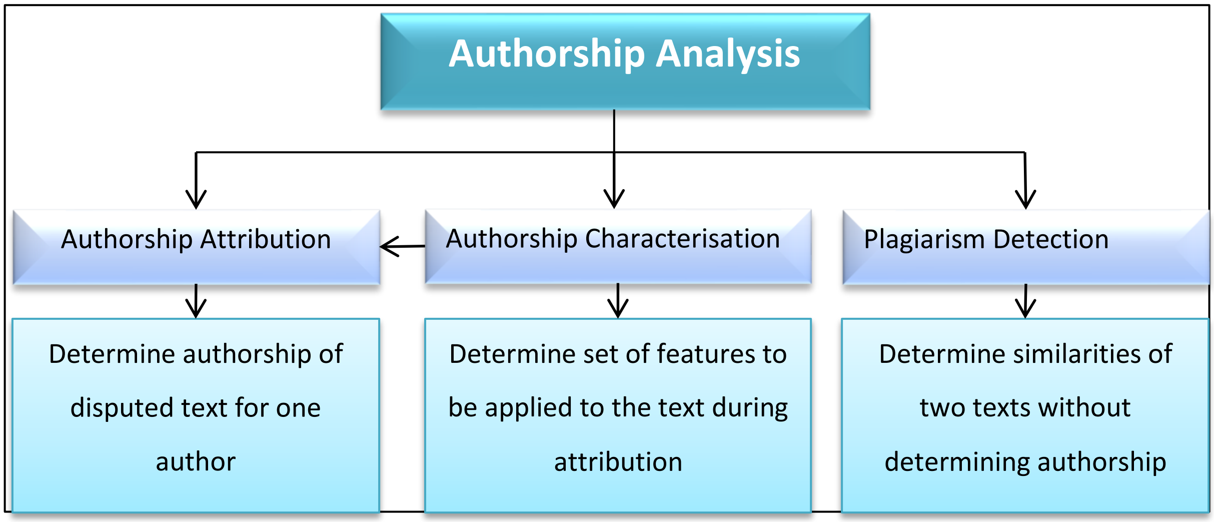

Authorship Analysis
Authorship Attribution
Authorship Characterisation
Plagiarism Detection
Determine authorship of disputed text for one author
Determine set of features to be applied to the text during attribution
Determine similarities of two texts without determining authorship

Figure 3 - 1 Authorship Analysis Sub-Divisions (De Val, Anderson, Corney and Mohay, 2001)

Authorship identification (also known as authorship attribution or stylometry) is the process of determining the author of a piece of work. The starting point of authorship attribution begins with the problem of identifying an author from a list of individuals for a given piece of work. According to Koppel, Schler, Argamon and Messeri (2006) this is the most straightforward version of the problem of authorship attribution although this problem is not new and authorship attribution has been around for decades. The following paragraphs will outline a brief history of authorship attribution.

Historically, authorship attribution studies were concerned with literary, historic and religious texts and many of the authorship analysis was around literature articles with disputed authors at the centre (McCombe, 2002). Smith (2008) suggests that stylometry possibly originated in 1851 when Augustus de Morgan proposed that it might be possible to identify biblical authors. However authorship attribution was marked by the nineteenth century study of Mosteller and Wallace (in 1964) on the authorship of the disputed Federalist Papers (Smith, 2008). The papers are a set of one hundred and forty-six eighteenth century essays, written by three different authors, namely John Jay, Alexander Hamilton and James Madison, and in most cases the author is known; however, the authorship of twelve papers is disputed (McCombe, 2002). Mosteller and Wallace used a Bayesian analysis of ninety function words and sixty other words to give odds for assigning each text to a given author and the conclusion was made that Madison was the most likely author of the twelve papers; however, more importantly this conclusion was an independent one (Smith, 2008; McCombe, 2002).

Mosteller and Wallace’s method continues to be used although Burrows altered this method after he described his method of stylometric analysis in a series of papers published in the late 1980s and early 1990s (Holmes 1998, as cited in Smith, 2008). The Burrows method essentially involves computing the frequency of each of a list of function words and performing principle component analysis to find the linear combination of variables that best accounts for the variations in the data (Smith, 2008). Rather than analyse this result statistically, the data is plotted and inspected visually for trends (Holmes 1998, as cited in Smith, 2008). This simple but effective method continues to be used today, partly because of the ease with which the results are communicated and interpreted (Smith, 2008).

In the early 1990s a technique called Advanced Machine Learning Technique was developed (Smith, 2008). These techniques, including genetic algorithms, support vector machines, and hidden Markov models were all applied to classic problems in authorship attribution, and to new data sets made available by the World Wide Web (Smith, 2008). Many of the studies confirmed the results from historical examination: for instance, three anonymous plays were compared to Shakespeare and Marlowe's compositions using neural nets, with results similar to those made by historians (Merriam and Matthews 1994, as cited in Smith, 2008). Moreover, Advanced Machine Learning Language achieves the same result as made by historians and this technique has many advantages over other techniques as it can be automated to further improve the efficiency of the authorship attribution.

In the authorship attribution field not all techniques have been successful in identifying an author. One such technique is Morton's 'Cusum' technique that caused much controversy in the early 1990s. The Cusum technique involves first calculating, for each sentence, the difference between a feature's value and the feature's average value for the whole document (McCombe 2002). These deviations from the mean are then summed cumulatively (hence 'Cusum'), and the resulting sequence of values is plotted. Morton determined experimentally that Cusum plots for different features were identical to the Cusum plots for sentence lengths in texts by a single author (Smith, 2008). Therefore an author's habits are replicated consistently in every paragraph he has written. Moreover, Morton claimed that authors could be reliably distinguished from each by applying the following method: the insertion of a handful of sentences written by one author into the paragraphs of another author produces an obvious skewing of the Cusum plot.

Morton's claims attracted the attention of lawyers, who wanted to utilise Morton's technique as a forensic technique to evaluate the authenticity of confessional statements (Holmes 1998, as cited in Smith, 2008). The research was used in several high-profile British court cases, and when challenged to demonstrate his technique on live television, Morton failed to "distinguish the writings of a convicted felon and the Chief Justice of England" (Grieves as cited in Juola and Sofko, 2006). Therefore the claims of certain techniques need to be analysed and tested thoroughly before being utilised. Widespread acceptance is not always an approval of a technique. The Cusum technique is also criticised due to its time consuming nature; it requires many manual processes to analyse and interpret data (McCombe, 2002). Hence, techniques such as Advanced Machine Learning Technique have been growing in use and are favoured as they can automate many processes.

Stylometry techniques are not only criticised, but stylometry itself is also questioned as in the case of USA (plaintiff) vs. Thomas James Zajac (defendant) in which Federal Bureau of Investigation special agent James R. Fitzgerald provided expert testimony that, based on research presented in a peer-reviewed journal article on authorship attribution, the defendant had written the threatening letters (Waddoups, 2010). However Zajac called William G. Eggington to testify about linguistics and he believed that Fitzgerald's conclusion was problematic based on the study performed by Carole Chaski (Waddoups, 2010). In the opinion on the judge,*"Neither Fitzgerald nor the Government has been able to identify a known rate of error, establish what amount of samples is necessary for an expert to be able to reach a conclusion as to probability of authorship, or pinpoint any meaningful peer review"* (Chaski 2001).

Juola and Sofko (2006) argue that the increase in research in recent years has led to the development of improved statistical techniques in conjunction with the wider availability of computer-accessible corpora and has made the automatic inference of authorship at least a theoretical possibility. However Juola and Sofko (2006) also note that the inaccuracy of techniques such as the Cusum technique can render any judgement moot because taking into account that people are not always consistent and will occasionally incorrectly spell words or use different spellings, therefore any test applied must be able to handle such distributions (Juola and Sofko, 2006). Hence, thorough testing of the technique to be used is needed and techniques such as the Cusum technique amplify this need in order to review and criticise the authors' claims.

Despite some problems utilising stylometry to identify an author, Baayen, van Halteren, Neijt and Tweedie (2002) performed an experiment to determine if two authors can be distinguised based on their writing styles. The results of the experiment showed that two authors can have considerable authorical structure and this supports the theory that authors may have textual fingerprints (Baayen, van Halteren, Neijt and Tweedie, 2002). Hence research in the stylometric field has pointed towards a 'textual fingerprint' and continuous experimentation must be done in order to examine the validity and existence of the theory.

If the hypothesis only supports a textual fingerprint for authors who do not conciously change their writing style, then how does it respond to authors who deliberately change their writing style. In the studies mentioned above, there are a number of different features utilised in the attribution of authors and according to Rudman (1998), approximately 1000 style markers have been identified. However, there is no concensus by the researchers on which feature set to use and this complicates the identification of an author based on style marker. Therefore, extensive experimentation using style markers for author identification is needed.

Thus the problem of authorship attribution lies with the technique that is used as well as the approach taken by the examiner. In short, the argument for authorship attribution outweighs the negative criticisms of the techniques that have failed and therefore this study will examine tools and techniques that employ the stylometric approach of authorship attribution as previous research has indicated that there is value in employing this approach. The following section addresses the identification of an author in the context of an e-mail.

3.2.2 Authorship Attribution in the E-mail Context

This section highlights the problem of authorship identification in the e-mail context. Prior to discussing the authorship attribution one has to understand how e-mail works and therefore a brief overview will be given. E-mail relies on two basic communications protocol: Simple Mail Transfer Protocol (SMTP), which is used to send messages and Post Office Protocol (POP3), which is used to receive messages and the simplified version of the e-mail life cycle is depicted in Figure 3 - 2 (Katakis, Tsoumakas and Vlahavas, 2007).

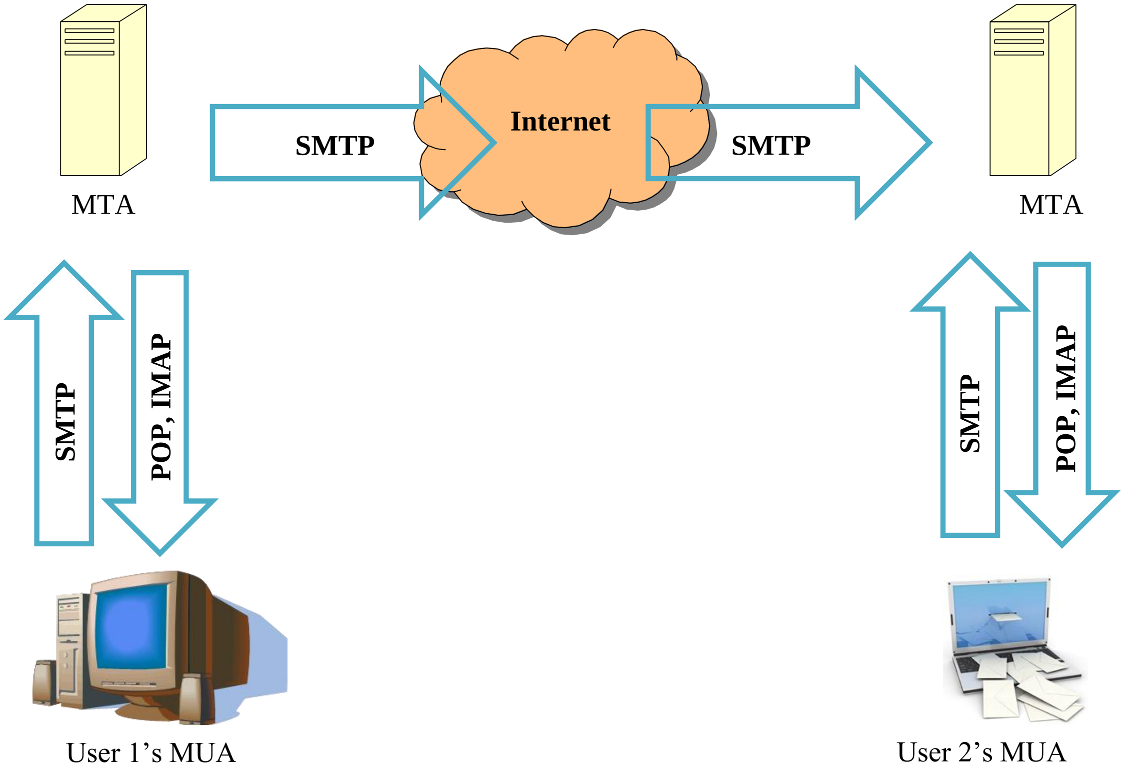


Figure 3 - 2 E-mail life cycle (Katakis, Tsoumakas and Vlahavas, 2007)

Katakis, Tsoumakas and Vlahavas (2007) give four important mail aspects:

1. Mail User Agent (MUA) - Responsible reading and writing e-mail messages. The MUA is usually implemented in software usually referred to as ‘e-mail client’. Two popular e-mail clients are Microsoft Outlook and Mozilla Thunderbird. These programs transform text messages into the appropriate Internet format in order for it to reache its destination.
2. Mail Transfer Agent (MTA) - Accepts a message passed to it by either an MUA or MTA and then decides on the appropriate delivery method and the route that the mail should follow. It uses the SMTP protocol to send the message to another MTA or an MDA.
3. Mail Delivery Agent (MDA) - Receives messages from MTAs and delivers them to the user’s mailbox in the user’s mail server.
4. Mail Retrieval Agents (MRA) - Fetches e-mail messages from the user’s mail server to the user’s local inbox. MRAs are often embedded in e-mail clients.

The e-mail message itself is composed of two sections, namely a header and a body. The header contains information about the destination of the message and the body is the text of all information contained within the e-mail (Katakis, Tsoumakas and Vlahavas, 2007). Now that the basic understanding of how an e-mail works has been discussed, the problem of authorship attribution can be analysed within the e-mail context and this discussion follows.

In light of the problem of authorship identification with respect to stylometry, authorship attribution in the e-mail context has further implications. E-mail is a relatively new medium of communication and in many institutions and organisations it has become the standard method of communication and in some cases the preferred method (Berendt and Draheim, 2007). There are a number of reasons for this shift that resulted in the popularisation of e-mail communication. One reason is that e-mail is relatively easy to use and messages can be created and sent in a matter of minutes as opposed to the traditional ‘written letter’ which takes time to compose and send (Gupta, Mazumdar and Rao, 2004). Organisations use e-mail as an audit trail for communication between departments e.g. communication between the programming department and the business consulting department where change requests on behalf of the client are recorded in a formal e-mail. Another example is that of the minutes of a meeting being distributed via e-mail after they have been transcribed. Therefore the e-mail message is a useful tool that has become embedded in business practice and has contributed towards a more efficient work place.

Furthermore given the ease with which an e-mail message can be created, a number of problems present themselves. Consider the following scenario where an employee authors a derogatory e-mail against his employer. The suspected employee can deny authorship of the said e-mail on the basis that his workspace computer was open and any of his colleagues could have sent the e-mail from his computer (Chaski, 2005). E-mail is also targeted by cyber criminals where their methods range from simple anonymity to identity theft and impersonation (Hadjidj et al., 2009).

Hadjidj et al. (2009) state that there are two limitations to why e-mail communication is exposed to such criminal users; the first is that there is no mechanism for encryption on the sender side and no integrity check on the recipient side and; secondly, the commonly used e-mail protocol i.e. Simple Mail Transfer Protocol lacks a source authentication mechanism. Broucek and Turner (2002) concur that e-mail is inherently insecure as a communication medium and that the majority of employees are unaware that the content of an e-mail unless encrypted can be accessed through is transfer from sender to receiver. Zhang, Liu, Zhang and Wang (2006) note that the simplicity of the SMPT allows it to be easily faked and hence spammers employ anonymous servers to send junk mail (also known as spam).

Moreover, Broucek and Turner (2002) point out that many e-mail systems still utilise the efficient but simple POP3 (post office protocol) which sends passwords in clear text unencrypted across computer networks thereby encouraging spoofing security breaches. Hadjidj et al. (2009) agree that the path along which an e-mail travels can easily be forged or anonymised. Therefore given these security issues e-mail messages can be compromised at any point in their transmission from sender to receiver. These issues further complicate the identification of an author. Therefore it is not only necessary that there is a method of identifying an author but also to ensure that the e-mail message is valid that there are efficient tools to verify the message. The first step in author identification is the creation of some kind of method to establish authorship, hence authorship categorisation is performed. This will be discussed in the following section.

3.3 Authorship Categorisation

In order for authorship attribution to occur, a subset of criteria needs to be established. These criteria are realised through the authorship categorisation process (De Val, Anderson, Corney and Mohay, 2001). This is a process of allocating a set of rules to identify the author and De Val et al. (2001) state that author categorisation utilises other fields such as author characterisation and similarity detection. Author characterisation determines the author profile and this includes characteristics such as educational and cultural history whereas similarity detection calculates the degree of similarity between two pieces of work irrespective of author identification (De Val et al., 2001). De Val et al. (2001) refer to authorial features as a set of distinctive attributes that can uniquely identify an author and this is designated as stylistic evidence.

De Val et al. (2001) argue that studies which base their author attribution on character or words such as vocabulary richness can be consciously controlled and hence opted to employ syntactic patterns thought to be unconsciously generated. An example given by De Val et al. (2001) is the 'function' words "if" and "the" whose frequency is unaffected by the subject matter. However Chaski (2005) examined three methods of author attribution, namely biometric analysis; qualitative language analysis; and lastly computational stylometric analysis. Biometric analysis employs the use of a profile created from the actual keystrokes of a user. Gupta, Mazumdar and Rao (2004) proposed a biometric based authentication method that employed the use of capturing the keystrokes of a user and generating a biometric profile based on the pattern established from the capture process. However this method of author categorisation is non-linguistic, although Chaski (2005) suggests that biometric profiles may be affected by the linguistic characteristics of the phonotactics of language.

The qualitative language analysis approach assesses the errors and idiosyncrasies of the user based on the digital forensic investigator's experience (Chaski, 2005). Koppel and Schler (2003) employed the following set of features in identifying an author: lexical features, part of speech tags and idiosyncratic usage; the latter considered syntactic, formatting and spelling usage, and various subsets of these were categorised. The third method as described by Chaski (2005) is quantitative and computation stylometric analysis in which countable language features such as word length and vocabulary frequency are used to create a basis for authorship categorisation. De Val et al. (2001) criticises linguistic techniques citing that they do not quantify linguistic patterns and fail to discriminate between authors with a high degree of precision, although this method of authorship attribution testimony as admissible evidence in legal proceedings has been identified in a number of cases. It is for this reason that many of the automatic computation methods of authorship attribution utilise stylometry and this ensures that a baseline for all methods of computational stylometry is set. The following section examines some of these tools and an in-depth analysis is performed for a number of the tools employing stylometric analysis.

3.4 Data Mining Tools that Aid in Author Identification

During digital forensic investigations, various tools and techniques are used by the digital forensic investigator in order to preserve the crime scene as well as to gather as much information and data about the event that occurred. Moreover, the anonymous nature of e-mail has made authorship identification a problem and while there are no proactive mechanisms to protect e-mail, there are techniques that can be used to determine authorship of e-mails. These techniques are used in conjunction with various tools in order to accomplish the task of authorship identification. Table 3 - 1 presents some tools and their strengths and weaknesses. Little has been done to show how data mining techniques can be used to identify authors from their e-mail style where authorship is disputed (Corney et al., 2001). The following section will discuss some of these techniques as well as identify possible strengths and weaknesses and point towards a technique that can support an investigation without detracting from the credibility of the digital forensic investigator.

Corney et al. (2001) state that the writing style of an e-mail message is situated somewhere between the informal spoken word and the written formal letter. More significantly, the e-mail message is a form of record of an act, intention, instruction or even an attitude (Corney et al., 2001). This therefore illustrates the purpose and importance of an e-mail message. However, Corney et al. (2001) suggest that e-mail messages can be spoofed or anonymised. Therefore, the problem arises when using an e-mail as a formal record and hence if an author denies ownership of a message, it would be crucial to establish the author with the use of some method. The process of analysing e-mail is called e-mail mining, which is a process of discovering useful patterns from e-mails (Nagwani and Bhansali, 2010). According to Katakis, Tsoumakas and Vlahavas (2007) e-mail mining can be considered as an application of the upcoming research area called text mining on e-mail data and hence it has attracted researchers from areas such as Machine Learning, Data Mining, Natural Language Processing and Computational Linguistics.

Table 3 - 1 Strengths and Weaknesses of E-mail Data Mining Tools (Own Compilation)

Tool Name	Strengths	Weaknesses
Support Vector Machine and Stylometry Used to determine authorship of e-mails (Corney et al., 2001).	• Based on Structural Minimisation principle • Provides a systematic way of determining the relative effectiveness of raw style markers	• Does not yield admissible evidence • More experimentation to determine sensitivity of authors to style markers
Writing-Style Features and Classification Techniques Used to address online messages and said author (Zheng et al., 2006).	• Experimental approach able to identify author • Structutral and content specific features allow identification of authors • Uses three classification techniques • Applied to multiple languages: English and Chinese	• Identification of optimal set of features for online messages • More experimentation needed • Validation of proposed technique in the field
Integrated E-mail forensic analysis framework Java based application used to determine authorship of e-mails (Hadjidj et al., 2009).	• Theoretical foundation based on statistical analysis, text mining and stylometry together with social networking techniques • E-mail geographic localisation – used to localise information relating to suspect e.g. e-mail server	• Level of cohesion of techniques needs to be increased in order to obtain more credible results • Further investigation is prompted for e-mail social networks
AutoMiner Noval data mining technique using frequent patterns and comparing it to write-print of an individual (Iqbal et al., 2008).	• Unique identifier for authorship identification – namely write-print which is dynamically extracted • Accuracy of 86 – 90% • Robust method for determining authorship	• As minimum supported threshold of intervals (features) increase, the accuracy decreases • Manual examination of write-prints as many frequent patterns are not obvious
EnCase Enterprise Edition 4.19a Designed to integrate with enterprise security architecture, providing enhanced access control and audit functions, and enabling digital forensic investigators to process many systems on a network simultaneously (Casey and Stanley, 2004).	• Tool of choice for enterprise investigations • Extracts more data than PDIR • Does not alter data on remote system • Uses System called SAFE to manage security • Data acquisition of 3.5 MB/s • Gives information about which files are opened • Can integrate with intrusion detction systems	• Data acquisition slow due to SAFE system initially reading device • Require administrator privileges • Cannot view data on network shares limiting amount of data • Provide most information possible
ProDiscover IR 3.5 Designed to examine one system at a time and is useful for focused investigations involving a small number of computers (Casey and Stanley, 2004).	• Alters last accessed date/time stamps when performing some processes • Has optional encryption and password protection • Only presents information that is verifiably complete	• Has optional encryption and password protection not enabled by default for servlet • Data acquisition of 5.5MB/s • Require administrator privileges • Cannot view data on network shares limiting amount of data

Text mining applies the same analytical functions of data mining to textual information (Dörre, Gerstl and Seiffert, 1999). A distinction can be made between e-mail mining and text mining and the key data characteristics are highlighted in Table 3 - 2. Therefore text mining is the data mining of e-mail textual data which can be specifically named as e-mail mining based on key characteristics.

Table 3 - 2 Key Data Characteristics of E-mail Mining (Katakis, Tsoumakas and Vlahavas, 2007)

No.	Characteristic
1.	E-mail includes additional information in the headers of e-mail that can be exploited for various e-mail mining tasks.
2.	Text in e-mail is significantly shorter and, therefore, some Text Mining techniques might be inefficient in e-mail data.
3.	E-mail is often cursorily written and, thus, linguistic well-formalness is not guaranteed. Spelling and grammar mistakes also appear frequently.
4.	In an e-mail message, different topics may be discussed; a fact that makes mail classification more difficult.
5.	E-mail is personal and therefore generic techniques are difficult to apply to individuals.
6.	E-mail is a data stream and concepts or distributions of target classes may change over time. Algorithms should be incremental in both ways: instance-wise and feature-wise, as new features (e.g. words) may appear.
7.	E-mail will probably have noise. HTML tags and attachments must be removed in order to apply a text mining technique. In some other cases, noise is intensively inserted. In spam filtering for example, noisy words and phrases are inserted in order to mislead machine learning algorithms.
8.	It is rather difficult to have public e-mail data for experiments, due to privacy issues. This is a drawback especially for research since comparative studies cannot be conducted without public available datasets.

Katakis, Tsoumakas and Vlahavas (2007) state that most e-mail mining tasks are accomplished by automatic e-mail classification typically using machine learning techniques. Corney et al. (2001) have used stylometry in conjunction with a learning machine technique called Support Vector Machine (SVM) to determine authorship of a suspect e-mail. Corney et al. (2001) add that SVM based identification gives useful results on text samples as small as 100 to 200 words although there are clear problems when applying stylometry to e-mails given the size of the e-mail text. However despite this Corney et al. (2001) reason that they have employed the use of stylometric features that have previously been recognised as successful when applied to ordinary text.

SVM is based on the structural risk minimisation principle and the idea behind it is to find a hypothesis that guarantees the lowest true error for the classification problem (De Val et al., 2001). Corney et al. (2001) used an experimental approach and proceeded by selecting one hundred and eighty-four style markers from a list of features previously used by other authors in the stylometric field. Since the e-mail messages do not contain a constant number of words, the features were normalised where possible i.e. a ratio of frequency of some property (e.g. number of upper case letters) to a summary property (total number of letters) (Corney et al., 2001).

These tests were first carried out on non-email data and then carried out on the e-mail corpus. The SVM algorithm provides a systematic method of determining raw style markers and can be used to create an authorship identification tool based on the result of the experiment carried out (Corney et al., 2001). Although, this approach will not yield admissible evidence in court as yet, the SVM approach has set about creating the foreground for court admissibility i.e. The Daubert criteria (i.e. empirical testing, known error rates, standard procedures etc.)(Corney et al., 2001). Other authors such as Koppel and Schler (2003) employed experimentation of authorship attribution using SVM. They applied their tests to a corpus of e-mail because sufficient material with a sufficient number of authors i.e. 480 e-mails with an average length of 200 words, written by 11 different authors over a period of a year (Koppel and Schler, 2003).

Other reasons for choosing this corpus was due to the number of errors that could be found in the e-mails as they were not written in an overly polished style; the material was homogeneous as there was a single topic of automatic information extraction; and lastly the material was in the public domain. These criteria are important as they establish the context for the experiment and hence, provide a baseline for comparison with other studies.

Koppel and Schler (2003) used two types of classification algorithms i.e. linear SVM and decision trees and three classes of features namely lexical, parts of speech tags and idiosyncratic usage on the e-mail corpus. Koppel and Schler (2003) followed a method of detecting errors where the texts were run through the MS-Word application to check for errors; however the authors found the MS-Word spell checker to be inadequate and hence prepared their own scripts. Figure 3 - 3 and Figure 3 - 4 illustrate the results in terms of efficiency.

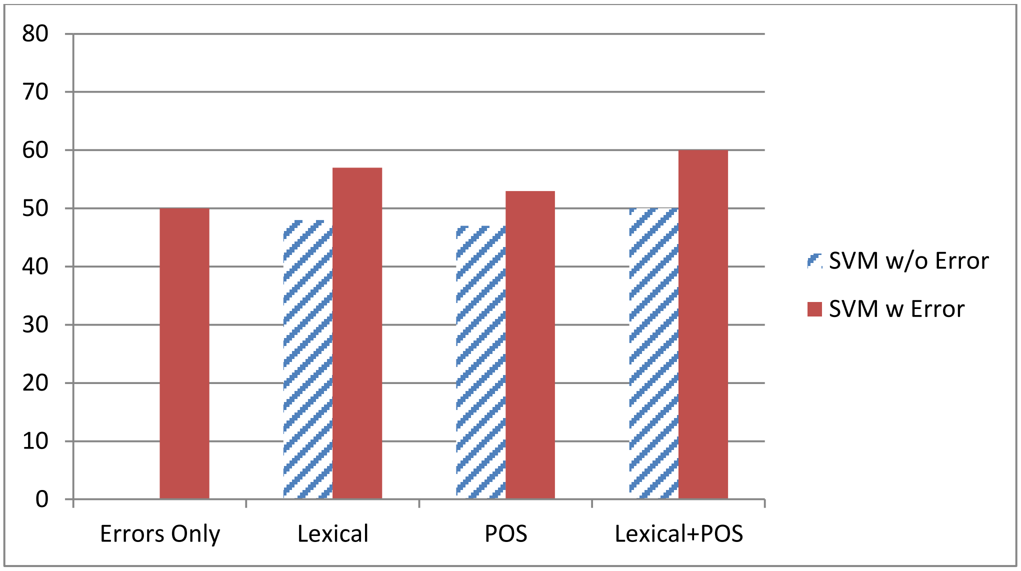


Figure 3 - 3 SVM results (Koppel and Schler, 2003)

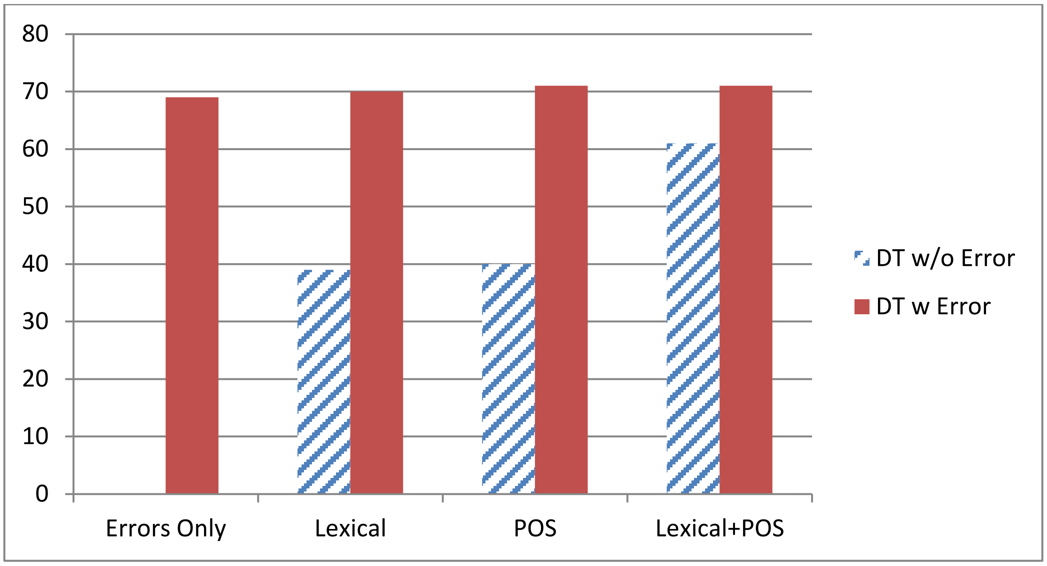


Figure 3 - 4 C4.5 Results (Koppel and Schler, 2003)

As can be seen from the graphs, SVM is more efficient than C4.5 for lexical and part of speech (POS) features alone; however once the errors are included C4.5 becomes more effective although the change for C4.5 from having no errors to including errors is dramatic. The purpose of the experiment was to determine the effectiveness of certain features and this was demonstrated with the use of e-mail texts and the algorithms. In essence the algorithm does not hold as much value as the feature sets used in the experiment because the feature sets to some extent, determine the outcome; however, it can be seen that SVM is the more consistent algorithm whereas C4.5 will give a better accuracy when specific features such as errors are included.

Koppel and Schlers' (2003) main objective with the experiment was to determine a so called 'smoking gun' for authorship attribution. The results show that the error rate features are responsible for accuracy when the C4.5 algorithm is employed when compared with other features which hardly contribute to the accuracy. Koppel and Schlers (2003) give an example of how error types can contribute to author attribution; an author uses British spelling where organisation is spelt with an 's' not a 'z', hence the error type *confusing 's' and 'z'* was helpful in identifying the author.

Therefore the feature types that human expert may exploit for authorship attribution can be identified and exploited in an automated version. Koppel and Schlers (2003) note that the feature set identified and compiled in their experiment can be greatly enhanced; however, they also suggest that greater improvement will be encountered with a larger corpus. Therefore traditional techniques can be used in automated methods such as SVM to determine authorship; nevertheless more testing of the features set to improve the accuracy is needed.

3.5 A Unique Method of Identifying an Author

Notwithstanding this, a framework was developed using stylometry, to address online messages specifically in order to identify the author of such messages (Zheng, Li, Chen and Huang, 2006). The process can be divided into four steps: message collection; feature extraction; model generation and finally author identification as can be seen in the Figure 3 - 5. A crucial element that forms part of the framework is that the authors have investigated the use of the framework and the algorithm in an attempt to establish the performance in a multilingual context. For this reason the framework was evaluated using two languages namely English and Chinese. Chinese was chosen as there are a number of characteristics that differentiate it from English i.e. word boundaries. Another element of the framework is that online messages are more casual than formal publications and that this is why authors are more likely to leave behind their own ‘write-print’.

The write-print of an individual is similar to the ‘textual fingerprint’ that Baayen et al. (2002) highlighted in their examination of stylometry. Zheng et al. (2006) employed WEKA (Waikato Environment for Knowledge Analysis) data mining tools. The WEKA tool was chosen as it implemented Platt’s sequential minimal optimisation algorithm which is the fastest algorithm for SVM design (Keerthi, Shevade, Bhattacharyya and Murthy, 2001). The WEKA tool is written in Java to ensure that it is available across all computer platforms and to enable automatic documentation (Witten, Frank, Trigg, Hall, Holmes and Cunningham, 1999).

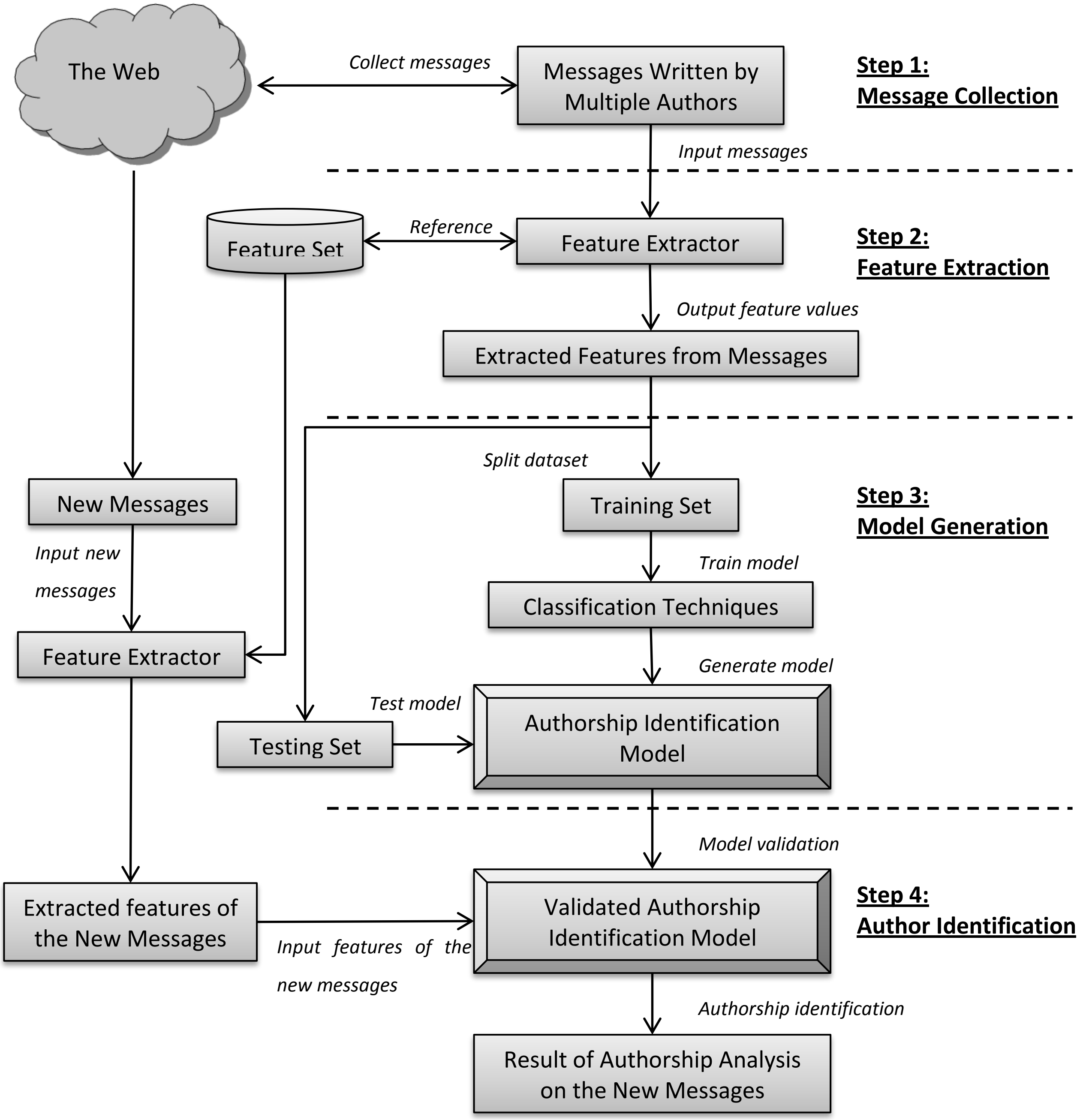

The Web
Collect messages
Messages Written by Multiple Authors
Step 1: Message Collection
Input messages
Feature Set
Reference
Feature Extractor
Step 2: Feature Extraction
Output feature values
Extracted Features from Messages
Split dataset
New Messages
Training Set
Step 3: Model Generation
Input new messages
Train model
Classification Techniques
Feature Extractor
Generate model
Test model
Testing Set
Authorship Identification Model
Model validation
Extracted features of the New Messages
Validated Authorship Identification Model
Step 4: Author Identification
Input features of the new messages
Authorship identification
Result of Authorship Analysis on the New Messages

Figure 3 - 5 Framework for authorship identification (Zheng et al., 2006)

During Step 1 of the framework of Zheng et al. (2006), the authors collect a set of online messages from potential authors in order to create a profile of their writing styles. This was done as in previous studies. Online messages are in an unstructured format and Step 2 of the framework extracts features of the authors based on predefined writing styles. Step 3 is the process of creating the classifier, and this is achieved by dividing the dataset into subsets called the training set and the testing set. The training set is used to train the classifier whereas the testing set is used to determine the classifiers prediction power, and if the testing set is validated it can be used to identify the authorship of the online messages. Zheng et al. (2006) suggest that an iterative training and testing process may be required to develop a good authorship prediction model.

After the authorship identification model is developed from the training set, it is used to predict the authorship of unknown online messages. The result of authorship identification will help the digital forensic investigator focus his or her effort on a small set of messages and authors. Zheng et al. (2006) evaluated their framework using an experimental approach in order to determine the effect of feature types and classification techniques on authorship identification of online messages. The results of the study showed that as more feature types were added, the accuracy of identification increased. It was also observed that the classifiers SVM outperformed Neural Networks (NN) which in turn outperformed C4.5 in authorship identification.

Zheng et al. (2006) noted that the results for the Chinese dataset were consistent with these results and that the best accuracy was achieved with SVM and all feature types. However, there was a decrease in accuracy from the English dataset suggesting that the result was influenced because different writing styles were used and the automated feature extraction was not as accurate as it was for English. Moreover, the accuracy increased as the number of authors decreased or the number of online messages increased.

The reason for the small number of messages per author is because online users utilise several channels and different user names; however the basis of the experimentation was to establish the prediction power of authorship identification for online messages using different parameter settings. The results for the English dataset are shown in Figure 3 - 6. The graph plots all three classifiers together as well as the number of authors. Thus it was demonstrated that the framework has the potential to trace unknown authors in cyberspace.

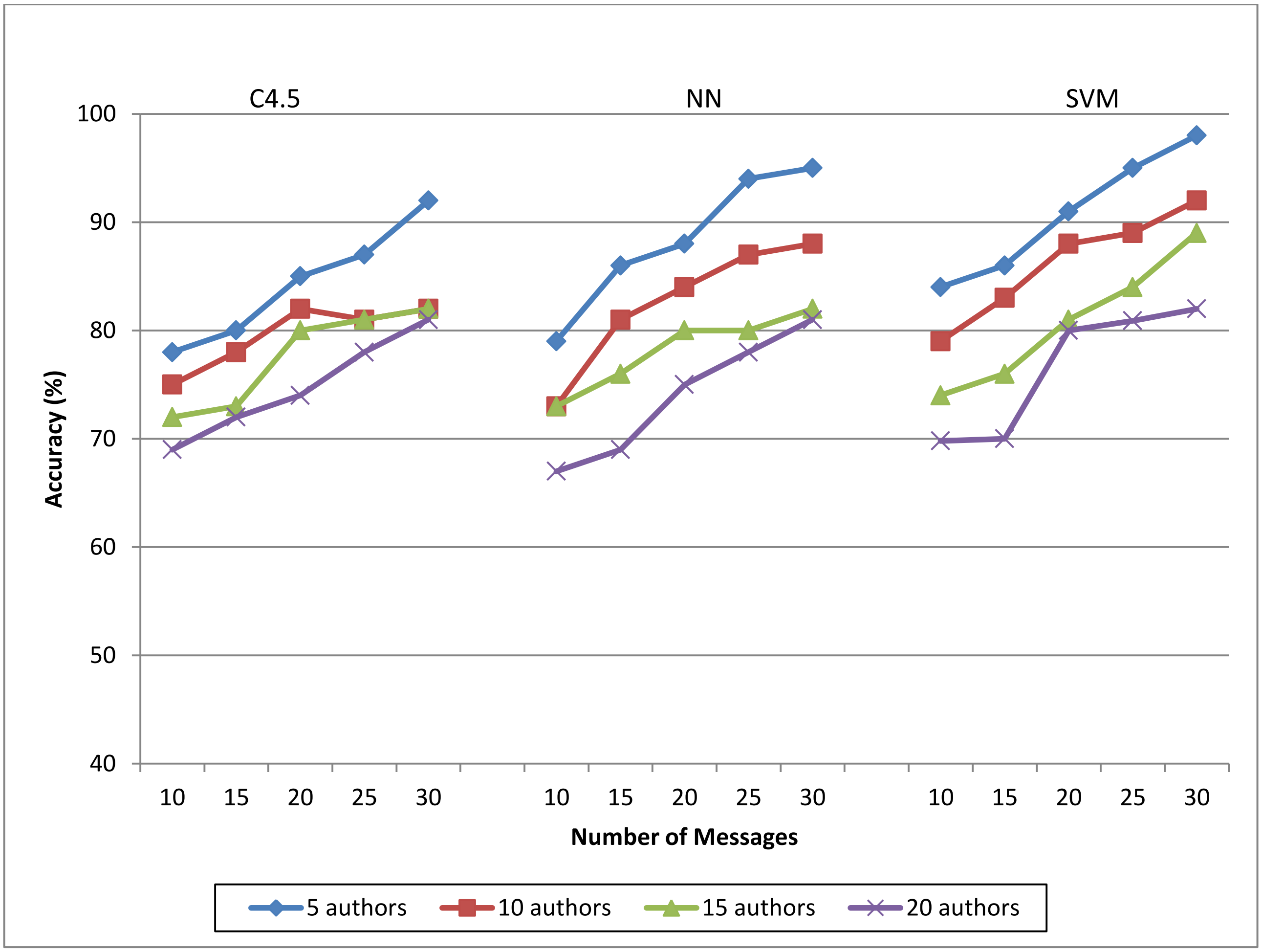


Figure 3 - 6 Comparison of Classifiers for the English Dataset Identification (Zheng et al., 2006)

Based on the studies and experiments of Koppel and Schler (2003) and Zheng et al. (2006), the SVM classify is the most accurate classifier when compared to C4.5 and Neural Networks. Additionally, the feature set used impacted on the accuracy of the classifiers prediction power, hence when SVM is utilised, the feature set needs to be as comprehensive as possible. Iqbal, Hadjidj, Fung and Debbabi (2008) state that C4.5 only utilises one attribute when employed for authorship attribution and it therefore fails to consider all features of an author; however this is overcome by SVM which considers all features although it does not explain how it arrived at the conclusion of one specific author.

3.6 Mining Write-Print

A novel method of data mining called AutoMiner has been proposed by Iqbal et al. (2008) and is an algorithm which can determine authorship of an e-mail by extracting frequent patterns from the e-mail and comparing it to the 'write-print' of a suspected individual. Frequent patterns have been used to identify consumer behaviour patterns and even hidden patterns in DNA sequences but this is the first application for authorship attribution (Iqbal et al., 2008). Iqbal et al. (2008) define an individual's write-print as combination of features that frequently occur in a written e-mail. When generating the write-print of an individual Iqbal et al. (2008) seek unique features that can identify the author and do not utilise all features as some could possibly be shared by another author.

Iqbal et al. (2008) posit that this approach has the following merits not contained in existing work:

- Justifiable evidence: the write-print is a unique set of features that guarantees the identification of only one author.
- Flexible writing styles: the frequent pattern mining technique can adopt all four types of commonly used writing style features.
- Feature optimisation: it is a combination of features and the support to each write-print in the feature set determines the contribution.
- Generic application: the technique is used in the experimentation on real world data i.e. the Enron e-mail corpus.

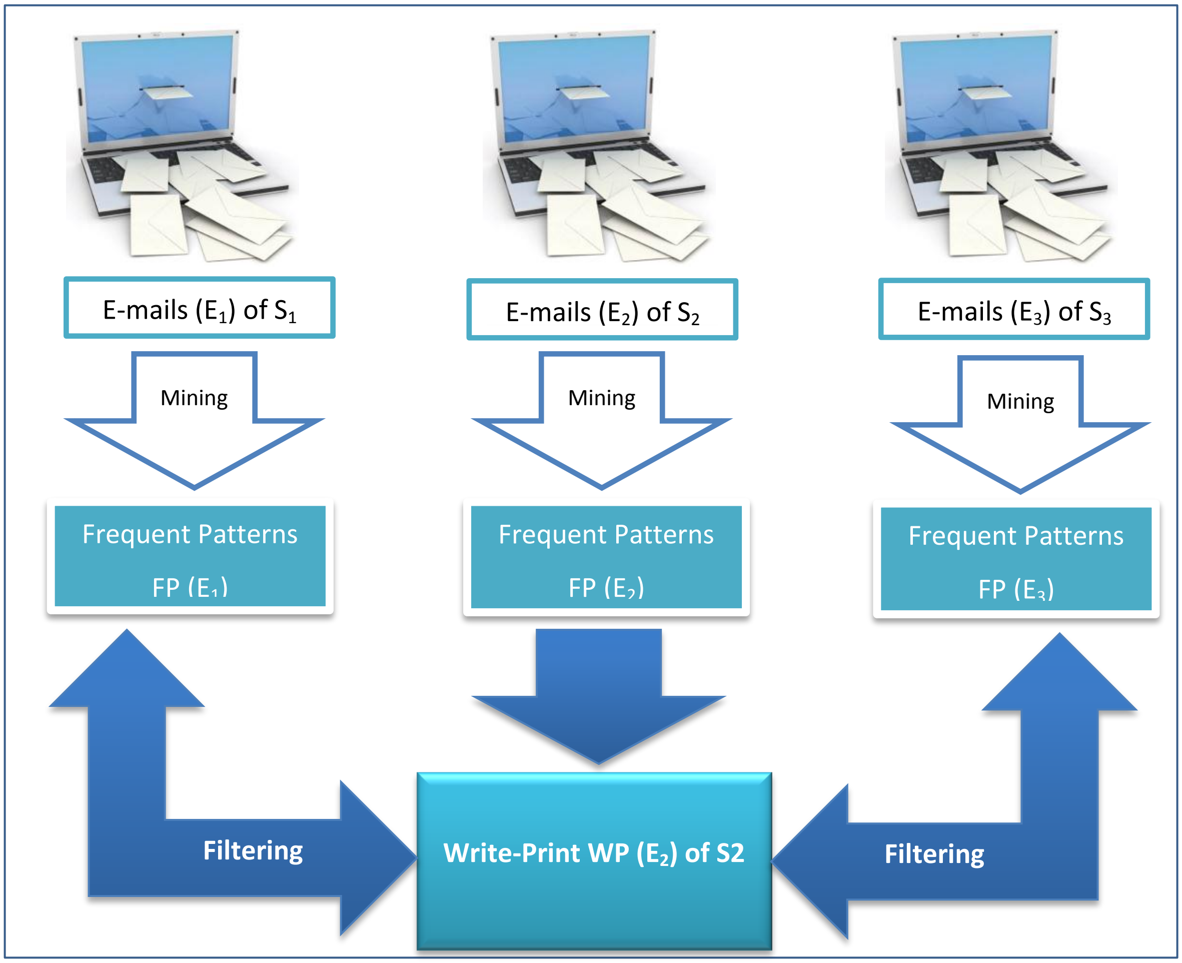


Figure 3 - 7 Mining Write-Print (Iqbal et al., 2008)

Iqbal et al. (2008) proposed the AutoMiner method in Figure 3 - 7. The method differs from the traditional method of write-print in that it is composed of a combination of features dynamically generated based on the pattern embedded in the e-mail and therefore allows the model to produce write-prints for different suspects (Iqbal et al., 2008). Iqbal et al. (2008) declare that the frequent pattern of the suspect only appears in that e-mail authored by the suspect and not in other suspect's e-mails and because of this it is argued that the evidence gathered in an investigation is more reliable and convincing.

The evaluation of this method was done through the Enron Dataset, which consisted of 200 399 e-mails from 158 employees after cleansing. Employees were randomly selected and a selected number of e-mails were used for the evaluation, of which 2/3 were used for training and 1/3 for testing. AutoMiner was then employed to determine the write-print of the e-mails from the training set and then used to determine the author of each e-mail. Six employees and a total of 120 e-mails were used to determine the accuracy of the method which was found to be between 86% and 90%. This successfully demonstrated that the write-print of an individual can lead to the identification of a single author.

Hadjidj et al. (2009) employed the AutoMiner method in the development of a framework based on a combination of established techniques: statistical analysis, text mining and stylometry together with e-mail social networking. The framework is called Integrated E-mail Forensic Analysis Framework (IEFAF). IEFAF is composed of five sub-modules

- Inter-database browser
- Statistics explorer
- Data mining explorer
- Weka submodule
- E-mail explorer

The authorship attribution occurs in two steps; the first e-mail grouping is conducted using content and stylometry based clustering of the data. The second step is the classification phase which entails feature extraction from the body of the e-mail followed by model generation and application (Hadjidj, et al. 2009). However this framework consists of a number of techniques to determine authorship and therefore in order to obtain more credible results, the cohesion of these techniques must be increased.

A C Sharp (C#) application to determine authorship of an e-mail with the use of stylometric features was proposed by Goodman, Hahn, Marella, Ojar and Wescott (2007). The program has 3 phases: data collection, feature extraction and classification. Tests performed showed that the application correctly identified 80% of e-mails (Goodman, et al., 2007). Therefore, a number of techniques can be used effectively in combination with stylometry.

The problem that digital forensic investigators are facing is that with all the information that a computer can store for one e-mail message, they are looking for only a minute trace of evidence to indicate authorship (Goodman, et al., 2007). Thus without techniques such as Autominer, the investigators will have a difficult job finding evidence.

Many digital forensic investigators have compiled their own toolbox of executables in order to be prepared for all eventualities (Casey and Stanley, 2004); however, the digital forensic investigator needs to be wary of which tools can be applied. Digital forensic investigators need to consider the tools they use as some of these tools can alter the state of the system from which evidence is being recovered; often they have to obtain evidence from remote live systems (Szezynska, Huebner, Bem and Ruan, 2009) and this increases the complexity of forensic retrieval of evidence.

It is for this reason that Casey and Stanley (2004) compare two tools namely ProDiscover IR (PDIR) 3.5 and EnCase Enterprise Edition (EEE) 4.19a. These tools are used for incident response and to preserve evidence on live remote systems. Generally both tools do not alter data on the remote system; however PDIR changes last accessed date/time stamps (Casey and Stanley, 2004). There are a range of tools that have been developed for specific tasks (Case, et al., 2008). The Forensic Toolkit (FTK) and Encase are two such examples and are commercial digital forensic suites used for the analysis of captured disk images (Arthur and Venter, 2004). There are also offline memory and log analysis tools used for memory acquisition, such as BodySnatcher (Case, et al., 2008). It is therefore imperative that the appropriate use of these tools is determined by the digital forensic investigator.

Digital forensic tools are not being developed fast enough to keep pace with the variety of forensic targets (Casey and Stanley, 2004). Ayers (2009) shares this opinion; however Ayers (2009) propose a set of requirements for the development for new tools while Arthur and Venter (2004) suggest some improvements to tools such as FTK and EnCase and believe that the prosecution of cybercrimes will increase if suggestions are researched.

3.7 Conclusion

Authorship attribution is an on-going problem that has existed for a number of decades. The original problem was the identification of authors from disputed texts. However, with changes in technology this problem has taken on a new facet, that is the identification of the author of in the context of an e-mail. Researchers have performed experiments with the analysis of e-mail data and the most common technique used is that of stylometric analysis, albeit in combination with a number of different applications.

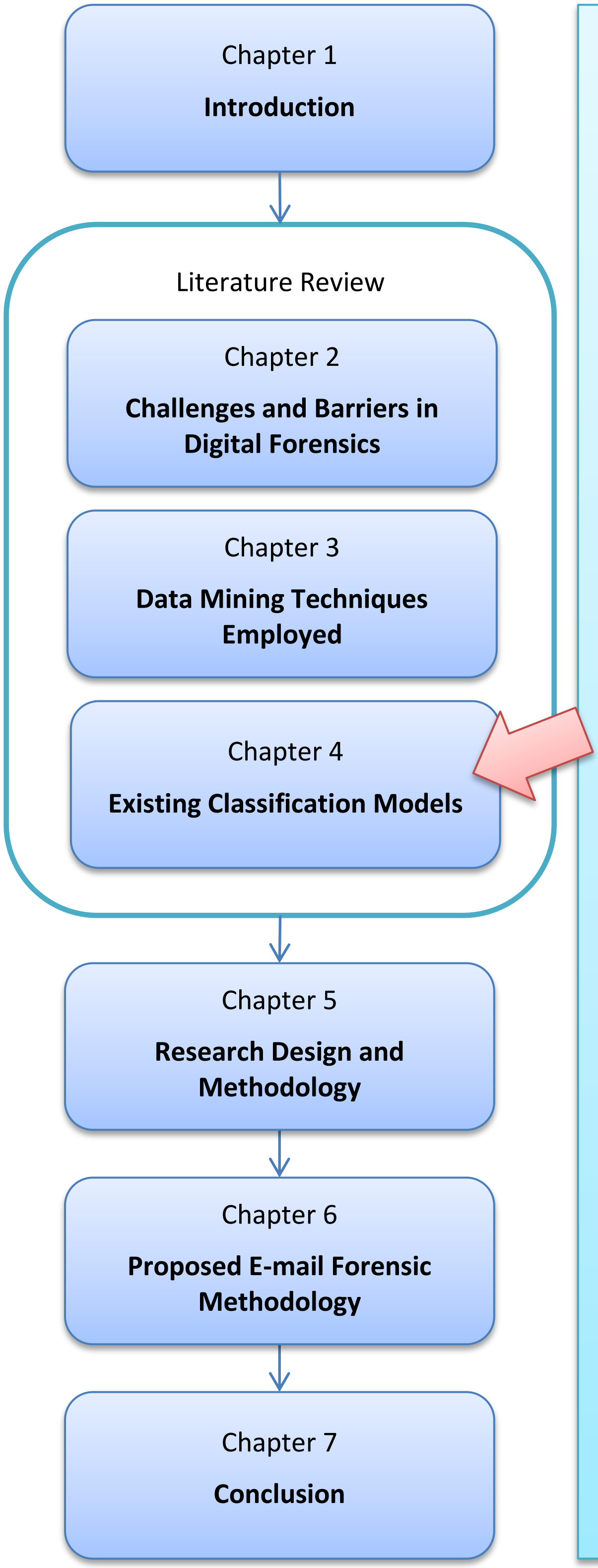
A new area of growth in the authorship identification field is that of an automated advanced learning machine technique called Support Vector Machine (SVM). This technique has a number of advantages over traditional techniques and many researchers have employed this method of data mining in their approach to author identification. When combined with a number of feature types, SVM outperforms other machine learning techniques and hence produces the highest accuracy for author identification.

New approaches to authorship attribution have been proposed; however these have not yet been employed in digital forensic investigations as there is a need to improve on the results and for the methods to be approved by the field experts. Hence further testing and refinement of these methods are needed in order to improve accuracy of the results as well as to improve credibility of evidence recovered via the method in question.

The tools available are varied and some are developed to perform a specific task; however, the most popular and widely used tools are the commercial tools such as EnCase and Forensic Toolkit. As these tools and methods form part of an overall process, it is necessary to examine the entire investigation process; therefore the next Chapter will discuss classification models and their application in the process of authorship identification.

4. EXISTING CLASSIFICATION MODELS

CHAPTER 4



Chapter 1

Introduction

Literature Review

Chapter 2

Challenges and Barriers in Digital Forensics

Chapter 3

Data Mining Techniques Employed

Chapter 4

Existing Classification Models

Chapter 5

Research Design and Methodology

Chapter 6

Proposed E-mail Forensic Methodology

Chapter 7

Conclusion

4.1. Introduction

4.2. A Basic Framework

4.3. An Examination of Existing Process Models

4.4. A Synthesis of Information

4.5. Conclusion

4.1 Introduction

The explosion of growth of technology and in particular in the computing world, has resulted in highly sophisticated equipment. This has in essence intensified the criminals' potential to perform criminal activity (Reith, et al., 2002). In light of this, law enforcement agencies have been busy trying to keep up with the criminal element that is persistent in abusing technology. In order for digital forensic investigators to perform their job, there are a number of steps that need to be well thought out and dealt with (Eloff, et al., 2006). These steps are encompassed in digital forensic classification models. There have been a number of models proposed demonstrating the complexity of the field. Each proposed model focuses on the investigative process or on a particular phase in the investigation (Eloff, et al., 2006).

To date there is a lack of agreement on the models (Broucek and Turner, 2002). The preliminary literature review identified the absence of an all-encompassing framework; consequently, there have been a number of models and frameworks developed. This Chapter examines, compares and contrasts methodologies that focus on digital forensic crime scene investigation. The frameworks, models and methodologies will be outlined and their processes described. Firstly the basic process of the digital forensic investigation is established.

4.2 A Basic Process Framework

The earliest framework or process of a digital forensic investigation was suggested in 1995 by Mark Pollitt in order to deal with potential digital forensic evidence (Pollitt, Personal Communication, 2010). The process defined the relationship between the science of recovering evidence with the law and set the tone for legally admissible evidence. The process of admitting evidence in a court of law was defined by the process of recovery of evidence and four distinct steps were defined in this methodology: Acquisition; Identification; Evaluation and Admission as Evidence. This early methodology set the scene for the digital forensic fraternity and many new models have been proposed since.

However as with many other disciplines, a definition is the starting point from which all other elements are derived. Likewise with digital forensics the starting point was a definition that was derived at the first Digital Forensic Research Workshop (DFRWS). This definition was the combined effort of a workshop whose primary focus was to establish digital forensics as a discipline and thereby provide a scope for and to build credibility (Palmer, 2001).

The definition provided by Palmer (2001): *"The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations"* resulted in a basic framework that was developed as a linear process; and the model is depicted in Figure 4 - 1.

The process was defined with the premise that practitioners utilise an investigative process in executing their duties, and that researchers use it as an aid in identifying shortfalls in technology (Palmer, 2001). Since the framework was established using the process that digital forensic investigators utilise, it was noted that not all processes were considered 'forensic'. The processes in green were subject to the least confusion but there was discussion as whether the processes of 'collection' and 'preservation' were sub-categories of each other.

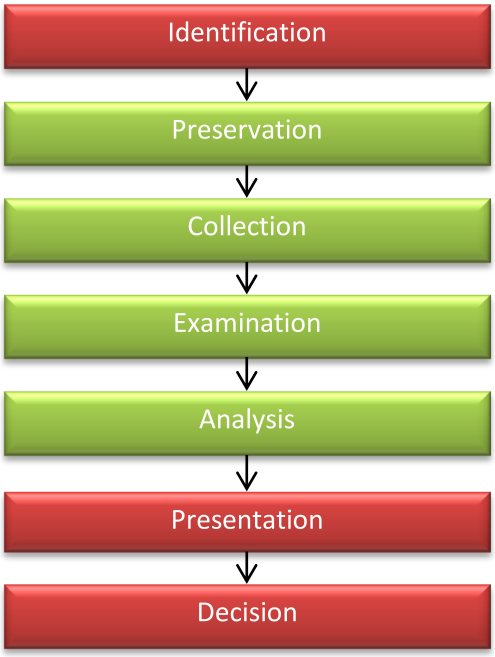


Figure 4 - 1 DFRWS Forensic Process (Palmer, 2001)

Many other definitions have been put forward over the years; however, it is not feasible to analyse each definition in this study; nonetheless the contribution made by other domains to the initial taxonomy by means of adding to the initial definition and the domain of digital forensics can be seen in Figure 4 - 2. Therefore, digital forensics has multiple contributing disciplines and this is what has increased the difficulty in defining a process for digital forensic investigators to follow as there are many factors to consider e.g. the technology (Information Systems and Computer Science) that must be examined in a digital forensic criminal case must be obtained lawfully (by law enforcement) with the individuals privacy in mind (Social Science) and examined within the bounds of the relative legislation (Law).

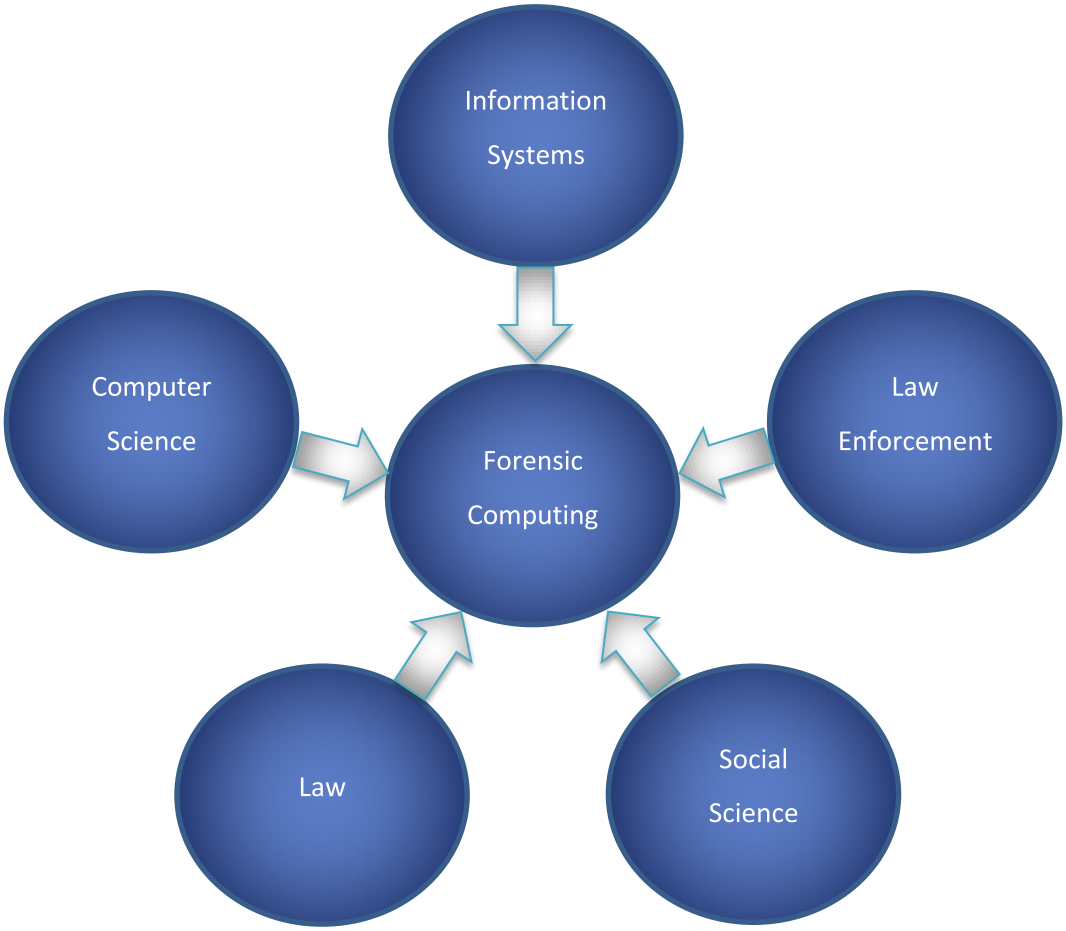


Figure 4 - 2 Forensic Computing Domain (Broucek and Turner, 2006)

The expansion of the domain started with the first DFRWS and the basic forensic framework, as seen in Figure 4 - 1, was a result of that; Reith et al. (2002) extended this basic model to include nine steps and this was aptly named the Abstract Digital Forensic Model depicted in Figure 4 - 3. This model was inspired by the DFRWS model and therefore shares many of its processes. This model is an abstract model as it does not define the technology involved and hence this allows for a standardisation of the forensic process to occur. The aim of the model was to determine the key aspects of the digital forensic process and in particular, the protocol for an FBI physical crime scene search (Reith, et al. 2002). This model also address many of the challenges within the digital forensic discpline i.e. Legal, societal

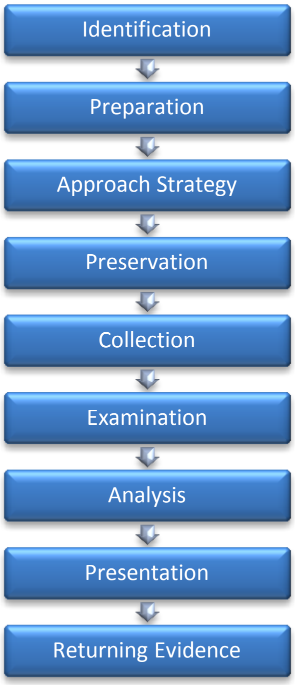


Figure 4 - 3 The Abstract Digital Forensic Framework (Reith, Carr, and Gunsch, 2002)

The first component of Abstract Digital Forensic Framework is the Identification of an incident from indicators and to determine its type. While this is not within the field of digital forensics, it is significant as it impacts other steps. The Preparation step is the preparing of tools, techniques and a search warrant as well as monitoring the authorisation and management support. The Approach strategy phase aims to collect as much untainted evidence as possible while minimising the impact on the victim. The Preservation phase aims to preserve all physical and digital evidence. The Collection phase is the recording of the physical crime scene and duplicating the digital evidence using standardised and accepted procedures. The Examination phase is the in-depth search of evidence related to the suspected crime and the construction of detailed documentation for analysis. The Analysis determines the significance and reconstructs fragments of data and draws conclusions from the evidence found. The Presentation phase is the summarised conclusions. The last step is the Returning of evidence to the proper owners and the removal of criminal evidence.

The steps in this model are modelled on the physical forensic processes; the additional steps of Preparation, Approach Strategy and Preservation identify further objectives of the digital forensic investigator. There is little difference between the Examination and Analysis phase and as with the original framework, these two processes can be grouped as one process. However, the main aim of this model is to aid in the standardisation of the digital forensic investigative process by creating a technology independent model.

Reith et al. (2002) state that additional sub-categories must be developed to identify sub-classes of digital technology e.g. method of collection will vary for different types of technologies. This model did not consider non-digital technologies therefore cannot be included for forensic analysis. Another disadvantage is that each sub-category added to the model will make it cumbersome to use. Although these disadvantages can be overcome, the model assumes that a strong chain of custody is maintained, and this must be addressed for any model to produce legally admissible evidence. Reith et al. (2002) argue that this is an assumption that any discussion of forensics involves a chain of custody but their model does not explicitly state this.

The creation of this framework was the basis for further work in the digital forensic discipline. The main aim was to establish digital forensics as a discipline and to improve on the founding principles. Hence the work performed at the first DFRWS can be called a success. The adaptation made by Reith et al. (2002) was to abstract the process from technology; many other forensic process models focus on the technology within each process. Although the model is independent of technology, it has opened other avenues of complications i.e. the sub-categories of technologies. Now that a basic framework of the digital forensic process has been established, the specific processes followed within different models will now be examined in-depth in the following section.

4.3 An Examination of Existing Process Models

This section examines the digital forensic process as defined by various researchers. Process orientated models are discussed followed by a framework that addresses the legal aspect specifically and then a live forensic model is discussed. Complex models are then examined and thereafter a comparison of the commonly reference models is outlined.

4.3.1 Process Orientated Models

Due to the nature of the digital forensic field, many digital forensic investigators follow a set of guidelines to enable successful discovery of evidence relating to the crime. It is for this reason that there are so many models for the digital forensic process. Digital forensic investigators follow a generalised methodology when conducting an investigation to ensure credibility and integrity of the digital devices (Arthur and Venter, 2004). The methodology is a stepwise process and is listed in Table 4 - 1. While this method is a sequential and a strict process, it ensures the integrity of evidence. All digital forensic investigators use a variation of this process although the overall method is similar.

The first step is to **Protect** the computer system from any harm during the forensic examination. This ensures the integrity of all the evidence on the system. This step is comparable with the Preservation step of Reith et al.'s (2002) abstract model. The second step is to **Discover** all files on the system deleted, hidden and encrypted. This **Discover** step is in line with the Examination step of the abstract model (Ciardhuain, 2004); however, this process is more specific in terms of identifying the objectives. The third step is to **Reveal** the content of all hidden and temporary files used by the programs and operating systems. This step can also fall into the Examination phase of the abstract model although it is specific in terms of stating files used by the operating system and programs. The fourth step is to **Access** the hidden files if possible and legally appropriate. The fifth step is to **Analyse** all relevant data found on the disk. The sixth step is to **Print** out an overall analysis of the computer system. The final step in the process is to **Provide** expert consultation as required in a court of law.

It can be seen by the description provided for each step of the generalised methodology that these steps are specific in terms of identifying the physical steps that must be taken, and these steps coincide with the digital forensic process followed by the digital forensic investigator. However, when compared with the Abstract Digital Forensic Framework, many steps in the generalised methodology are strictly defined. Although this is an advantage in terms of preserving the integrity of evidence, it also works adversely because these steps are difficult to follow when different digital devices are being examined and this is where the Abstract Digital Forensic Framework makes more sense i.e. This generalised methodology is useful when the steps are being followed during the process of examining the evidence, that is uncovering proof of the incident.

Furthermore, the generalised model is technical in nature and therefore technologically dependent and this is where the Abstract Digital Forensic Framework can be focused more closely. These steps can be incorporated into the Abstract Digital Forensic Framework from the collection phase until the presentation phase with the sub categories once they have been defined. But from a practical point of view not all forensic experts will proceed with a digital investigation in the same manner or utilise the same tools. Therefore it is reasonable to assume that many digital forensic investigators will adhere to their own process.

There are a number of models that have been proposed but not all follow the same process flow of information and evidence retrieval. Cardwell et al. (2007) divide digital forensics into three categories. The first step, namely **Litigation Support**, is the process of identification, collection, organisation and presentation of digital media while the second and third processes deal with the specific types of digital media i.e. **Digital Media Analysis** and **Network Investigations** (Cardwell, et al., 2007). Thus, it can be seen that in the first step of this model, the processes Identification and Collection are similar to Arthur and Venter's (2004) Discover and Recover processes. The second and third processes are very much specific in terms of technology and these steps are not directly comparable with any process in the Abstract Digital Forensic Framework.

Cardwell et al. (2007) deal with the methodologies in a practical way i.e. by detailing the steps in different categories; other methodologies include similar principles. One such methodology is the U.S. Department of Justice (U.S. DOJ) Process Model. The model consists of four phases and is an abstract model not specific to any technology or methodology and therefore is a generalised process, focusing mainly on core aspects (Reith et al., 2002). Hence this model will be more applicable in digital investigations as it can be adapted to the technology under examination. The processes are listed in Table 4 - 1. The steps of the U.S. DOJ Forensics Process Model are very similar to the Abstract Digital Forensic Framework.

Due to the abstract nature of the U.S. DOJ Forensics Process Model, one can directly compare this to the Abstract Digital Forensic Framework. It therefore allows for any gaps between the models to be identified and addressed accordingly. The Abstract Digital Forensic Framework is a more comprehensive process as more details about the forensic process are defined. A key feature of this model is that it ensures that preparation is performed at the beginning of the investigation and this is not addressed by the U.S. DOJ Forensics Process Model. The additional process of returning evidence to the owner as defined in the Abstract Digital Forensic Framework is another advantage. While this step may be assumed, it is important to explicitly state these steps in order to define a standardised process.

A very specific process model is that of Kruse and Heiser’s Methodology that includes three components, as seen in Table 4 - 1, ensuring the integrity of evidence during investigation (Eloff et al., 2006). There are a number of frameworks and methodologies that cover the digital forensic investigation differently and the above two are most commonly referred to in literature, and this adds to the complexity of the digital forensic process (Eloff et al., 2006). However, Lee, Casey, Reith, Carr and Gunsch are named as the most frequently quoted authors and their procedures are known to be the ‘standard’ procedures used during investigations (Ieong, 2006). Therefore, the need for a standardised process has become more evident.

Table 4 - 1 Classification Models

Classification Model	Processes/Categories/Classification
Generalised Methodology Arthur and Venter (2004)	• Protect • Discover • Recover • Reveal • Access • Analyse • Print • Provide consultation
Cardwell, et al. (2007)	• Litigation support • Digital media analysis • Network investigations
U.S. Department of Justice Forensics Cardwell, et al. (2007)	• Collection • Examination • Analysis • Reporting
Kruse and Heiser's Cardwell, et al. (2007)	• Acquiring the evidence • Authenticating the evidence • Analyzing the data
Forza Framework (Ieong, 2006)	• 6 questions: What, why, how, who, where, and when. • 8 roles: Case leader; System/business owner; Legal advisor; Security/system architect/auditor; Digital forensics specialist; Digital forensics investigator/system administrator/operator; Digital forensics analyst; Legal prosecutor
Liforac Model (Grobler & Von Solms, 2009b)	• Laws and regulations • Timeline • Knowledge • Scope
Lee, Palmbach and Miller (2001)	• Recognition • Identification • Individualisation • Reconstruction
Casey (2004)	• Recognition • Preservation, collection, documentation • Classification, comparison and individualisation • Reconstruction
The Digital forensic Research Workshop (Palmer, 2001)	• Identification • Preservation • Collection • Examination • Analysis • Presentation • Decision
Reith, Carr, and Gunsch (2002)	• Identification • Preparation • Approach Strategy • Preservation • Collection • Examination • Analysis • Presentation • Returning Evidence

Lee, Palmbach and Miller (2001) have proposed a model which consists of four stages focusing on the crime scene and not the entire investigation process; this model however does not extend to the electronic crime scene (Ciardhuain, 2004). Therefore this limitation will not allow for the preparation and presentation of evidence. Casey (2004) presents a model with four steps that is similar to Lee Palmbach and Miller (2001) but the model is only successful when applied to standalone systems and networked environments (Ciardhuain, 2004). During the Digital Forensic Research Workshop (DFRW) in 2001 a linear process model was developed; see Figure 4 - 1, which was driven by academia as opposed to law enforcement. This is important as there is no standardisation that has emanated solely from within the scientific community. This model is not comprehensive but is the basis for future work (Ciardhuain, 2004). This is an abstract model and is the key for a standardised process to be defined and proposes a model based on the DFRW model with some additional steps defined (Reith, et al., 2002); see Figure 4 - 3.

While standardisation of the digital forensic field is the focus, there are many aspects that must be considered as pointed out in Section 4.2. There are a number of contributing disciplines which have added to the complexity of the field. Even so, the models examined thus far have neglected to address the issue of the law. Due to the nature of digital forensics being interconnected with the law, all processes within the investigative framework must consider the legal issues surrounding it. This will be discussed in the next section.

4.3.2 Incorporation of Legal Aspects

Of the models examined, none explicitly state the legal implications of the processes; though Cardwell et al. (2007) name their first process ‘litigation support’ it is essentially the collection and presentation of evidence but is used in court cases. These models have been developed by traditional forensic scientists and this must be addressed as the challenges in the forensic discipline are changing. As stated in Chapter two (Section 2.2.4, pg. 24), the technological challenge faced is an evolving one; therefore the processes followed must be tailored or even adaptable to changes in technology. So too must the processes of the digital forensic investigation be adjusted or amended in order to comply with the legal challenges.

One such framework that addresses the legal requirement more thoroughly than the other models is a framework called FORZA (FORensics ZAchman) proposed by Ieong (2006). FORZA links all the common procedures as well as binding eight roles and responsibilities of individuals involved in the investigation process as illustrated in Table 4 - 2 (Ieong, 2006). Ieong (2006) argues that just as the IT Security field has a set of core values, namely Confidentiality, Integrity and Availability, so too should digital forensics. These fundamental principles are Reconnaissance, Reliability and Relevance (Ieong, 2006). These fundamentals incorporate the various processes within the digital forensic investigation as discussed below.

Reconnaissance essentially is the collection, recovery, decoding, discovery, extraction, analysis and conversion of data retained on different storage media or readable evidence (Ieong, 2006). The Reliability principle is the ability to extract, analyse, store and transport the evidence with emphasis on the chain of evidence as this enables evidence that cannot be repudiated or rebutted to be reliable and admissible for judicial review (Ieong, 2006).

Relevancy principles ensure that the evidence collected is relevant to the court case, and legal practitioners can advise digital forensic investigators as to what should be collected so that cost and time can be saved (Ieong, 2006). Ieong (2006) argues that the procedures of the digital forensic process must be bound to the practitioners of those procedures. The digital forensic investigation involves system owners, digital forensic investigators and legal practitioners, and Ieong (2006) further separates these into roles and responsibilities of these practitioners. The roles and responsibilities are depicted in Table 4 - 3.

Table 4 - 2 FORZA Framework (Ieong, 2006)

	Why (motivation)	What (data)	How (function)	Where (network)	Who (people)	When (time)
Case Leader (contextual investigation layer)	Investigation objectives	Event nature	Requested initial investigation	Investigation geography	Initial participants	Investigation timeline
System Owner (if any) (contextual layer)	Business objectives	Business and event nature	Business and system process model	Business geography	Organization and Participant relationship	Business and incident timeline
Legal Advisor (legal advisory layer)	Legal objectives	Legal background and preliminary issues	Legal procedures for further investigation	Legal geography	Legal entities and participants	Legal timeframe
Security/system architect/auditor (conceptual security layer)	System/Security control objectives	System information and security control model	Security mechanisms	Security domain and network infrastructure	Users and security entity model	Security timing and sequencing
Digital forensic specialists (technical preparation layer)	Forensics investigation strategy objectives	Forensics data model	Forensics strategy design	Forensics data geography	Forensics entity model	Hypothetical forensics event timeline
Forensic investigators/system administrator/operator (data acquisition layer)	Forensics acquisition objectives	On-site forensics data observation	Forensics acquisition/ seizure procedures	Site network forensics data acquisition	Participants interviewing and hearing	Forensics acquisition timeline
Forensic investigators/forensic analysts (data analysis layer)	Forensics examination objectives	Event data reconstruction	Forensics analysis procedures	Network address extraction and analysis	Entity and evidence relationship analysis	Event timeline reconstruction
Legal prosecutor (legal presentation layer)	Legal presentation objectives	Legal presentation attributes	Legal presentation procedures	Legal jurisdiction location	Entities in litigation procedures	Timeline of the entire event for presentation

Table 4 - 3 Digital Forensic Practitioners (Ieong, 2006)

Roles	Responsibilities
Case leader	Planning and orchestrating of the entire digital investigation process. Leads the case and determines whether it should proceed or not.
System/business owner	Owns the system being inspected. He/she is usually the victim and sponsor of the case. The Owner may also be the suspect
Legal advisor	The first legal practitioner that the case leader would seek for legal advice. He/she would advise the case leader whether it is applicable to proceed forward for legal disputes.
Security/system architect/auditor	Responsible for the system security structure and security controls design and implementation of security controls.
Digital forensics specialist	Planning of all operations within the investigation. Reconsiders all the inputs and requirements from legal advice to plan the entire investigation strategy. Decides whether it is necessary to contact third party vendors or an external consultant to perform specific part of investigation.
Digital forensics investigator/system administrator/operator	The main responsibilities of the investigator is to collect, extract, preserve and store the digital evidence from the systems.
Digital forensics analyst	Extract relevant data, analyse them against the hypothetical model proposed for investigation. Analysts may also have to perform various tests to prove/disprove the hypothetical model that emulates the case. They also have to reconstruct the timeline of the case based on the extracted data.
Legal prosecutor	Advises the case leader whether the collected evidence is sufficient, relevant, admissible and favourable to which party. Chooses the most suitable arena and leads the case onwards in the litigation process.

The interaction of these roles is shown diagrammatically in Figure 4 - 4. These roles may be performed by one individual or one individual may assume a number of different roles. Each role is placed in a layer that is connected through the six categories of questions, which has been derived from the Systems and Business Security Architecture (SABSA) framework (Ieong, 2006).

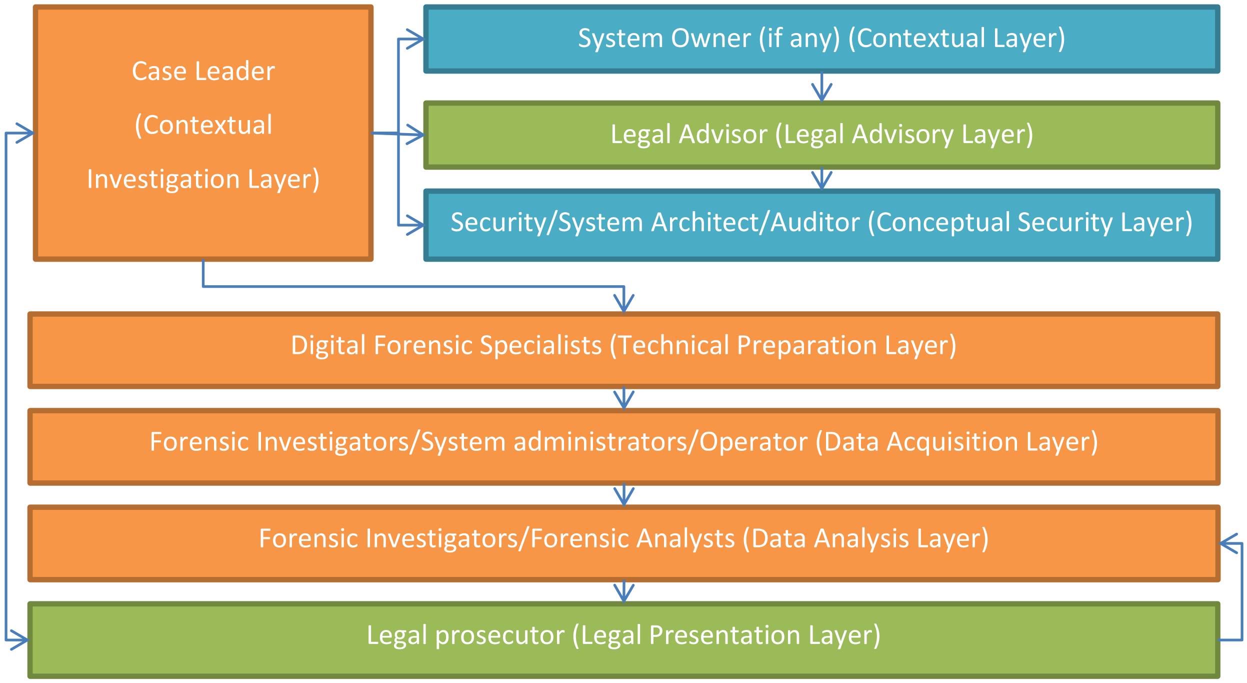

Case Leader (Contextual Investigation Layer)
System Owner (if any) (Contextual Layer)
Legal Advisor (Legal Advisory Layer)
Security/System Architect/Auditor (Conceptual Security Layer)
Digital Forensic Specialists (Technical Preparation Layer)
Forensic Investigators/System administrators/Operator (Data Acquisition Layer)
Forensic Investigators/Forensic Analysts (Data Analysis Layer)
Legal prosecutor (Legal Presentation Layer)

Figure 4 - 4 Process flow between the roles in digital forensics (Ieong, 2006)

In order to bind roles, responsibilities and procedures together, a technology-independent digital forensics investigation framework is required (Ieong, 2006). The defined roles are combined with the Zachman (Software Architecture) framework that poses six questions: What, why, how, who, where, and when, and together with the roles and responsibilities they produce the FORZA framework. The Zachman Framework is based on the Enterprise Architecture Framework which is a logical structure for classifying and organizing the descriptive representations of an Enterprise that are significant to the management as well as to the development of the Enterprise's systems (Zachman, 2004); (Pereira and Sousa, 2004).

Essentially the framework maps the different role players in the system development process to aspects of the process (Hay, 1997). However the Zachman Framework is not a methodology for the implementation of an object but it is the ontology for describing the enterprise and hence is a basis for architecture (Zachman, 2008). On the other hand the FORZA Framework establishes a greater level of interaction and level of responsibility for the legal discipline (Ieong, 2006). This is achieved through the legal prosecutors/advisor posing the six questions for example the 'What' questions that are part of the Legal presentation attributes are:

- What charge should be issued?
- What information should be included/excluded?

Therefore the legal advisor will provide advice to the investigative team. Ieong (2006) used a web hacking case to test the FORZA Framework, thus applying the various layers to the scenario. The FORZA framework is a practical and technologically independent framework that can be applied to real world scenarios and has also created a greater role for the legal fraternity to adopt during a digital forensic investigation.

The models discussed thus far have built on the original digital forensic investigative processes. Researchers have increasingly acknowledged the contribution of other disciplines and these have been reflected in the process models and proposed frameworks. A number of the models incorporate the same processes; however they are named differently, although not all frameworks approach the investigation process in the same manner.

Ieong (2006) proposed a unique framework in that it incorporates a new method of approaching an investigation. As new models become more diverse the problem of creating a standardised process becomes more apparent. Each new model is an attempt to overcome gaps in the process that existed in a previous model and other models such as Ieong's take a fresh look at the process to identify a new way forward. A gap that currently exists is the acceptance and practice of Live forensic acquisition. Hence, the following model addresses this aspect of Live forensic acquisition.

4.3.3 Adressing Live Forensics

There are limitations in the process models (Leigland and Krings, 2004) and four deficiencies were found at a digital forensic research workshop in 2001; procedural, technical, social and legal (Leigland and Krings, 2004). It is the scientific community's responsibility to standardise procedures and to certify individuals with a formal educational process (Meyers and Rogers, 2004). Many of the models that currently exist and that are widely used focus on traditional forensic acquisition of data; this is termed Dead Forensics or the forensic duplication method. Grobler and Von Solms (2009a) present a South African model for live forensic acquisition called Liforac which is illustrated in Figure 4 - 5. The Liforac model is practical and consists of four dimensions based on existing theories (Grobler and Von Solms, 2009a). The model is not a set of steps but rather a guideline for digital forensic investigators. However, the problem of inadmissibility is encountered as many courts do not accept live forensic evidence because of a lack of precedent and the innovative manner in which criminals exploit new technology (Grobler and Von Solms, 2009a).

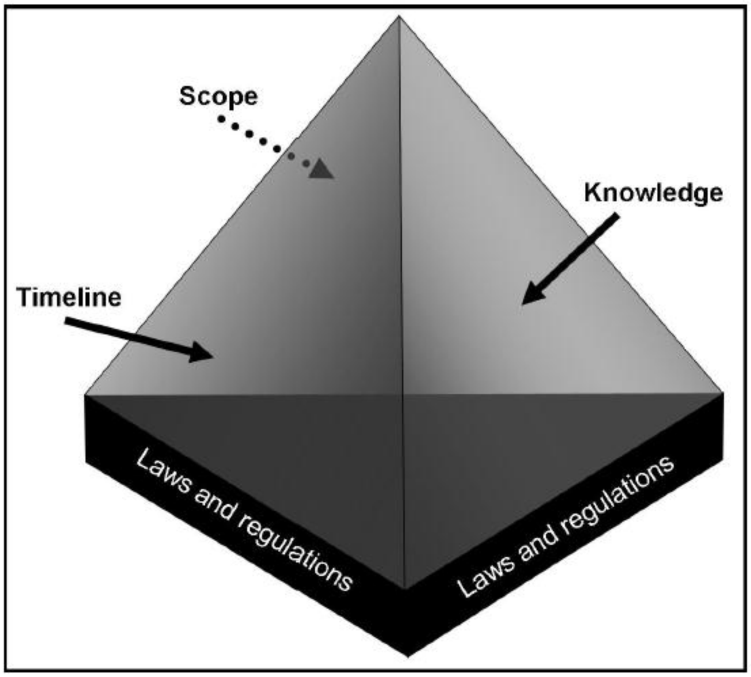


Figure 4 - 5 Liforac Model (Grobler and Von Solms, 2009b)

The Liforac model was not developed as a set of rigid steps but more as a set of broad guidelines to assist a digital forensic investigator (first responders) during acquisition of data (Grobler and Von Solms, 2009b). The four dimensions of the model are Laws and Regulations, Timeline, Knowledge and Scope and were determined by drivers that directed the decision to divide the model accordingly (Grobler and Von Solms, 2009b). The Laws and Regulations dimension is the basis for the model and considers what the digital forensic investigators need to know concerning the laws and regulations within the discipline (Grobler and Von Solms, 2009b). The Timeline Dimension addresses the sequence in which the digital forensic investigator needs to execute processes while the Knowledge dimension indicates the different levels of awareness and understanding digital forensic investigators must possess to perform Live forensics (Grobler and Von Solms, 2009b). The Scope dimension addresses practical problems related to Live forensics (Grobler and Von Solms, 2009b). Each dimension has sub-dimensions listed in Table 4 - 4.

Table 4 - 4 Liforac Dimensions and Sub-Dimensions (Grobler and Von Solms, 2009a)

Dimension	Sub-dimensions
Laws and regulations	➢ Common crime laws applicable to cyber crime ➢ Specific cyber laws. ➢ Court cases and precedents ➢ Definition of court admissibility
Timeline	➢ Implied processes ➢ Explicit processes ➢ Before the investigation ➢ During the investigation ➢ After the investigation
Knowledge	➢ Computer Science ➢ World Trends and Events ➢ Information Systems ➢ Social Sciences ➢ Forensic Sciences ➢ Law ➢ New technology
Scope	➢ Access to the machine ➢ Dependency on operating system ➢ Data modification ➢ Demonstrate the authenticity of evidence ➢ Court acceptance

Live forensics deals with the most volatile data first in order to preserve as much evidence as possible without compromising the data on the system. A clear distinction can be made between volatile and non-volatile data.

- ✓ Volatile data: all data that contains critical system details that provides the investigator with an insight as to how the system was compromised as well as the nature of the compromise. Examples include logged-in users, active network connections and the processes running on the system. (Lee, Savoldi, Lim, Park and Lee, 2009)
- ✓ Non-volatile data: reveals the status, settings and configuration of the target system, potentially providing clues to the method of compromise and infection of the system or network. Examples of this data include registry settings and audit policy.
  (Lee, Savoldi, Lim, Park, & Lee, 2009)

Live forensic acquisition of data has been growing as the need for different methods of acquiring data from different mediums arises. Lee, Savoldi, Lim, Park and Lee (2009) have proposed a new XML data collection framework for live forensics in Windows based computer systems called XLive. This framework has four components that comprise the architechture and these are:

- ❖ Scenario type analyser
- ❖ The data collection block
- ❖ The report manager
- ❖ The scenario generator

The framework is aimed at large scale digital evidence investigations and the automation of the collection of evidence allows digital forensic investigators to analyse terabytes of data. The framework can be adapted to network environments and the components of the framework can be adopted by other toolkits. The key advantage of this framework is that general live data which are not defined in any scenario, such as the RAM and page file dumps, network data, e-mails, and the list of processes can be collected (Lee, et al., 2009).

The problem encountered with the volatile data is that much of the data is low level and there is no general tool that could be used to manage the complex data, and for this reason a framework for the extraction and analysis of digital forensic data from volatile system memory called FATKit, was proposed (Petroni, Walters, Fraser and Arbaugh, 2006). FATKit is a cross platform debugger and is version independent. The platform and operating system are simply inputs into the system. It also offers the ability to analyse multiple images concurrently and to correlate information across machines regardless of dissimilarities between operating systems or hardware (Petroni et al, 2006). Now that models addressing specific aspects of the digital forensic investigation have been discussed, it is necessary to examine other models that focus on other aspects of the digital forensic process. This is necessary as it allows for a broader spectrum of model and frameworks to build a baseline.

4.3.4 Complex Models

This section examines two process models that address the sequence of the digital forensic investigation. Carrier and Spafford (2003) proposed a framework based on the physical investigation crime scene process. The phases of the physical crime scene have been applied to the digital crime scene and the Integrated Digital Investigation Process (IDIP) was developed as shown in Figure 4 - 6. The investigation process is segregated into five groups consisting of seventeen phases that focus on the reconstruction of events that led to the incident and stresses the importance of reviewing the entire task facilitating a faster examination.

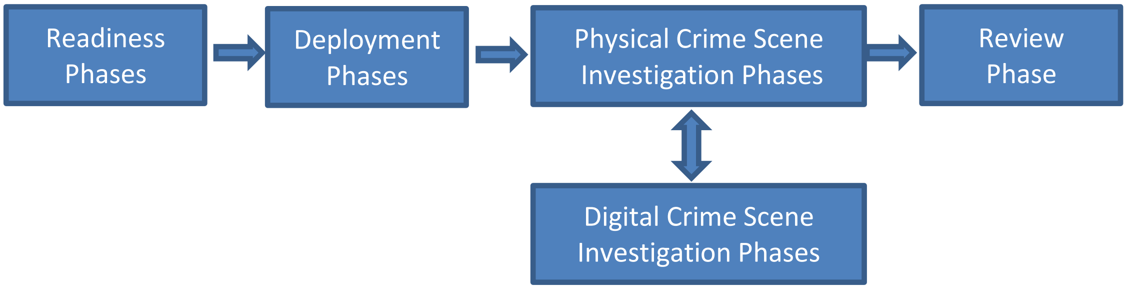

Readiness Phases
Deployment Phases
Physical Crime Scene Investigation Phases
Review Phase
Digital Crime Scene Investigation Phases

Figure 4 - 6 Phases of Integrated Digital Investigation Framework (Carrier and Spafford, 2003)

This model was further extended by Baryamureeba and Tushabe (2004) and was named the Enhanced Digitial Investigation Model (EIDIP). This model seperates the primary (focused on the computer) and secondary (focused on the physical crime) crime scenes and depicts the process as an iterative process as opposed to a linear process as shown in Figure 4 - 7. The deployment phase of the IDIP model was expanded to include the physical and digital crime scene and a new phase was added to trace back to the computer used to commit the offence (Baryamureeba and Tushabe, 2004). Additionally this model proposes that the reconstructiuon occurs after all investigations are completed in order to avoid any inconsistencies.

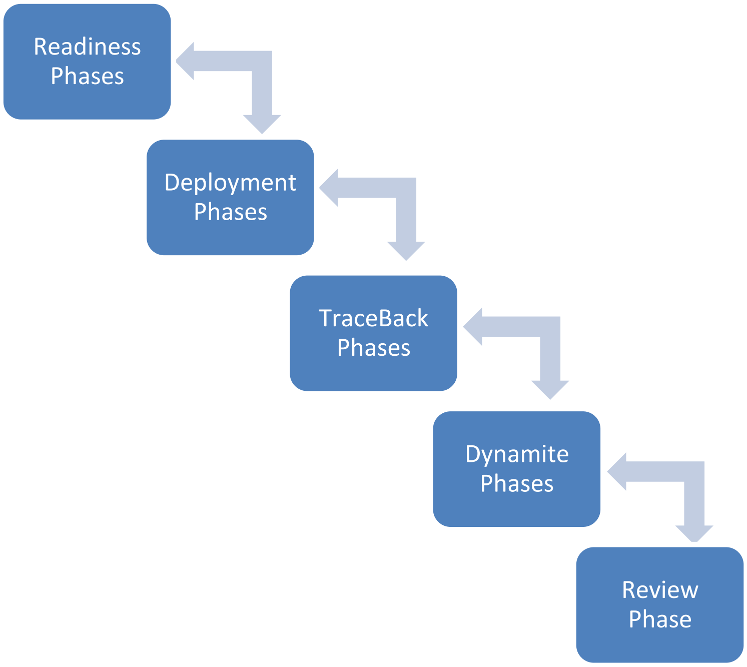


Figure 4 - 7 Enhanced Digital Investigation Process (Baryamureeba and Tushabe, 2004)

This section focused on the process models, frameworks and methodologies within the digital forensic discipline. Firstly simple process models were examined which covered the digital forensic process in a linear fashion according to the physical steps performed during an investigation, the prominent framework being the Abstract Digital Forensic Framework. Secondly the FORZA framework, which incorporates legal aspects, was examined. Furthermore this framework takes a unique perspective on the digital forensic investigation utilising the Zachman Framework. Thirdly Live forensic acquisition was addressed by highlighting the importance of data that can be recovered from a live system before it is shut down resulting in the loss of important data. Lastly two different models were examined each approaching the digital forensic investigation from a crime scene perspective. The next section is a comparison of the frameworks, models and methodologies examined and overlapping processes and unique aspects are identified.

4.4 A Synthesis of Information

A framework considered to be the most up to date is Ciardhuain's Extended Model of Cybercrime Investigations (Eloff, Kohn and Olivier, 2006). Each step within the framework is clearly defined and must be followed during the investigation process from the beginning during the preparation phase until the dissemination phase. This model explicitly represents the information flow in an investigation and captures the full scope of an investigation, rather than only the processing of evidence (Ciardhuain, 2004). The framework includes the following phases: awareness, authorisation, planning, notification, search and identify, collection, transport, storage, examination, hypotheses, presentation, proof/defense and dissemination. The framework also provides a basis for the development of techniques and tools to support the work of digital forensic investigators (Ciardhuain, 2004). Since this model is considered to be the most up to date model it it is beneficial to show the gaps that have been covered by the proposed changes.

Table 4 - 5 displays the comparison made with the most commonly referred to models to the most current model i.e. the Extended Cybercrime Investigation Model. The table is not a complete list of all models that have been examined. The table lists all the processes in the Ciardhuain Extended model and the table shows which of those processes corresponds with a process in the other models. Many of the models overlook the first three processes. This could be a possible reason for the lack of response by management leading to the inconsistencies that exist in digital investigations. The transport and storage phases are not considered by other models; these are crucial phases that need to be reflected in the investigation process as there is the potential for evidence to be damaged or tampered with.

The dissemination phase is also a separate phase that should be represented separately because once the investigation is complete it is necessary for the relevant information to be distributed to the correct individuals. This is also a crucial step because without the correct individuals receiving the relevant information timeously, there could be a delay in the legal proceedings if the evidence is to be utilised in a court of law.

Table 4 - 5 Comparison of Existing models to the Extended Cybercrime Investigation Model (Ciardhuain, 2004)

	Model				
Process	Lee	Casey	DFRWS	Reith, Carr, Gunsch	Ciardhuain
Awareness	✓				✓
Authorisation					✓
Planning					✓
Notification					✓
Search/Identify	✓	✓	✓	✓	✓
Collection	✓	✓	✓	✓	✓
Transport					✓
Storage					✓
Examination	✓	✓	✓	✓	✓
Hypothesis	✓		✓	✓	✓
Presentation	✓		✓	✓	✓
Proof/defence			✓		✓
Dissemination					✓

Accordingly there are issues that exist in these models. Selamat, Yusof and Sahib (2008) highlight three main issues identified from the examination of the frameworks. The issues are: process redundancies, area focus and framework characteristics. For example Baryamureeba and Tushabe (2004) and Reith, Carr and Gunsch (2002) have duplicated processes in their framework. Carrier and Spafford's framework is focused on the analysis of data in order to reconstruct past events (Selamat, Yusof and Sahib, 2008). Nonetheless these issues further emphasise the need for a standardised framework to be developed.

Selamat et al. (2008) propose a map of digital forensic investigation frameworks by grouping and merging processes that provide the same output. This is achieved through a process of identifying existing frameworks, constructing a phase name based on activities and process outputs and then finally mapping the processes of the different frameworks. The resultant map is shown in Table 4 - 6. The figure contains other models not examined here i.e. Beebe and Clark (2004), Kent et al. (2006) Rogers et al. (2006) and Freiling and Schwittay (2007).

The Framework Map shows that most of the frameworks contain the critical processes; however some frameworks do not contain phase one and five. This is important as identifying all processes of the digital forensic investigation allows for a holistic approach to be taken when developing a new framework. According to Selamat et al. (2008), the basis of a good digital forensic framework is that it should consist of all five phases.

Table 4 - 6 Digital Forensic Framework Map (Selamat, Yusof and Sahib, 2008)

	Phase	Phase 1	Phase 2	Phase 3	Phase 4	Phase 5
Digital Forensics Investigation Framework/Model	Pollitt, 1995		√	√	√	
	Palmer, 2001		√	√	√	√
	Reith et al., 2002	√	√	√	√	√
	Carrier and Spafford, 2003	√	√	√	√	√
	Stephenson, 2002		√	√	√	
	Baryamureeba and Tushabe, 2004	√	√	√	√	√
	Ciardhuain, 2004	√	√	√	√	√
	Beebe and Clark, 2004	√	√	√	√	√
	Carrier and Spafford, 2004	√	√	√	√	
	Kent et al., 2006		√	√	√	√
	Kohn et al., 2006	√	√	√	√	√
	K. Rogers et al., 2006	√	√	√	√	√
	Freiling and Schwittay, 2007	√	√	√	√	√
Output		Plan, Authorisation, Warrant, Notification, Confirmation	Crime Type, Potential Evidence Sources, Media, Devices, Events	File, Log Files, Events Log, Data, Information	Evidence, Reports	Evidence Explanation, New Policies, New Investigation Procedures, Evidence Disposed, Investigation Closed

4.5 Conclusion

In order for digital forensic investigators to perform the tasks necessary to recover digital evidence, they must follow steps that are predefined and contained within digital forensic classification models and frameworks. However, the problem encountered by digital forensic investigators is that there are a number of models that have been proposed and therefore, the task of following a predefined process has been complicated. The earliest methodology was the physical process of evidence recovery with the admission of the evidence in a court of law.

A definition given during the DFRWS is stated and the aim for digital forensics investigators was defined. Subject to some discussion the investigation processes were defined and this became the first methodology for digital forensics investigators to adopt. Additionally the contribution made by other disciplines was acknowledged and this highlighted the complexity of the digital forensic discipline. The initial methodology was further refined by creating a technology independent process and the abstract digital forensic framework was born in an attempt to standardise the digital forensic process. The main aim of the DFRWS was to create a standardised process for digital forensic investigators to follow and to establish digital forensics as a discipline.

A number of forensic process models and methodologies were examined in an attempt to assess the differences and the uniqueness of the methodologies. The drive of the digital forensic discipline is to create a technology independent model in order to accommodate a wider variety of evidence recovery efforts. As more models were proposed, more models included the legal aspects of the investigation as the main aim of a digital forensic investigation is the evidence that must be submitted in a court of law. A unique framework called FORZA illustrated how a framework that addresses the entire investigation process can be valuable. The framework creates a process flow of information and the responsibilities of the individuals within an investigation are assigned in order to create a greater chain of custody in order to address the concerns of the legal discipline.

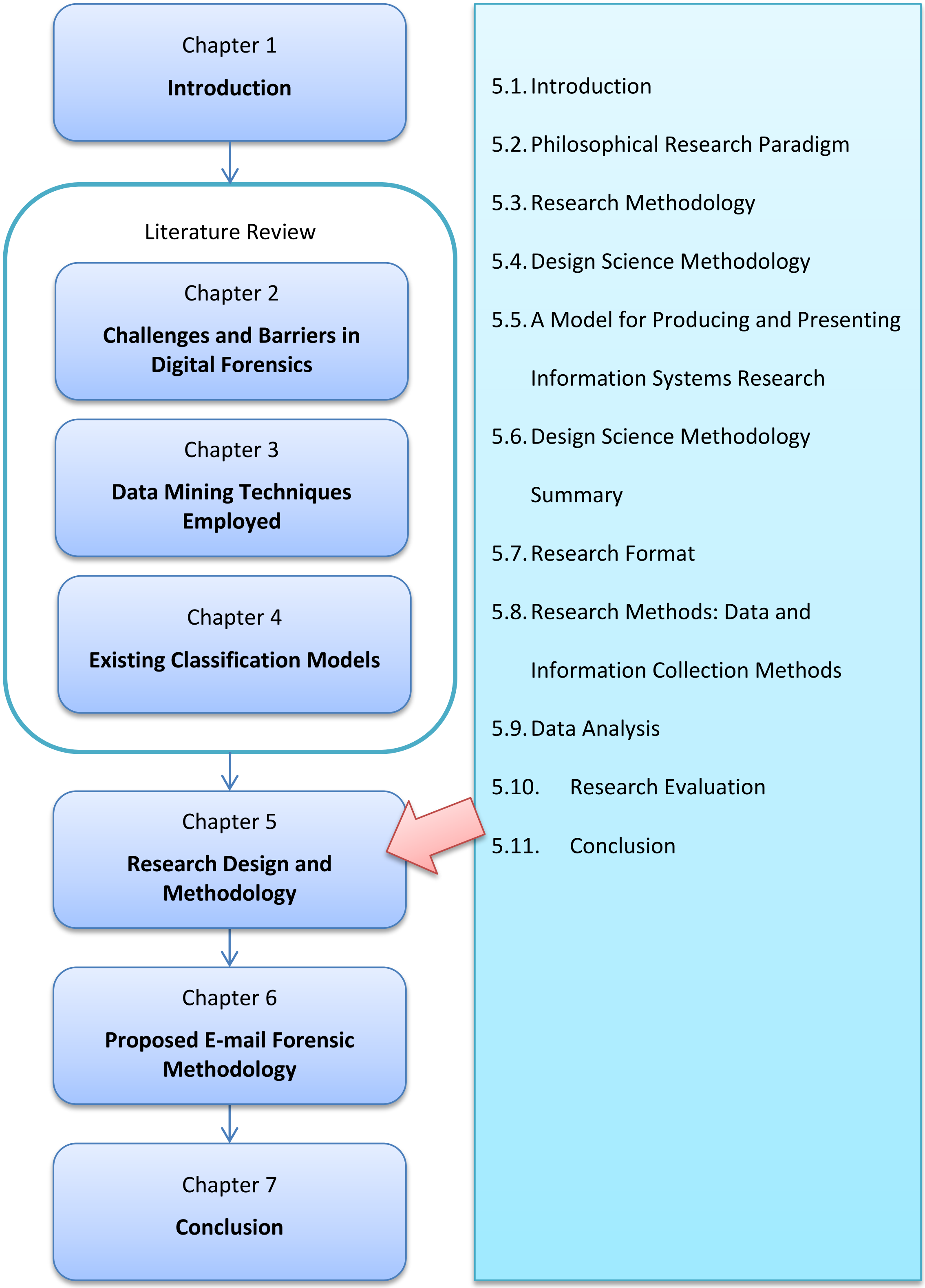
Due to the nature of digital forensics, various data types must be considered when recovering evidence and it is for this reason that many newer models now include an aspect of ‘live’ digital forensics. Essentially this allows the investigator to focus on evidence that may provide clues as to the nature of the infringement as well provide insight into how the system was compromised by accessing the volatile system memory. Additional models were examined and these were proposed in an attempt to reduce the inconsistencies that exist during the reconstruction process of the investigation. More specifically, the Liforac Model attempts to provide a guideline for first responders.

A comparison of the commonly referenced models with the most complete model was performed in an attempt to assess whether the gaps identified during the DFRWS have been covered and if the models being proposed contribute to the discipline. Finally, a map of the digital forensic process methodologies is presented. The aim of this is to establish a standardised process by eliminating redundancies and broadening the focus area of the models. Additionally this map provides a basis for a standardised digital forensic investigation process.

Whilst these models are being proposed and used in the digital forensic field, there is no best practice or standardisation of the procedures followed. Thus, many of the models are guidelines developed ‘ad hoc’ for performing investigations and this therefore highlights the importance of the standardisation of procedures and techniques used. The models discussed all focus on the processing of the digital evidence. However, there is a need to create a standard digital forensic methodology which will focus on the entire investigation process and the chain of custody. The following Chapter will describe the Research Methodology taken in developing the Proposed E-mail Forensic Methodology.

5. RESEARCH DESIGN AND METHODOLOGY

CHAPTER 5


Chapter 1
Introduction
Literature Review
Chapter 2
Challenges and Barriers in Digital Forensics
Chapter 3
Data Mining Techniques Employed
Chapter 4
Existing Classification Models
Chapter 5
Research Design and Methodology
Chapter 6
Proposed E-mail Forensic Methodology
Chapter 7
Conclusion

5.1 Introduction

Chapter four examined the various process models and methodologies that have been proposed. Many models contain similar processes derived from the physical digital forensic investigation process. Some models are unique in their approach to the digital investigation i.e. FORZA and Autominer, while others have been progressively enhanced e.g. Ciardhuain's Extended Process Model. A comparison was made of the most current model with other frequently referenced models in order to identify the gaps that were addressed. This Chapter will focus on the research design and methodology utilised in this study.

A discussion of the quantitative and qualitative research methods is given, in order to determine the best method for this study. The research method applied was influenced by the research project's objectives. By describing the theoretical aspects of the chosen method, the aim of this Chapter is to illustrate how the study was conducted and how the results were derived. It describes the way in which data is collected, measured and analysed and a discussion thereof is presented.

Empirical research methods are referred to by Moody (2002) as a collection of research methods where empirical observations or data are collected so that an answer to a particular research question may be obtained. This is the process that researchers undergo to obtain and apply relevant information in order to produce credible research. A research method is important for every researcher, as according to De Vos, Strydom, Fouché and Delport (2005) all research is conducted within a specific paradigm. Therefore, it is important that the correct research paradigm is chosen and properly described.

This Chapter is important to show the link between the chosen method, and how it addresses the research objectives. Since the research objectives and problems may vary, different methods can be utilised. This Chapter will deal with the methodological approach which will be presented, followed by a discussion concerning primary and secondary data collection methods adopted during this study. Finally, a brief description will be provided to show the approach used in the analysis of data, and to describe the credibility and dependability of this research project's findings.

5.2 Philosophical Research Paradigm

Due to the nature of research, all research is undertaken with some or other assumptions; according to Myers (1997) this is the basis for what constitutes 'valid' research and thereby determines what research methods are appropriate. Therefore the research method is affected by the selected research paradigm. Myers (1997) declares that in order to evaluate qualitative research it is necessary and important to know what the assumptions are. A paradigm is a set of beliefs that deals with ultimates or first principles (Guba and Lincoln, 1994). Guba and Lincoln (1994) argue that the basic beliefs of a paradigm can be summarised by the responses given to three fundamental questions and an answer given to any one question constrains how questions, taken in order, are answered. Guba and Lincoln (1994) utilise these questions to analyse the paradigms. The three fundamental questions are:

1. The Ontological question – what is the form and nature of reality and what is known about it?
2. The Epistemological question – what is the relationship between the knower and what is known?
3. The Methodological question – how can the inquirer go about finding what is believed to be known?

Guba and Lincoln (1994) suggest that there are four paradigms for qualitative research: positivism, post positivism, critical theory and contstructivism. However Myers (1997) adopts a threefold classification of the paradigms:

- Positivist,
- Interpretive
- and Critical.

Moreover Myers (1997) emphasises that whilst the three paradigms are philosophically distinct, in practice, they are not always clear cut. The differences between the paradigms is illustrated in Table 5 - 1.

Table 5 - 1 Differences between Paradigms (Myers, 1997)

Research Paradigm	Description	Aim
Positivist	Reality is objectively given and can be described by measurable properties independent of the observer	Test theories in order to increase the predictive understanding of the phenomenon
Interpretive	Assumption that access to reality is only given through social constructs e.g. language	Attempts to understand phenomenon through the meanings assigned by people
Critical	Assume that social reality is historically given and it is produced and reproduced by people	Main task is that of social critique in order to bring to light the restrictive conditions

The questioning and the synthesis of personal and somewhat subjective opinion is used to generate a forensic methodology. Hence this research project follows the interpretivistic paradigm and leans towards the intepretivistic portion of the continuum as indicated by Figure 5 - 1. Figure 5 - 1 illustrates the paradigms of research although it depicts the positivist and intepretivistic paradigms as two extremes on the continuum.

The extreme left denotes the area in which positivist or objective research paradigm is applied and followed. As one moves along the continuum, the assumptions of the one paradigm are relaxed and replaced by those of the other paradigm. This movement to the right is towards the interpretivistic research paradigm.

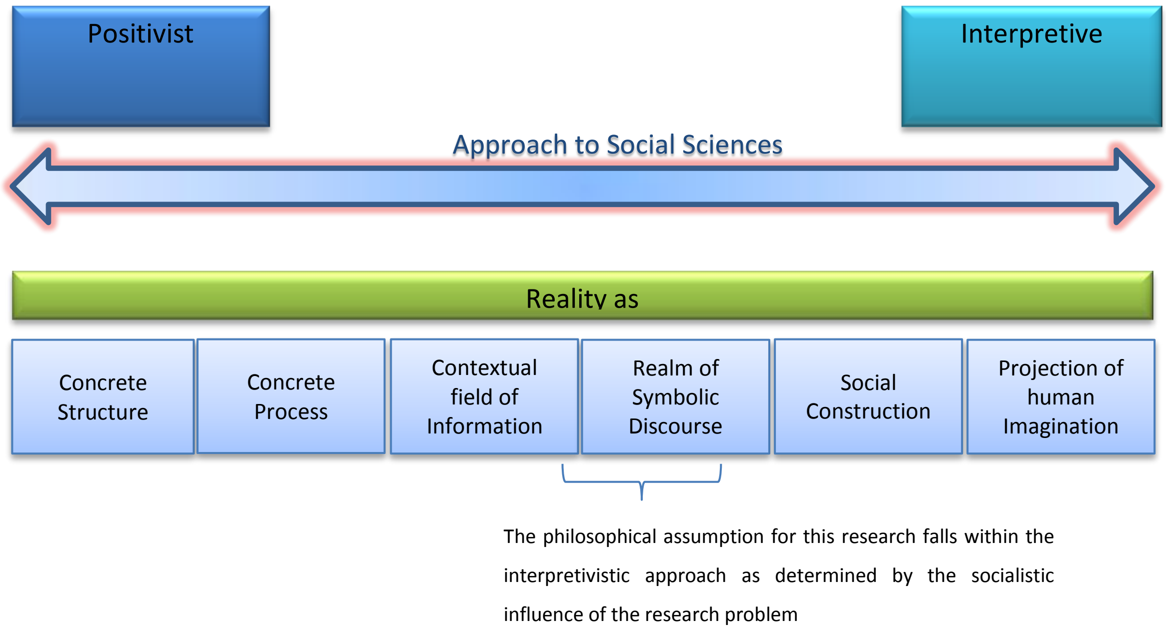

Positivist
Interpretive
Approach to Social Sciences
Reality as
Concrete Structure
Concrete Process
Contextual field of Information
Realm of Symbolic Discourse
Social Construction
Projection of human Imagination
The philosophical assumption for this research falls within the interpretivistic approach as determined by the socialistic influence of the research problem

Figure 5 - 1 Continuum of Core Ontological Assumptions (adapted from Morgan and Smircich, 1980)

The following section will now discuss the research methodology chosen and the reason for its implementation within this study.

5.3 Research Methodology

The purpose of this study is to develop a standardised e-mail forensic methodology for digital forensic investigators to follow using the qualitative method of research that has been derived from an interpretive research approach. For this purpose Design Science is used in order to generate an artefact (forensic methodology). A research method refers to the approach that is taken in conducting the research (Myers, 1997). A research design refers to the procedures that will be followed for collection and analysing data.

There is however two methods used to address a research project, namely quantitative and qualitative research methods and Myers (1997) indicates that this is the most common distinction made. The distinction is highlighted below.

- Quantitative Research – Myers (1997) adds that quantitative research methods were originally developed in the natural sciences to study natural phenomena and well accepted examples of these include survey methods, laboratory experiments and numerical methods. The objective of the quantitative research method is to obtain objective data that is measurable and can be analysed in a numerical manner.

- Qualitative Research – By contrast, qualitative research methods were developed in the social sciences to enable researchers to study both social and cultural phenomena of which examples include action research, the case study and ethnography. Qualitative research focuses on subjective data based on the opinions of the researchers and hence includes subjectivity.

Even though most researchers will perform either qualitative or quantitative research, there has been another approach that has been suggested by other researchers, that of triangulation (Myers, 1997). Essentially triangulation is a combination of one or more research methods. Myers (1997) also notes that there are other distinctions made about research methods e.g. subjective versus objective or etic (outsider) and emic (insider) perspective; however these are out of the scope of this research project.

Because of the nature of the digital forensic discipline this study will utilise the qualitative research method. Qualitative research is an appropriate research method because digital forensics is a growing discipline and many of the procedures followed cannot be measured quantitatively e.g. the digital forensic process and to an extent the recovery of digital forensic evidence. Furthermore, the opinions of the digital forensic experts will weigh heavily on the outcome of the Proposed E-Mail Forensic Methodology mainly because these experts can provide insight into the digital forensic investigation process, due to their implicit knowledge, something that cannot be achieved through the quantitative research method.

5.3.1 Delphi Technique

The Delphi Technique is applied within the qualitative research method. The Delphi technique was developed by Dalkey and Helmer (1963) at the Rand Corporation in the 1950s, and it is a widely used and accepted method for achieving convergence of opinion concerning real-world knowledge solicited from experts within certain topic areas. It is designed to be a group communication process that conducts detailed examinations and discussions (Hsu and Sandford, 2007). The technique is used to explore new concepts within and outside the Information Systems domain (Skulmoski, Hartman and Krahn, 2007). Common surveys try to identify "what is," whereas the Delphi technique attempts to address "what could/should be" (Miller, 2006).

The Delphi method is an iterative process that collects and distils the anonymous judgments of experts; no other data gathering and analysis techniques employ multiple iterations designed to develop consensus on a specific topic (Hsu and Sandford, 2007; Skulmoski, Hartman and Krahn, 2007). The iteration process is a means of offsetting the shortcomings of a conventional pooling of opinions (Powell, 2003). The iterative Delphi technique process is demonstrated in Figure 5 - 2.

The Delphi technique usually consists of a panel of experts in a specific domain and is characterised as a method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem (Okoli and Pawlowski, 2004). The size of the panel can vary from a few to 50 experts, but this depends on the number of experts available; however in Brockhoff's study of Delphi performance (1975 in Colton and Hatcher, 2004) suggests that groups with eleven participants were more accurate than larger groups; furthermore he found that for fact finding questions for dissertations, groups of seven had a higher performance (Colton and Hatcher, 2004).

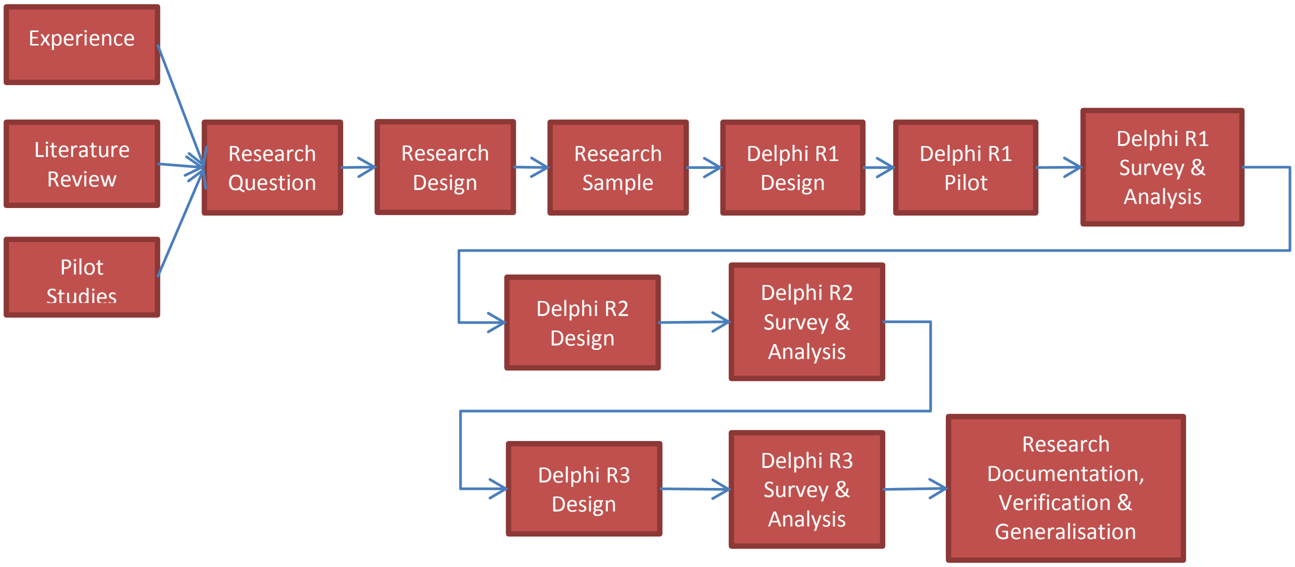


Figure 5 - 2 Three Round Delphi Process (Skulmoski, Hartman and Krahn, 2007)

Powell (2003) adds that the panel size does not need to be a representative sample for statistical purposes as the quality of the experts outweighs the numbers. Therefore it can be reasoned that one expert's opinion on a specific topic outweighs many responses received during a random questionnaire survey (Powell, 2003). Additionally Skulmoski, Hartman and Krahn (2007) add that the Delphi technique is well suited to the rigorous capture of qualitative data. According to Adler and Ziglio (1996 in Skulmoski, Hartman and Krahn, 2007) the participants should meet four "expertise" requirements:

- knowledge and experience of the issues under investigation
- capacity and willingness to participate
- sufficient time to participate in the Delphi
- effective communication skills

The Delphi method is a flexible research technique well suited when there is incomplete knowledge about phenomena (Skulmoski, Hartman and Krahn, 2007). Additionally it is not only a quantitative method but it is very suitable for qualitative research. Therefore the Delphi technique is well suited for this research project based on the above justifications. As the main aim of this study is to produce a standardised digital forensic process for e-mail authorship, a relatively new research methodology will be applied i.e. Design Science.

5.4 Design Science Methodology

This section describes the research methodology chosen for this study and a discussion on Design Science. This study follows the Design Science methodology, including research methods i.e. data collection and data analysis. Design Science attempts to expand human potential by creating ground breaking artefacts (Hevner and March, 2003). Benbasat and Zmud (1999 in Hervner and March, 2003) argue that the relevance of Information Systems (IS) research is directly related to its applicability in design, stating that empirical IS research should be implementable or stimulate critical thinking among IS practitioners. Furthermore, it is complex to design useful artefacts because of the advances in the domain areas. Design Science is a problem solving process and the fundamental principle is that the understanding of a problem and its solution are acquired in the building of an artefact (Hevner and March, 2003).

Hevner and March (2003) have proposed a conceptual research framework in order to assist understanding, execution and evaluation of Information Systems research, which is illustrated in Figure 5 - 3. This framework is used to assess what is being produced from each concept against each other in the context of business needs. The realm of IS research is at the convergence of people, organisations, and technology (Davis and Olson 1985 and Lee 1999 as cited in Hervner and March, 2003). Information Systems/Information Technology artefacts are broadly defined as *constructs* (vocabulary and symbols), *models* (abstractions and representations), *methods* (algorithms and practices), and *instantiations* (implemented and prototype systems) (Hevner and March, 2003); (March and Smith, 1995) .

Design Science creates and evaluates IT artefacts intended to solve identified organisational problems. As field studies enable behavioural science researchers to understand organisational phenomena in context, the process of constructing and exercising innovative IT artefacts enables design science researchers to understand the problem addressed by the artefact and the feasibility of their approach to its solution (Hevner and March, 2003). The design process is a sequence of expert activities that produces an innovative product (i.e. the design artefact). The evaluation of the artefact then provides feedback information and a better understanding of the problem in order to improve both the quality of the product and the design process. This build-and-evaluate loop is typically iterated a number of times before the final design artefact is generated.

In Figure 5 - 3 the environment is the space where the problem resides, and in the information systems discipline this is made up of the organisation, people and technology. And together these represent the business' needs and hence the 'problem' (Hevner and March, 2003). These business problems are addressed by IS research. Hevner and March (2003) argue that just as behavioural science addresses the problem by developing and justifying the theories that explain the problem, Design Science addresses the problem by building and evaluating artefacts that are designed to meet the business problem. Hevner and March (2003) draw the comparison between behavioural science and Design Science and reason that the goal of behavioural science is truth and the goal of Design Science is utility, and as argued above, truth and utility are inseparable and therefore an artefact may have some utility due to some undiscovered truth.

Furthermore behavioural science tries to understand reality whereas Design Science attempts to create things to serve a human purpose (March and Smith, 1995). This then leads to the reassessment and refinement of the process in future research (Hevner and March, 2003). Information System research draws on a knowledge base and this is composed of foundations and methodologies and the appropriate and applicable use of methodologies is achieved through rigor (Hevner and March, 2003).

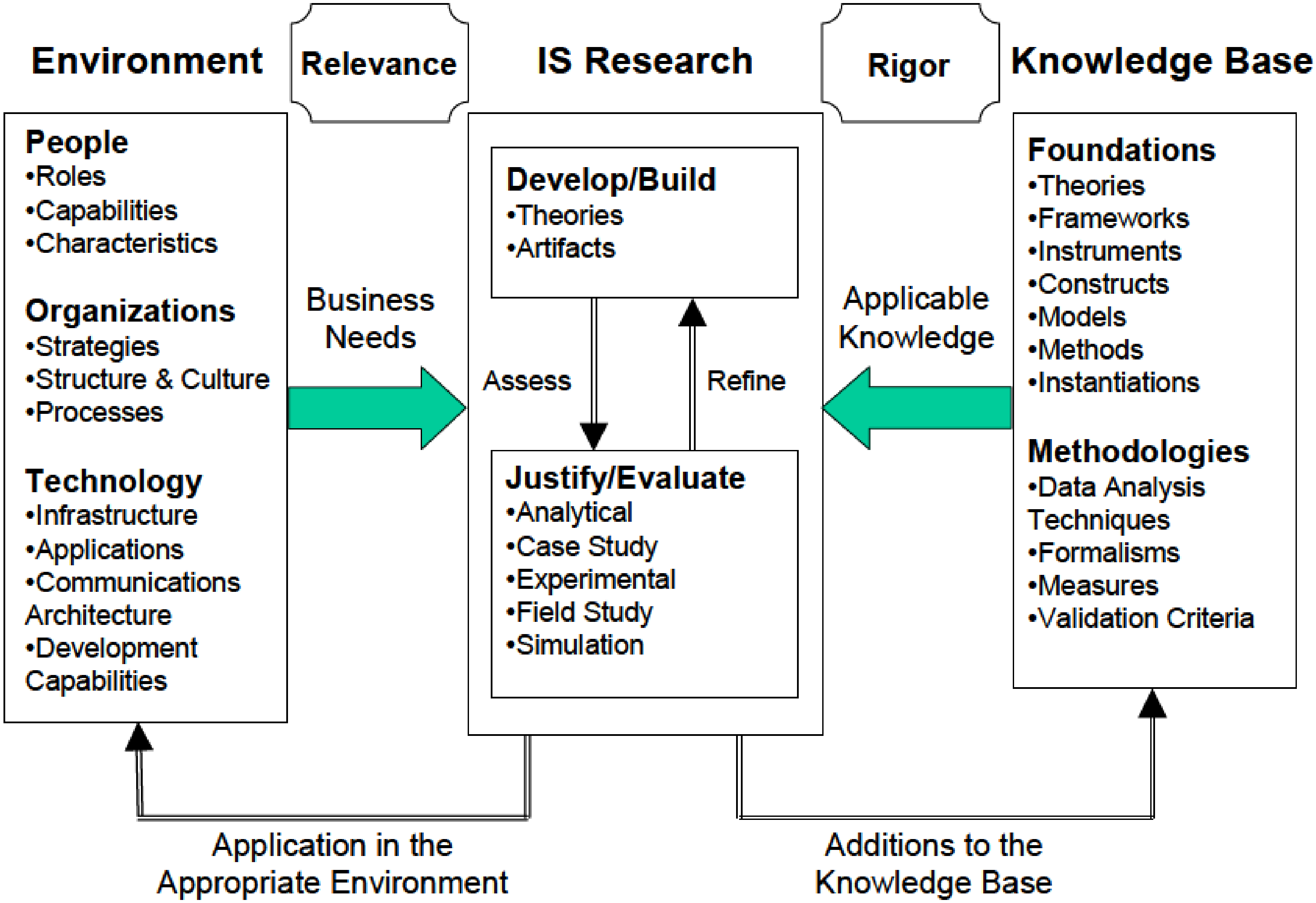


Figure 5 - 3 Information System Research Framework (Hevner and March, 2003)

An important aspect of Design Science is emphasised by Hevner and March (2003) and that is that Design Science addresses important unsolved problems in a unique and innovative way or alternatively solves problems in a more effective and efficient way. It is for this reason that Design Science was chosen to address the problem of standardising the digital forensic investigative process. With the overall framework in mind Hevner and March (2003) have presented a set of guidelines in order to assist Information System researchers because the theory of design in information systems is a constant state of revolution (Kuhn, 1996 in Hevner and March, 2003). These seven research guidelines are presented in Table 5 - 2.

Table 5 - 2 Research Guidelines (Hevner and March, 2003)

Guideline	Description
Design as an Artefact	Design science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation.
Problem Relevance	The objective of design science research is to develop technology-based solutions to important and relevant business problems.
Design Evaluation	The utility, quality, and efficacy of a design artefact must be rigorously demonstrated via well-executed evaluation methods.
Research Contributions	Effective design science research must provide clear and verifiable contributions in the areas of the design artefact, design foundations, and/or design methodologies.
Research Rigor	Design science research relies upon the application of rigorous methods in both the construction and evaluation of the design artefact.
Design as a Search Process	The search for an effective artefact requires utilizing available means to reach desired ends while satisfying laws in the problem environment.
Communication of Research	Design science research must be presented effectively both to technology-oriented as well as management-oriented audiences.

The following sub-section identifies and explains a Design Science Information System research model developed to guide Information System researchers.

5.5 A Model for Producing and Presenting Information Systems Research

The information systems research framework as developed by Hevner and March (2003) illustrates how IS research is conducted and gives overall guideline as to how IS research should be conducted. The model has been developed by Peffers et al. (2006) as was briefly discussed in Chapter two. Peffers et al. (2006) looked to influential research to determine what common processes could be found in the literature; however, they note that some of the literature examined had different processes in mind and hence was not trying to compare apples and oranges but rather to draw out common processes. From this process Peffers et al. (2006) Design Science Reseach Process was developed and is illustrated in Figure 5 - 4 and a discussion on the model's activities follows:

1. Problem identification and motivation

   This activity involves the definition of the research problem and the justification of the value of the solution. If the problem is conceptualised an effective artefact can capture the problem's complexity (Peffers, et al., 2006). Peffers et al. (2006) explain that the justification of the value of a solution accomplishes two things, namely: it motivates the researcher and the audience to pursue the solution and to accept the result and also adds to the understanding of the researcher's perception of the problem.

2. Objectives of a solution

   The objectives of the solution must be derived from the problem definition and these may be quantitative or qualitative.

3. Design and development

   An artefactual solution is broadly defined as constructs, models, method or instantiations (Hevner and March, 2003). The appropriate artefact is defined by the desired functionality of the solution.

4. Demonstration

   This activity involves the demonstration of the artefact's ability to solve and define the problem. This could be in the form of an experiment, case study, proof or an appropriate activity.

5. Evaluation

   This activity involves the observation of how well the artefact supports a solution to the problem. The evaluation can be performed in terms of the objectives identified in activity two.

6. Communication

   Peffers et al. (2006) state that the problem, its importance and that the artefact, its utility and novelty, the rigor of its design, and its effectiveness to researchers and other relevant audiences, such as practising professionals, when appropriate must be communicated. Peffers et al. (2006) add that this could be in the form of scholarly research publications.

Peffers et al. (2006) state that this model is structured in a nominally sequential order although there is no expectation that the researcher should always proceed in a sequential order. Hence, the model caters for entry into the research process at various points as depicted in Figure 5 - 4**Error! Reference source not found.**. For example a design and development centered approach would start with activity three and would result from the existence of an artefact that has not yet been formally thought through as a solution for the explicit problem domain in which it will be used, and which may have come from another research domain used for a different problem. In this scenario the researcher will apply the process retroactively and start at activity four.

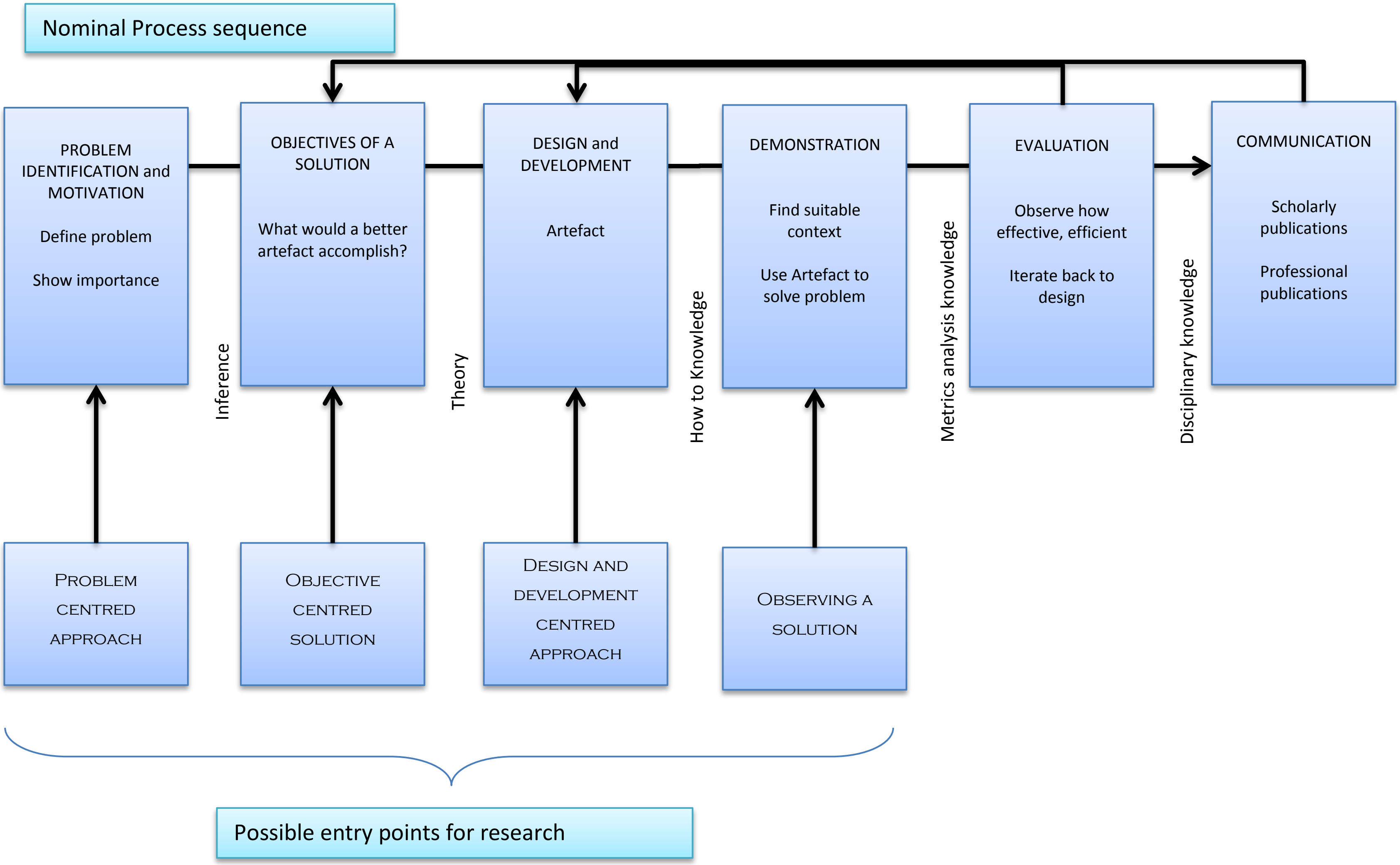


Figure 5 - 4 Design Science Research Process Model (Peffers, et al., 2006)

5.6 Design Science Methodology Summary

The guidelines set out by Hevner and March (2003) provide a roadmap for the researcher to follow in order to overcome the lag between academic research and the adoption by industry. Peffers et al. (2006) have proposed a mental model for researchers to follow in an attempt to address the limitations in the current literature. The proposed research model is sufficiently complete and robust and serves as a template for future research. This study follows the guidelines proposed by Hevner and March (2003) and more closely attempts to utilise the conceptual model proposed by Peffers et al. (2006).

This study adopts the Design Science as part of the research methodology and the resultant proposed solution is a artefact in the form of a methodology (i.e. the Proposed E-Mail Forensic Methodology) derived from previous research and has been refined by the process of interviews and open critisism from experts in the digtal forensic domain. Based on the research methodology chosen the best method fit for this study was determined.

In this instance interviews with experts in the digital forensic domain and previous research are the main sources of data. The expert review method is the best warranted method for this study as it considers many opinions of various experts and researchers who of necessity cannot meet physically and within reasonable costs of the study.

5.7 Research Format

There are various types of research formats namely: descriptive, explanatory and explorative (Collis and Hussey, 2003). Explorative research is defined as new research where there has been little or no research about the research problem. Explanatory research demonstrates a change in one variable causing a predictable effect on another variable. Descriptive research generally answers problems related to ‘what’ questions and describes a phenomenon as it exists (Collis and Hussey, 2003). Since this research examines existing literature through the secondary data and addresses the interviews through primary data, the descriptive method is the most applicable research format. Additionally the facts are clearly defined and the problem has been well identified by previous researchers and hence the Proposed E-Mail Forensic Methodology is derived through the descriptive research format.

Arguments in a research project can be categorised into inductive and deductive reasoning. Reasoning is the cognitive process of drawing inferences from given information (Goel and Dolan, 2004). In a deductive argument the conclusion asserts no more than what is contained in the premise while in an inductive argument more is inferred in the conclusion than is contained in the premise (Cothran, 1999). Furthermore valid deductive arguments offer sufficient proof for their conclusions and tend to be argued from a general perspective to a specific perspective. This study adopts the inductive reasoning approach. Additionally the argumentative method is philosophical in nature and will, therefore, in its pursuit of knowledge, equip society with an awareness and a better understanding of the concepts, as they are used in a variety of contexts, such as sciences, morals, history and the like (Hirst, 1974). Hamm (1989) has correctly argued that this discourse will, in the end, ensure that everyone has a better and clear understanding of the issues at stake.

5.8 Research Methods: Data and Information Collection Methods

There a numerous techniques that can be applied to a research project in order to collect data and information. Research studies utilise one or more techniques in an attempt to assess the research problem and in order to create some form of solution. There are two types of data sources applicable to research projects namely: primary and secondary data sources. Most researchers utilise both sources to meet the demands of their research project. A primary data source refers to data which is unpublished and which the researcher has gathered from the participants directly. Secondary data sources are any previously published materials, such as books and journal articles. This study utilises both primary and secondary data sources.

5.8.1 Primary Data Collection

The primary data sources utilised in this research project are interviews, more specifically expert review. Hence this is a qualitative research technique.

5.8.1.1 Interviews

The reason that interviews were selected as a primary data source was because in the digital forensic discipline experts in the domain have implicit knowledge of the processes and techniques applied during an investigation. This type of knowledge is difficult to transfer and therefore a deeper understanding of the problem can be gauged from the interview process. There are two types of interviews i.e. structured and unstructured interviews. Both types were used during the primary data collection process of this research project.

A predefined list of questions was organised and formed part of the structured interview. Additionally the unstructured interview was held in order to gain further insight into the responses given by the interviewee. A large portion of the interviews was unstructured due to the responses obtained from the interviewee. An Informal survey was used to obtain feedback from experts in the digital forensic field in order to further refine the proposed model.

These 10 experts were presented with the study's findings and asked to reflect and comment on the findings as a step to further refine the Proposed E-Mail Forensic Methodology. This approach also corresponded to the Delphi technique which follows an iterative cycle of refinement. The first step was validated by three experts who provided a critical review of the Proposed E-Mail Forensic Methodology. However as indicated in section 5.8.3 the population was relatively small and many experts live abroad and therefore the second round of interviews was held at a Information Security Conference. A third round of reviews was performed using South African experts who formed part of the Special Investigation Unit. During each round of the technique, feedback was obtained and this was used in the refinement of the Proposed E-Mail Forensic Methodology.

5.8.2 Secondary Data Collection

Data collected by another person is termed secondary data. The secondary data examined in this study included literature survey of internet sources, frameworks, methodologies, journal articles, past research projects, reports as well as books. The literature survey was performed initially to determine the problem and the research objectives. This is the most important aspect of the secondary data collection phase as this is where the body of knowledge in the discipline is expanded upon.

5.8.3 Population and Sample of Experts

Since the Information Security population is relatively unknown and due to the interpretive nature of this study, the sample size of the population is relatively small. Many of the experts that were consulted were gathered at a conference, many of whom are internationally recognised and ordinarily would not be able to provide feedback because of their busy schedules. It is for this reason that most of the experts were consulted at this venue. Moreover all experts consulted met the 'expertise' criteria discussed in Section 5.3.1. Additional experts who reside in South Africa were consulted separately and this added an extra validation and refinement step to the Proposed E-Mail Forensic Methodology.

5.9 Data Analysis

The qualitative data from the experts was summarised and changes were made according to the feedback of the experts as a further stage in refining the proposed solution. Their feedback either supported or opposed the proposed solution and this added to the integrity of the project.

5.10 Research Evaluation

During a research project, errors can be communicated during the process of gathering information and therefore the researcher must remain vigilant in order to maintain the integrity of the research project. Hence research evaluations assist in maintaining the credibility and integrity of a research project. Lincoln and Guba (1985 in Oates, 2006) highlight a set of criteria for interpretivist research that is an alternative to, but parallel to, those for the positivist approach and this is highlighted in Table 5 - 3.

Table 5 - 3 Quality in positivist and interpretivist research (Lincoln and Guba, 1985 in Oates, 2006)

Positivism	**Interpretivism**
Validity	Trustworthiness
Objectivity	Conformability
Reliability	Dependability
Internal validity	Credibility
External validity	Transferability

The following evaluation applies to the research project. The expert reviews carried out through the Delphi technique was evaluated in terms of the trustworthiness of the expert as well as the information supplied. The credibility of the expert weighed heavily on the results of the information obtained i.e. an expert with far more experience and in-depth knowledge was assigned a higher weighting than an expert with less experience and knowledge. Hence this added to the dependability and trustworthiness of the expert review obtained which translated into the proposed digital forensic e-mail methodology.

Additionally the dependability of the research is substantiated by the literature review where well known and established literature was utilised. To an extent conformability and transferability have been met in the proposed digital forensic e-mail methodology as the process of gathering credible literature and utilising expert opinion to develop a standardised procedure for e-mail investigations can be repeated for another type of digital forensic investigation type i.e. recovery of data from smart-phones. Since all five areas on interpretivistic research have been met to varying degrees, the credibility and dependability of the research project has been demonstrated and this contributes to the soundness of the study.

5.11 Conclusion

This Chapter described the research methodology used in this research project. A description of the two methods of research was described i.e. qualitative and quantitative research and a further method called triangulation was briefly discussed. The research method adopted during this research was a qualitative approach and because the main aim of the research was to provide a standardised approach to a digital forensic investigation, a relatively new paradigm of digital forensics was also adopted.

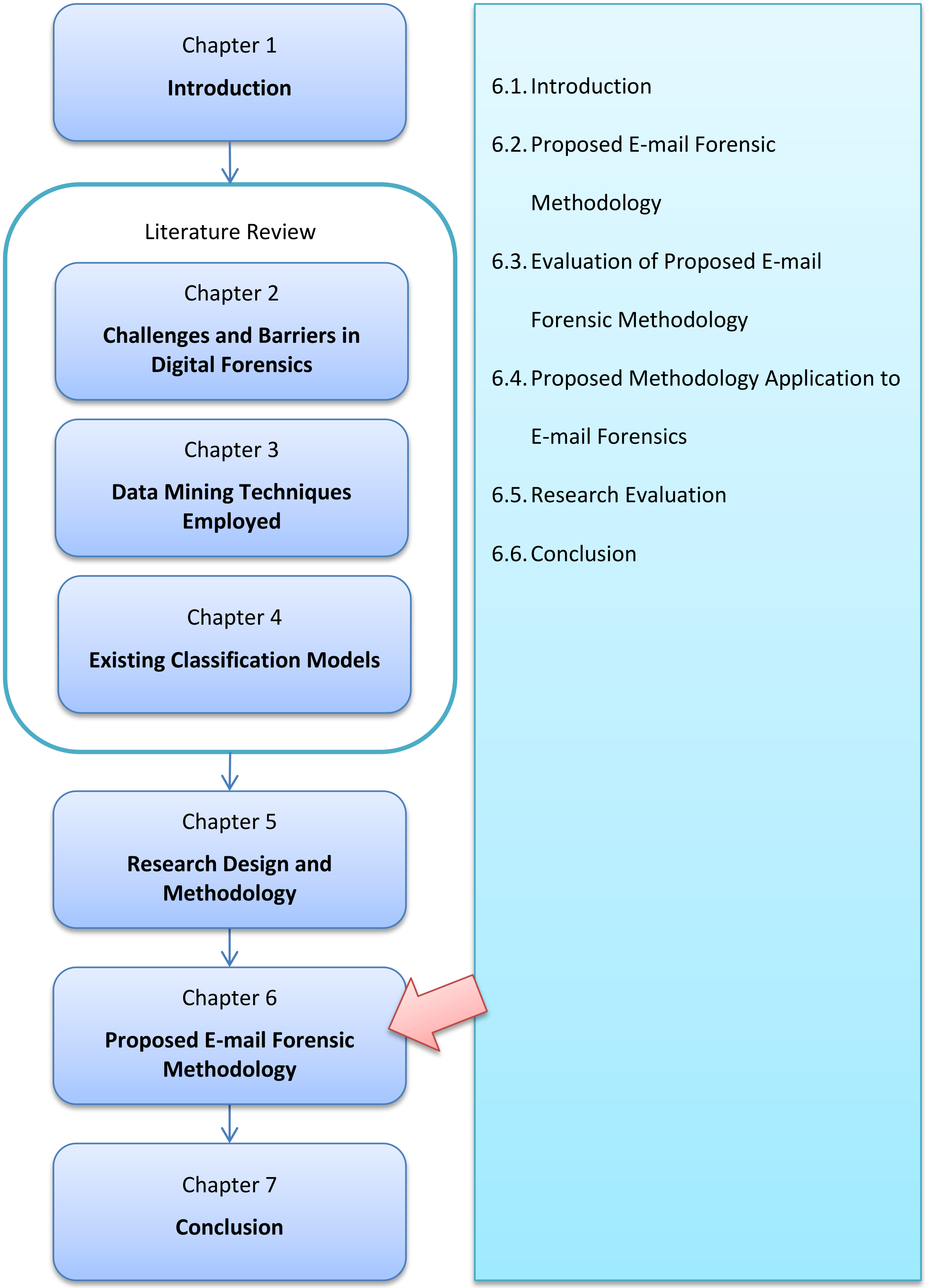
The different research paradigms were discussed in-depth in order to determine the best approach that the study should adopt. The positivistic, interpretivistic and critical paradigms were examined and it was determined that the study closely follow the interpretive paradigm. A discussion on Design Science ensued and the various approaches within the concept were expounded. Hevner and March (2003) guidelines were examined and Peffers et al. (2006) Design Science research process model was discussed. It was determined that the study follow closely the guidelines but was more in line with the conceptual mode proposed by Peffers et al. (2006).

The research techniques were then discussed and justification was argued for the chosen techniques, that of the interviews. Interviews were the best technique to extract rich, in-depth knowledge from the experts in the field. The key aspect of secondary data collection was also explained and finally the data analysis was covered.

This Chapter covered the process of obtaining the information and the approach taken in proceeding with the study. The following Chapter will introduce the proposed digtal forensic e-mail methodology and a discussion thereof will follow.

6. PROPOSED E-MAIL FORENSIC METHODOLOGY

CHAPTER 6



Chapter 1

Introduction

Literature Review

Chapter 2

Challenges and Barriers in Digital Forensics

Chapter 3

Data Mining Techniques Employed

Chapter 4

Existing Classification Models

Chapter 5

Research Design and Methodology

Chapter 6

Proposed E-mail Forensic Methodology

Chapter 7

Conclusion

6.1. Introduction

6.2. Proposed E-mail Forensic Methodology

6.3. Evaluation of Proposed E-mail Forensic Methodology

6.4. Proposed Methodology Application to E-mail Forensics

6.5. Research Evaluation

6.6. Conclusion

6.1 Introduction

The previous Chapter discussed the research design and methodology. This study followed the Interpretivistic research paradigm, as with most Information Systems research, and emphasis was placed on the Delphi technique, in which research findings are criticised and feedback is given through an iterative cycle. Design Science was chosen as part of the research methodology. Design Science allows for an artefact to be generated through an intensive research process. This Chapter presents a new more comprehensive E-Mail Forensic Methodology for digital forensic investigations.

The overarching Proposed E-Mail Forensic Methodology is a combination of the previous models and the processes followed during a digital forensic investigation. The aim of the Proposed E-Mail Forensic Methodology is to provide a comprehensive process flow to follow and to further standardise the digital forensic investigative process. The Proposed E-Mail Forensic Methodology is derived from models that have been proposed in other countries; furthermore the predominant practices have their origins in the USA which is at the forefront of the digital forensic discipline.

This research aims to maintain a South African context by addressing the South African laws that regulate the digital forensic profession; moreover South African process models have already been examined in Chapter four. This Proposed E-Mail Forensic Methodology is a generalised investigative process that can be applied irrespective of the technologies utilised and the evidence to be recovered. In the following section the Proposed E-Mail Forensic Methodology is introduced and explained.

6.2 Proposed E-Mail Forensic Methodology

In this section the Proposed E-Mail Forensic Methodology is presented, Figure 6 - 1, followed by a discussion. The Proposed E-Mail Forensic Methodology is a high level view of the entire investigation process. The key elements and processes adopt the nomenclature of previously examined models. The tasks of the Proposed E-Mail Forensic Methodology accurately represent those physical processes that form part of an investigation. The Proposed E-Mail Forensic Methodology aims to provide a base for all e-mail investigations that involve digital evidence and therefore this Proposed E-Mail Forensic Methodology can be expanded upon and refined.

The following paragraphs provide a full description of the Proposed E-Mail Forensic Methodology in sequential order. The original digital forensic model proposed by Carrier and Spafford (2003) is the consolidation of all preceding models and is the basis for all subsequent models. The reason for this is that while the actual phases of a digital forensic investigation are dependent on specific charateristics of an incident, the general nature of the model allows for it to be applicable to many types of investigations (Walters and Petroni, 2007).

The first process in the Proposed E-Mail Forensic Methodology is the Preparation Process. This process contains two elements namely Awareness and Readiness. The organisation must be aware (Ciardhuain, 2004) of the need for an investigation and ensure that the operations and infrastructure can sustain an investigation (Baryamureeba and Tushabe, 2004). In order for an organisation to be aware of the need, there must be a trigger, and this will come from an event/s that has compromised data/information and has violated the law (Stephenson, 2002). The awareness of a need for an investigation is derived from the receiver of the e-mail. The aim of this process is to allow the organisation to prepare for a forensic investigation larger than that of an investigation dealing primarily with evidence recovery. If there is no awareness, then no investigation can take place.

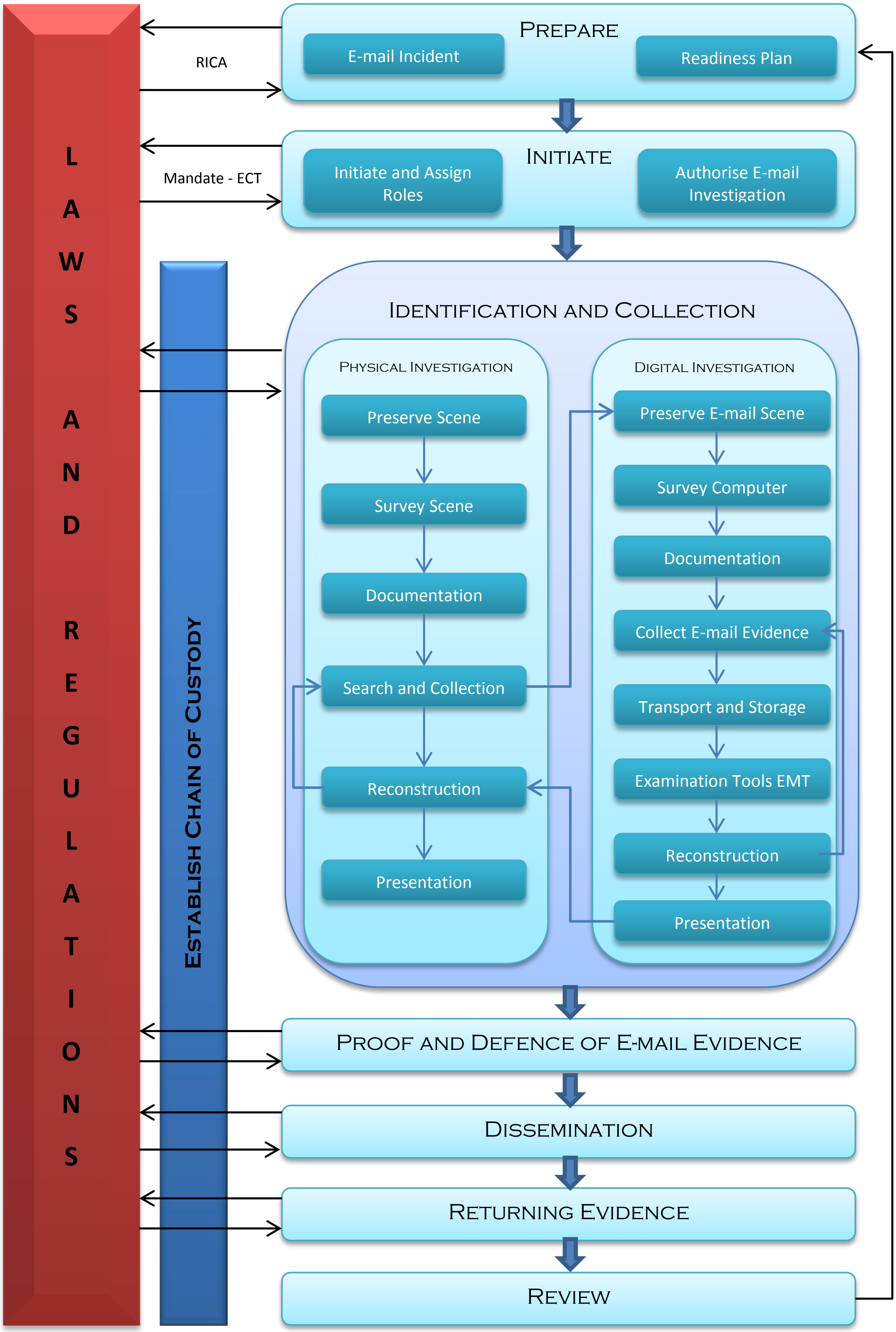


Figure 6 - 1 Proposed E-Mail Forensic Methodology

The impact the investigation may have on the business processes must be established. Therefore the readiness of the organisation must be taken into account. This is a crucial step; it is not reasonable or even feasible for an investigation to cost more than the perceived benefit of the outcome i.e. it is pointless shutting down operations in order to determine who sent an e-mail that is not threatening or dangerous in nature. However if an employee’s computer needs to be used in a digital forensic investigation and that employee may be rendered unproductive because no work can be performed without it, then the company may foresee the need to have a backup option available e.g. spare computer or roaming laptop. Thus, this is a risk assessment task performed at business level (Rowlingson, 2004). Drawing from Rowlingson (2004), the goals of a forensic readiness plan should be to:

- Recover legally admissible evidence without interrupting business processes
- Keep the cost of the investigation proportionate to the incident
- Ensure that the evidence makes a positive impact on the outcome of legal action

It is necessary to make these goals a priority. The organisation should be able to practically implement the plan which should contain activities that are clearly defined (Rowlingson, 2004). This initial step should be explicit in a forensic methodology because it defines the relationship with the events clearly as it impacts other steps and therefore ensures that the correct approach to the investigation is taken (Ciardhuain, 2004). The third goal of the forensic readiness plan is to have an impact on legal action, hence the plan must consider and establish a relationship with the law; this must be incorporated within the organisations security policy (Jordaan, 2010).

From the digital forensic investigator’s perspective, this first step corresponds with the preparation of tools needed to perform certain tasks during the investigation if the investigator is “in house” (Carrier and Spafford, 2003). As discussed in Chapter three, the digital forensic investigator would have a ‘toolbox’ that contains both tools and techniques in order to ensure retrieval of forensic evidence. From a readiness point of view, all digital forensic investigators have some form of readiness plan in place or follow a readiness process; therefore this is a practical step that is already implemented (Jordaan, 2010). The readiness plan would also be included in the company policy e.g. when an employee signs employment contracts they are required to sign a security and confidentiality agreement, and also a condition of terms of use i.e. use of company resources e.g. use of personal and company e-mail. If an employee breaches the terms, the company is in a position to take action. Hence the check (arrows in Figure 6 - 1) from the preparation process to the law i.e. in terms of RICA.

Once the incident has been identified, it is crucial for the organisation to initiate an investigation as soon as possible to minimise its impact and the damage or loss incurred. From a South African perspective this process of initiation is provided in the ECT Act where the mandate can be derived (Jordaan, 2010). Therefore without the mandate the evidence is useless. At this point the roles and responsibilities of the forensic team need to be established. Hence using the FORZA Framework, the eight roles can be specifically listed. The FORZA Framework was discussed in-depth in Chapter four and the roles and responsibilities can be found in Table 4 - 3. Each role can be carried out by one individual or each individual can assume more than one role, but what is important is that responsibilities for these roles are carried out accordingly.

Further these roles are easily identifiable; however, accountability is established once the investigation begins. In order for the investigation to begin, the necessary authorisation must be obtained from key individuals within the organisation i.e. the system administrator may only require verbal authorisation from management whereas law enforcement will need legal authorisation such as warrants (Ciardhuain, 2004). Furthermore, it is crucial that the authorisation is obtained in order to acquire the necessary data and in a South African context the applicable laws i.e. ECT Act and RICA must be adhered to but individual privacy and constitutional rights must be respected or the data becomes useless and inadmissible in a court of law (Jordaan, 2010). The organisation may also be required to notify individuals and other parties involved in the investigation but this will depend on the scope of the investigation e.g. phone call to local authorities to report that a crime has occurred or notification to an employee that their equipment is subject to an investigation. It is worth emphasising that the type of crime and the type of evidence that should be recovered also affects the initiation process (Jordaan, 2010). For example, some organisations monitor their employee e-mails, and any form of unbecoming activity that is inferred can be determined if there are key words, or phrases in the e-mails and hence the organisation can take the necessary disciplinary steps in house. However if there is correspondence from a web based e-mail account i.e. Gmail/Google mail etc., a search warrant may be necessary to obtain certain e-mail records and the police and digital forensic investigators must be actively involved.

The next process is to Identify and Collect Evidence. There are two sub processes that occur concurrently: the physical crime scene and the digital crime scene investigations. The reason that this distinction is made is because of the type of evidence recovered during the investigation i.e. physical evidence (physical objects) or digital evidence (digital data). Carrier and Spafford (2003) argue that "*an object has information about the incident because it was a cause or effect in an event related to the incident*" and therefore separates the investigation process into two aspects, i.e. physical and digital, because each has elements relating to the incident that initiated the investigation. At this point the digital forensic investigator should establish a chain of custody.

Hence the Physical crime scene investigation process deals with the physical aspect of the crime such as evidence that could have been left behind by the suspect relating to the incident identified e.g. flash drives. The Physical crime scene is classified into primary and secondary scenes; the primary scene is where the first criminal act occurred and the secondary scene is where additional criminal activities were carried out (Lee, Palmbach and Miller, 2001); there are however other methods of classifying a crime scene although this will not be discussed. The goal is to provide a link between the suspect and the incident; the following steps have been identified by Carrier and Spafford (2003) and are conducted by the law enforcement crime scene expert. This process contains six phases as outlined below.

The **Preservation Phase** is the securing of the incident site, which entails closing of exits and restricting access to the scene. This follows standard crime scene protocol in order to preserve the crime scene. This phase is indifferent to the type of crime, i.e. digital incident or non-digital incident, and activities include securing exits and identifying witnesses while the main aim of this phase is preserve the crime scene so evidence can later be identified and collected (Carrier and Spafford, 2003).

The **Survey Phase** occurs when the first responder identifies key pieces of physical evidence that contribute to the hypothesis of what crime has occurred. The objective of this phase is to identify obvious pieces of evidence in order to establish an initial hypothesis of the crime e.g. in a breaking and entering case, identifying how the suspect entered the building must be established and key pieces of fragile evidence must be collected immediately (Carrier and Spafford, 2003). The discovery of physical evidence at a location can link the suspect to a specific criminal act e.g. in the World Trade Centre bombing of 1985 the remains of a truck chassis with the vehicle identification number (VIN) intact was found; the truck was traced to an individual who rented the truck and later reported it as missing (Lee, Palmbach and Miller, 2001). Lee et al. (2001) demonstrate that crucial pieces of evidence can be linked to a suspect utilising the Linkage Theory as illustrated by Figure 6 - 2. In a digital incident examples of key pieces of evidence could be Personal Digital Assistants (PDA), cell phones, Compact Disk Read Only Memory (CD ROM), removable media (flask drives and external hard drives disks) (Carrier and Spafford, 2003).

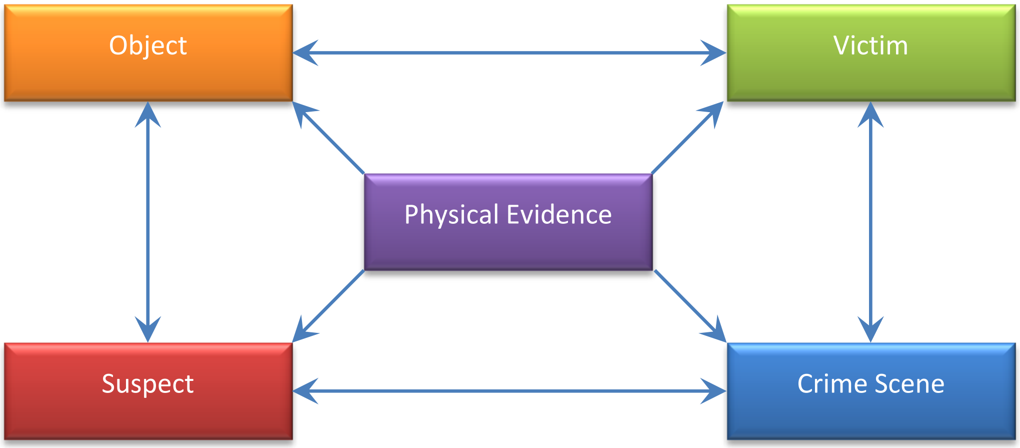


Figure 6 - 2 Linkage Theory (Lee, Palmbach and Miller, 2001)

The **Documentation Phase** objective is to collect as much information thereby preserving aspects of the crime scene; tasks include photographs and documentation of the crime scene (Carrier and Spafford, 2003). In a digital incident the connections to the computer must be photographed and documented and more importantly, the state of the computer must be documented. Moreover anything that could be useful in the later stages of the investigation must be recorded accordingly and Carrier and Spafford (2003) highlight that the final incident report is not generated during this phase.

The **Search and Collection Phase** is an in-depth analysis of the physical crime scene. This includes targeting logs of access to the crime scene and evidence such as the computer (Carrier and Spafford, 2003). Each type of evidence has a specific type of procedure on how it should be collected and this is the last phase that occurs at the physical crime scene. At this point the digital crime scene investigation begins and for a digital incident the computer will be collected as it is considered physical evidence, i.e. the arrow from the Search and Collection phase of the physical investigation to the Preservation phase of the digital forensic investigation; however the collection procedures must reflect how the volatile data was collected from the running system and how the computer was powered down.

The **Reconstruction Phase** is the process of analysing all the collected evidence and formulating a theory as to the events that transpired that led to the incident in question. A scientific method is used with the evidence to test the theory of the events. The results from the digital investigation are fed into the physical crime scene investigation at this point, i.e. the arrow linking the Presentation phase of the digital forensic investigation to the Reconstruction phase and this is where the link between the suspect, and the crime scene is made (Carrier and Spafford, 2003). However, if there is evidence missing the Search phase, will be resumed, the arrow linking back to the Search phase and the phases will be followed again.

Carrier and Spafford (2003) note that this phase is not the same as the recreation of the crime scene; this is about developing a model of the actual crime scene. An example of this phase is the creation of a event timeline whereby all evidence collected is matched to events and the suspect e.g. data entry logs of a individual are linked to the computer that was used in the digital incident. The **Presentation Phase** is the final part of the physical investigation as it involves presentation of the evidence to corporate management or a court of law, and this depends on the owner of the system who initiated the investigation (Ciardhuain, 2004). The theory and supporting evidence is presented to the relevant people (Carrier and Spafford, 2003).

The digital crime scene begins during the Search and Collection Phase of the physical crime scene and the results fed back into the physical crime scene investigation at the point of the Reconstruction Phase. The digital crime scene is defined by Carrier and Spafford (2004) as the virtual environment created by hardware and software where digital evidence about a crime or incident exists. The goal of the process is to identify electronic events on the system. Another reason why the distinction between the physical and digital crime scene is made is because the digital crime scene investigation may require specialists who have the expertise in computer forensics to perform the necessary tasks. This digital crime scene follows the same six processes of the physical crime scene but is tailored to the digital crime scene as the computer is treated as the crime scene and therefore it is searched for evidence.

According to Carrier and Spafford (2003) each digital device is treated as a separate crime scene as this enables analysis on different devices to be performed at different locations. During the physical crime scene investigation a distinction is made between primary and secondary crime scenes as suggested by Lee et al. (2001); this is also applicable to digital crime scenes for example if a server is hacked; this would form the primary crime scene and the log server that was hacked in order to modify audit logs would be classified as the secondary crime scene (Carrier and Spafford, 2003). The phases are discussed below; however two additional phases have been added according to the previous models that have been examined.

The Preservation Phase is the securing of the incident site, which includes closing off the computer from the network, and maintaining the integrity of log files in the system (Carrier and Spafford, 2003). Carrier and Spafford (2003) list certain tasks that are performed in this phase such as isolating the computer form the network, collecting volatile data before the system is turned off and identifying suspicious processes that are running to name a few.

Log files are crucial as they can be considered as eye witnesses to the crime and therefore they should be secured and back up copies made, and Carrier and Spafford (2003) state that the entire digital enviroment is preserved in this phase not just he digital evidence. Here, Carrier and Spafford (2003) name an advantage of the digital world, a complete image of the system can be made and be sent to the lab for forensic analysis and Carrier and Spafford (2003) draw the comparison to the physical crime scene where photographs are taken in order to replicate the crime scene when analysis is performed at a later stage.

During the Survey Phase two types of digital evidence are identified: live data and static data because there are differences in evidence recovery as well as the impact of the law on the specific type of data (Grobler and Von Solms, 2009b). A live data survey is conducted together with the capturing of images of the digital system.

A static data survey occurs when the digital forensic investigator creates an image of the system using software; sometimes more back-up copies of the image are made in order to safeguard against accidental loss (Carrier and Spafford, 2004). A live data survey is carried out when for example, a critical server cannot be turned off and then the investigator would collect data by running data collection tools and then save the output to a CD. The goal of a live data survey is to keep the system running and to maximise the data collected without making changes to the system (Grobler and Von Solms, 2009b).

The Documentation Phase is not a specific phase and is performed when the evidence is found but a chain of custody needs to be established early on in the investigation for the evidence to be used in a court of law (Grobler and Von Solms, 2009b; Ieong, 2006). The photographs and images taken of the system during the preservation/survey phase, form part of the documentation process. Carrier and Spafford (2004) state that digital evidence exists in abstraction layers and this must be documented accordingly; some examples include documenting the file with its full path name and sectors used on the hard disk whereas network data can be documented with the source and target addresses at various network layers.

The Search and Collection Phase includes copying digital evidence and making use of technical and non-technical investigators on hand; specific tasks must be performed on the data using standardised and accepted procedures (Carrier and Spafford, 2003; Jordaan, 2010). This phase uses the results from the survey phase and focuses on additional analysis types e.g. a keyword search can be performed after keywords are identified from other evidence. During this phase deleted files are recovered and a timeline of the user's activity is established and different search techniques are utilised in this phase just as there are different techniques during the physical crime scene investigation (Carrier and Spafford, 2003).

Once all necessary steps have been executed at the crime scene and the evidence has been collected, the next step is the Transport and Storage of evidence. Evidence is transported to a safe place where further analysis is performed. During this step the integrity of the evidence must be ensured in order for the evidence to remain valid; transport to a secure location where the evidence is stored reduces the risk of evidence tampering and thus the integrity of the evidence is maintained (Baryamureeba and Tushabe, 2004). It is not always necessary to transport the physical evidence as stated earlier during the survey phase; the example of the critical server that cannot be removed is a case in point. However the data recovered from the server must maintain its integrity and this data is normally transported to a lab for further analysis (Jordaan, 2010).

The Examination Phase is an in-depth analysis of the digital evidence and involves the application of digital forensic tools and techniques used to gather evidence and to further scrutinise the evidence in order to support or reject a hypothesis. As elaborated in Chapter two, the terabyte hard drives have become the norm in the industry and hence the amount of data recovered would be a large volume. Jordaan (2010) gives an example where one page with one letter depressed until the page is full could easily contain 250 plus pages of data. As a result of the large volumes of data the digital forensic investigators require automated techniques to examine it (Baryamureeba and Tushabe, 2004; Jordaan, 2010). This phase is the most intense phase as the majority of investigative time is spent in this phase (Baryamureeba and Tushabe, 2004).

The Reconstruction Phase uses scientific methods of testing and rejecting theories based on the digital evidence; however, if information is missing the Search phase will commence again, the arrow linking back to the Search phase of the digital forensic investigation. During this phase all the evidence is mapped together in order to provide a cohesive set of facts that support/reject the theory or theories (Carrier and Spafford, 2003). The data will point to certain events and these events may or may not provide supporting evidence to the incident in question (Carrier and Spafford, 2004). The hypotheses are tested and the data is scrutinised and tested in order to evaluate if the event is supported by the data and that the data in fact shows that the event occurred (Carrier and Spafford, 2004).

The last step is the Presentation Phase where the digital evidence is presented to the physical crime scene investigators. The results of the digital crime scene are used in the reconstruction of the crime during the physical crime scene (Carrier and Spafford, 2003). Carrier and Spafford (2003) also note that the digital crime scene investigation of the system does not involve data from other digital sources. Recall that each digital source is treated as its own crime scene. Therefore this phase presents the findings of each digital source to the physical crime scene investigators and Carrier and Spafford (2003) add that the physical and digital crime scene investigators may be the same team and this is in line with the FORZA framework roles identified earlier. Additionally Jordaan (2010) states that sometimes specialists are required during an investigation and their expertise is utilised during the specific phase where evidence is recovered i.e. the search and collection phase.

It is important to further emphasise the impact the law has on the investigation process. During the Identification and Collection phase any evidence collected must be done according to the relevant laws and legislation in place in the presiding jurisdiction. For example once the Identification and Collection phase is complete a concrete hypothesis is usually established and in most cases the suspect is clearly identified by the data and information of this phase (Jordaan, 2010). Jordaan (2010) further comments that at this point most suspects will either plead guilty to the charges or accept a deal due to the incriminating evidence and as a result, the case does not go to court.

The next process is Proof and Defence. Opposing theories will also be presented and therefore there is a need to provide substantiated proof of the events that occurred as well as defence of the theory of the events that occurred (Ciardhuain, 2004). This is where the benefits of having a standardised digital forensic process enables, either the conviction of a suspect or the exoneration of an innocent individual. According to Ciardhuain (2004) the presented hypothesis will not go unchallenged and a contrary hypothesis as well as supporting evidence will be presented to the court and here the digital forensic investigators will have to prove the validity of their hypothesis. A successful challenge will result in the revaluation of the evidence in order to contruct a better hypothesis; each hypothesis will be supported by the evidence presented and be based on existing laws and legislation and in rare occasions set a new precedent.

The following process of Dissemination involves the sharing of information in order to provide a basis for future investigations i.e. court precedents. Therefore the information will influence future investigations (Ciardhuain, 2004). Conversely, there are various policies and procedures that need to be followed as prescribed by the organisation and the law in order to share information relating to a crime. Certain information may only be made available in the investigating organisation while other information may be more widespread (Ciardhuain, 2004). Hence collecting and maintaining such information is a key aspect for digital forensic investigators and Hauck, Atabakhsh, Ongvasith, Gupta and Chen (2002 in Ciardhuain, 2004) provide an example of this; it is a system called Coplink which provides real-time support for law enforcement.

Coplink is essentially an analysis tool based on a large collection of information from other investigators. Thus investigators are more effective if they have information and Coplink takes advantage of this by capturing connections between people, places, events and vehicles based on past crimes (Hauck, Atabakhsh, Ongvasith, Gupta and Chen, 2002). Therefore the facilitation of information dissemination is performed through the normal channels of media but also through specific tools generated for this purpose and also to facilitate public opinion (Roydhouse, 1999).

The rule of sub judice applies to information where the court cases are on-going. It is also crucial to understand that information about sub judice cases cannot be disseminated as this would have an impact on the case and hence, information can only be disseminated in accordance with the rule of sub judice (Roydhouse, 1999). Roydhouse (1999) comments further that it would not only be impossible to regulate the Internet's communication technologies but also to agree on the regulation in the first place. Therefore careful consideration must be paid when information is disseminated as these are subject to controls which must be adhered to.

The process of Returning Evidence allows the investigation to come full circle in terms of addressing the physical and digital evidence removed for analysis. The evidence must be returned to the proper owners and the criminal evidence removed (Reith, et al., 2002). Once the court case has been concluded it is necessary to return all physical evidence that was removed from the crime scene.

The final process is the Review Phase and is performed after the investigation to determine how effective certain processes and techniques were and whether the digital and physical investigators worked well together (Baryamureeba and Tushabe, 2004). This phase is used to identify areas of improvement and to refine the processes. This is the linkage from the

Review phase to the Preparation phase. The objective is to focus on poor practices and errors encountered during the evidence recovery process. As with other disciplines this is a continuous improvement item that is constantly changing.

Due to the nature of the digital forensic process and the applicable laws, mistakes can be costly and therefore criminals may not be prosecuted. Hence it is of the utmost importance that due diligence is applied during the investigative process and the chain of custody is maintained; the review process allows for further improvements thereby creating a better process of digital forensic examination (Baryamureeba and Tushabe, 2004). Baryamureeba and Tushabe (2004) name the elements that are examined in this process as data protection, data acquisition, imaging, extraction, interrogation, ingestion/normalisation, analysis and reporting. Therefore the entire investigative process must be examined for improvements. In the Proposed E-Mail Forensic Methodology the emphasis lies on the laws and legislation and therefore every aspect of the digital forensic framework must be scrutinised for faults and improvements. Now that the entire Proposed E-Mail Forensic Methodology has been discussed it is necessary to evaluate the whole process. The following section focuses on the evaluation of the Proposed E-Mail Forensic Methodology.

6.3 Evaluation of the Proposed E-Mail Forensic Methodology

As with many models and methodologies that are proposed, the evaluation of the benefits and value must be demonstrated. This section highlights the uniqueness and the added value the Proposed E-Mail Forensic Methodology adds to the digital forensic discipline, specifically e-mail forensics. The Proposed E-Mail Forensic Methodology, Figure 6 - 1, is a hybrid of previously proposed and well tested models utilised in the digital forensic discipline. Hence a comparison of these models is not beneficial because although this methodology draws from those it is a more comprehensive process and as stated previously, the Ciardhuain (2004) model is deemed to be the most current and therefore this methodology is based on it. It is worth noting that other models define similar processes; however they are named differently. Chapter four covers the various models highlighting their differences and similarities.

The uniqueness of this Proposed E-Mail Forensic Methodology is that it demonstrates all the processes of the digital forensic investigation from start to finish and the emphasis is on the law and legislation as the crux of an investigation is to provide legally admissible evidence in a court of law. Previous models did not highlight the importance of this aspect which was taken to be an implicit aspect of an investigation. However to effectively illustrate the digital forensic process the law must be explicitly stated and recognised. Additionally all digital forensic investigators are aware that a chain of custody must be maintained during an investigation; however there is little mention in previous works about maintaining a thorough chain of custody. Again this is a legal aspect of the process as without a chain of custody the evidence is useless.

6.3.1 Advantages and Disadvantages

I. Advantages

The main advantage is the definition of the processes within the digital forensic investigation in totality, thereby allowing for better prosecution in a court of law. Inclusion of the Integrated Digital Forensic Process Model has allowed for greater definition of the forensic investigation. The Proposed E-Mail Forensic Methodology has included the collection of 'live data' during the digital forensic investigation and therefore has allowed for a greater spectrum of evidence recovery.

An additional advantage is that this Proposed E-Mail Forensic Methodology explicitly names the law and legislation as a presiding aspect. This is important as other models recognise the law as an implicit aspect. There are also various processes within the Proposed E-Mail Forensic Methodology that are iterative and various researchers have recognised these aspects but these were not illustrated accordingly. Additionally the Proposed E-Mail Forensic Methodology is proactive in design by incorporating the preparation and awareness processes so that the forensic process is not merely reactive. This Proposed E-Mail Forensic Methodology will provide a further baseline for researchers to establish standardised digital forensic process as it can be abstracted from the e-mail aspect of the investigation.

II. Disadvantages

As with many of the digital forensic methodology, the Proposed E-Mail Forensic Methodology is an abstract one that is standardised to cater for the broader spectrum of investigations. This does not detract from the value that will be gained using the Proposed E-Mail Forensic Methodology as covering the steps and processes during an investigation will improve the reliability and admissibility of evidence. The inclusion of live data in the evidence recovery process creates the problem of inadmissibility of evidence as the courts have yet to establish precedent.

Furthermore the law is stated as an explicit aspect yet there are no guidelines that point an investigator from another country to the presiding laws and legislation. Moreover the Proposed E-Mail Forensic Methodology is a very high level view of the forensic process; it does not give detailed guidelines as to specific tasks that can be executed within a process.

6.4 Proposed Methodology Application to E-Mail Forensics

In order to demonstrate the effectiveness of the Proposed E-Mail Forensic Methodology, the additional objective of the study needs to be satisfied i.e. how the Proposed E-Mail Forensic Methodology can assist during an investigation and hence how it will be applied to a hypothetical scenario. Since the Proposed E-Mail Forensic Methodology is a high level depiction of the digital forensic investigation the focus will be on an investigation whose ‘smoking gun’ is the e-mail that has been sent.

Using a hypothetical scenario, based on actual events, of a company whose directors have been sent a threatening and anonymous e-mail one can gauge the novelty that the Proposed E-Mail Forensic Methodology contributes. While this is a hypothetical scenario the events are taken from an actual investigation that cannot be named in this forum. If an employee is suspected of sending the e-mail, due to the content of the e-mail i.e. only knowledge certain people would have, the first step is to follow the readiness plan since the company will be aware of the need to investigate. The awareness would arrive from the consequent reading of the e-mail by the receiver i.e. the director.

At this point the director should alert either the authorities or inform the other directors of the company if the matter is to be handled internally. The second sub process of the Preparation Process is the Readiness Plan or Process and this is demonstrated by both the organisation's policies and procedures and where available the organisation's Readiness Plan. The Readiness Plan should contain attributes of the organisation that allow the reaction to such an event. If the organisation does not have the capability to handle the investigation in-house i.e. the system administrator is not familiar with the processes that must be followed during an investigation, then the organisation will have to seek assistance. The assistance could be in the form of a digital forensic investigator or another company that specialises in such investigations. At this point the investigation is initiated. For the purpose of this scenario, the organisation will utilise the assistance of a digital forensic investigator. The Preparation phase is applied to the digital forensic investigator who would have the necessary knowledge as well as the tools needed to recover the evidence. Therefore the Initiation Process will involve the hiring of a digital forensic investigator. At this point all the roles of the investigation can be established except for the legal prosecutor as the case has not yet proceeded past the Identification and Collection phase; the roles are displayed in Table 6 - 1.

Table 6 - 1 Identified Roles from given scenario

Person	Roles
Director who received the E-mail/organisation	• System/business owner
System Administrator of the organisation	• Security/system architect/auditor
Hired digital forensic investigator	• Case leader • Digital forensics specialist • Digital forensics investigator/system administrator/operator • Digital forensics analyst

The process of Initiation also involves the Authorisation sub-process. This will take the form of written or verbal communication to the digital forensic investigator. Since the organisation owns the equipment on which the director received the e-mail, authorisation is implicit when the digital forensic investigator is hired and tasked with the investigation. The investigator would need to liaise with the system administrator in order to obtain access to the e-mail server of the organisation. If the organisation had reported the incident to the authorities then the law enforcement agencies tasked with the investigation may require a search warrant to legalise the investigation.

The next step in the investigation is the Identification and Collection of evidence. Due to the nature of the e-mail, it emerged that only a few employees would have the knowledge and information in order to author such an e-mail. With this information the digital forensic investigator can cycle through the physical crime scene activities. The suspected employees' computers can be confiscated by the organisation; however, there must be evidence to suggest that an employee has the knowledge contained in the e-mail before any equipment can be confiscated. All the evidence encountered during the Search phase must be documented accordingly and the digital forensic investigator would start a chain of custody so that each piece of evidence is handled with a proper ownership log. This chain of custody must be maintained throughout the investigation.

During the Examination Phase, the digital forensic investigator will need to follow further steps specific to e-mails i.e. determine the author of the e-mail; extract evidence supporting the conclusion on authorship. Firstly, the digital forensic investigator traces the origin of the e-mail though network forensic processes by performing hash functions and looking for e-mail header data. The digital forensic investigator will identify evidence, such as the actual e-mail message that was sent, Word documents, text files, etc. that exist both on the director's machine and any of the employees' machines.

Using techniques such as data mining as discussed in Chapter three and date time stamps, the investigator can determine from the list of suspects, the computer that was used and this can be seized for further analysis. To further add to the digital forensic investigator’s case, the application of techniques such as write-prints and stylometry can be applied to the e-mail and thereby aid in the determination of the author of the e-mail.

The digital forensic investigator can utilise the E-mail mining toolkit (EMT), which is a tool, as those discussed in Chapter three, used to perform e-mail mining were discussed (Stolfo and Hershkop, 2005). EMT is an automated tool designed to analyse large sources in minutes and present to the digital forensic investigator and analysts e-mails ranked and logically ordered for their direct inspection (Stolfo and Hershkop, 2005). If any evidence is removed from the crime scene it must be part of the chain of custody and be secured so that Transport and Storage of the evidence does not affect the integrity.

The link between the suspect and the evidence can be made by matching the employee’s writing style to that of the e-mail in question. During the Identification and Collection Process it is essential that the employees’ rights to privacy are respected in terms of the privacy legislation; additionally the RICA prevents organisations and law enforcement agencies from intercepting communications under certain circumstances and therefore if the organisation in question obtained a lead to a suspect via their internal monitoring system without reasonable grounds, the evidence may be deemed inadmissible. Hence the company cannot use any information obtained by e-mail message that has been monitored especially if the receiver has not yet received or even read it because according to the RICA act the message is considered in transit and therefore the information has been illegally obtained (Jordaan, 2010).

If the link made between the suspect and the evidence is conclusive then the digital forensic investigator would present the findings to the person or the panel of the organisation that initiated the investigation. The digital forensic investigator would need to prove that the evidence supports the events leading up to the incident as well as show that all other contrary theories can be dispelled. In this case the organisation would take action against the employee or else lay a criminal charge and take further action. The evidence would be utilised to expose the criminal act as well as the suspect who performed the criminal act. The digital forensic investigator may be called to substantiate the evidence in a court of law.

The process of Dissemination would occur differently in these two instances. If the organisation took action internally the information would not be released and shared, although it could be used in a paper that could be published. However if the case went to court the organisation would only be sharing the information with the relevant authorities and then the information would be considered sensitive and since the case would be sub judice, the information would not be released to the public (Jordaan, 2010). Again the dissemination of information is subject to controls such as legislation that governs such information distribution as well as the policies that exist within the organisation. The returning of evidence would occur if the investigator removed any physical items of evidence from the organisation or if the matter went to court; the evidence would be held until the court case was completed.

The review process occurs from the investigator's side where any elements that required improvement would be noted and carried forward to the new investigation. However the organisation itself could also review the processes it has in place; the readiness plans could be revised or implemented if they did not previously exist.

6.5 Research Evaluation

The approach adopted for the validation process was the Delphi technique through the use of expert review. The Proposed E-Mail Forensic Methodology was presented initially to a group of experts who provided critical review. These experts were of international stature. The initial feedback was favourable as reviewers commented on the South African perspective applied in the methodology context. The following is an account of the feedback, both positive and negative, weighed against the Proposed E-Mail Forensic Methodology during the refinement stages.

6.5.1 First Round of Expert Reviews

Reviewer number one commented that the Proposed E-Mail Forensic Methodology was based on a strong literature review and contained key existing works in the domain of digital forensics. The reviewer also added that the novelty and value of the Proposed E-Mail Forensic Methodology must be emphasised as many models existed but none had become a de facto standard in the discipline. However the reviewer added that it was difficult to reason why this Proposed E-Mail Forensic Methodology would be favoured over more accepted models. Additionally, the reviewer added that some validation of the Proposed E-Mail Forensic Methodology would establish support.

Reviewer number two commented that the Proposed E-Mail Forensic Methodology was "*as usual for such models, pretty abstract and does not add much to existing forensic practice*". Therefore the reviewer rated this methodology as low in originality. The reviewer also stated that the specific Proposed E-Mail Forensic Methodology contained no real link between e-mail forensic and the general digital forensic investigation process. The reviewer did however rate the presentation of the Proposed E-Mail Forensic Methodology as 'good'. The reviewer also commented that a more technical aspect in the Proposed E-Mail Forensic Methodology would be beneficial as technically minded people would provide additional scrutiny.

Reviewer number three added that the Proposed E-Mail Forensic Methodology was about an interesting and important topic in digital forensics; however the overall Proposed E-Mail Forensic Methodology was about the generalisation and standardisation of the digital forensic process instead of e-mail forensics. Moreover the reviewer commented that the number of references used included important works in the digital forensic discipline.

6.5.2 Second Round of Expert Reviews

Reviewer number four found that the entire Proposed E-Mail Forensic Methodology would go some way to establishing a standardised digital forensic process. Additional comments were favourable to the extent that the standardisation aspect of the Proposed E-Mail Forensic

Methodology should be emphasised. Reviewer 4 believed that the model was an important step in the digital forensic discipline.

Reviewer number five concurred with reviewer number two and three in criticising the Proposed E-Mail Forensic Methodology's lack of addressing e-mail forensics. The technical aspect of the e-mail process should be incorporated; however the reviewer did note that the Proposed E-Mail Forensic Methodology would then not aid the standardisation process. Reviewer five suggested that the Proposed E-Mail Forensic Methodology be applied to a specific scenario in which the case revolved around e-mail forensics and e-mail authorship attribution. The reviewer also added that it is *"as complete as one can illustrate the process without becoming technical"*.

Reviewer six commented on the South African perspective that was taken in regard to the aspect of live forensic acquisition. Additionally the reviewer commented favourably on the law aspect that was incorporated and explicitly illustrated in the Proposed E-Mail Forensic Methodology and that the law was highly relevant to all digital forensic investigations. Reviewers seven, eight and nine had no additional comments above those already recorded by reviewers five and six. The only exception was that reviewer eight found the Proposed E-Mail Forensic Methodology interesting in that it could be abstracted and utilised for a generic digital forensic investigation.

6.5.3 Third Round of Expert Reviews

Reviewer ten (one of the foremost experts in digital forensics in South Africa and internationally recognised) added that a South African perspective must be addressed in the literature review. Additionally the reviewer was happy with the manner in which the forensic process was covered and that the Proposed E-Mail Forensic Methodology was comprehensive and added to the field by moving towards a standardised process. Moreover the reviewer was pleased with the inclusion of the readiness aspect of the investigative process.

The reviewer also commented strongly on the emphasis of the authority required to initiate an investigation. The reviewer also stated that the Proposed E-Mail Forensic Methodology captures the key aspect of the digital forensic process, that the process was accurately depicted; since the digital forensic discipline is a mix of all other disciplines, the inclusion of the law as well as behavioural sciences was appreciated in this context. Furthermore the reviewer added that the methodology was very close to the actual process practised and was happy that many of the processes were depicted as an iterative process.

Some of the comments provided by many reviewers concurred with the sentiments expressed by others. The reviewers' comments weighed heavily as all are either practitioners or researchers in the digital forensic discipline as well as considered as experts and therefore all comments and suggestions as well as criticisms were taken into account and incorporated in the study which led to the final Proposed E-Mail Forensic Methodology presented in this study. Each iteration of expert review was incorporated into the Proposed E-Mail Forensic Methodology and the advice given was used.

Based on the above evaluations, it can be concluded that the Proposed E-Mail Forensic Methodology provides a good basis for understanding the process of digital forensic investigations. Additionally the study set out to meet the objective of creating a standardised process for digital forensic investigators to follow and to apply to an e-mail forensic case. Based on the expert review the study has accomplished the objectives. The Proposed E-Mail Forensic Methodology was designed so that the investigation process could be followed irrespective of the type of investigation performed i.e. e-mail forensic analysis; this demonstrates the usefulness of a generic Proposed E-Mail Forensic Methodology as no current standard exists. Therefore this provides a way forward for other researchers and practitioners to further refine the Proposed E-Mail Forensic Methodology.

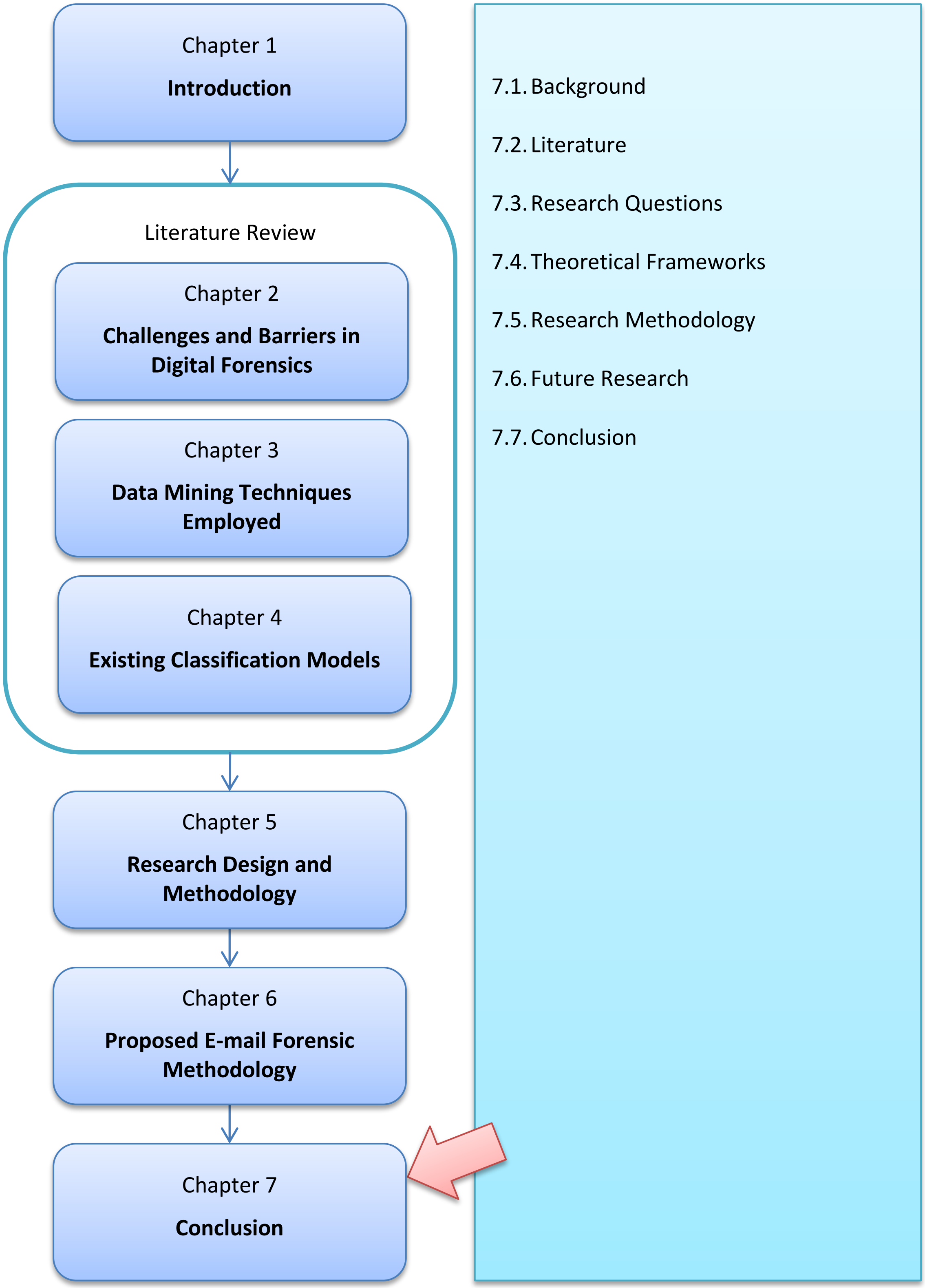
6.6 Conclusion

This Chapter presented the Proposed E-Mail Forensic Methodology. A brief introduction was given followed by an in-depth discussion about the model addressing each component of the Proposed E-Mail Forensic Methodology. The Proposed E-Mail Forensic Methodology was evaluated in terms of addressing the objectives of the study. The advantages and disadvantages of the Proposed E-Mail Forensic Methodology were discussed in order to show the uniqueness and value that it adds to the digital forensic domain.

The Proposed E-Mail Forensic Methodology's usefulness was demonstrated by the application of it to a hypothetical scenario based on actual events. The research findings were presented and the expert review and Delphi technique were discussed; the Proposed E-Mail Forensic Methodology was refined accordingly based on the feedback obtained. The feedback was mostly favourable in terms of the standardisation process taken to present the Proposed E-Mail Forensic Methodology; however criticism faulted the e-mail forensic aspect, as most reviewers expected the Proposed E-Mail Forensic Methodology to explicitly address the problem of e-mail forensics. This was addressed by the inclusion of the scenario in which the Proposed E-Mail Forensic Methodology was evaluated against. The following Chapter provides the conclusion to the study and gives a summary of all the Chapters.

7. CONCLUSION

CHAPTER 7



Chapter 1

Introduction

Literature Review

Chapter 2

Challenges and Barriers in Digital Forensics

Chapter 3

Data Mining Techniques Employed

Chapter 4

Existing Classification Models

Chapter 5

Research Design and Methodology

Chapter 6

Proposed E-mail Forensic Methodology

Chapter 7

Conclusion

7.1. Background

7.2. Literature

7.3. Research Questions

7.4. Theoretical Frameworks

7.5. Research Methodology

7.6. Future Research

7.7. Conclusion

7.1 Background

The previous Chapter presented the Proposed E-Mail Forensic Methodology and discussed the findings of the study. The Proposed E-Mail Forensic Methodology was developed using the secondary data and primary data formed part of the validation process in order to have expert review of the Proposed E-Mail Forensic Methodology. This Chapter will provide a summative conclusion of this research project. A background to the study in terms of the literature, research questions and methodology will be given and the contribution made will be presented followed by the research findings. The research objective is addressed by highlighting the progress made during this study. The limitations of this study are discussed and the recommendations for future research are provided. Finally concluding remarks are made pertaining to the study.

7.2 Literature

The literature examined in this study suggested that the challenges facing digital forensic investigators are ever changing and evolving. Various views were taken into account to establish some of these challenges and they were subsequently organised into categories. The key issues within a category were examined and the challenges were discussed. A comparison of the various categories was made and the similarities and dissimilarities were identified.

Through this discussion a new consolidated set of categories were established namely: Technological; Educational; Societal; Procedural and Legal. Additional challenges were identified and added to the categories in an attempt to update the challenges listed previously. All other categories pointed towards the legal category as all efforts during an investigation must fulfill the legality requirement.

E-mail Authorship attribution was identified as a problem for digital forensic investigators. While the original problem of authorship attribution has its roots in pure text, the rapid advances of technology have taken this problem into a new dimension i.e. identification of an author of an e-mail. It was also established that digital forensic investigators require specific tools in order to determine the author of an e-mail, and this is achieved through the process of data mining and utilising specific applications and with the aid of stylometry, determine the author of an e-mail.

There are various automated tools; Support Vector Machine has shown advantages over traditional techniques and when combined with other feature types, SVM outperforms other machine learning techniques. Additionally new techniques have been proposed; however they have not yet been employed as their results need to improve in order to aid credibility. Over and above the techniques that digital forensic investigators utilise, commercial tools are also available although the advances in the discipline are fast outpacing these tools making them obsolete.

One has to consider the entire digital forensic process as a flowing process and the tools and techniques used do not occur in isolation. During an investigation, digital forensic investigators follow a set of predefined tasks and activities and these are contained in digital forensic classification models. These models are in the form of frameworks, methodologies, and models and include predefined steps that a digital forensic investigator must adhere to during an investigation in order to obtain legally admissible evidence.

However, there are a number of models and following a predefined process has become more complicated. The number of models and the contributing disciplines makes it more difficult for investigators to adhere to a predefined process as investigators are required to be multi-skilled. A number of unique models were also examined and one specific framework called FORZA aids in creating a better process by identifying key roles and responsibilites and information flow. As the number of proposed models increased there was no standardisation of the processes and procedures to follow. It was established that there is a need to create a standard methodology which will focus on the entire investigation process and chain of custody. The following section addresses the research objectives and consequent sub-questions.

7.3 Research Questions

The main research objective is to produce an e-mail forensic methodology that will aid a digital forensic investigator in determining authorship of an e-mail and produce legally admissible evidence in a court of law in the process. In order to address this objective three sub-questions were posed.

1. What are the challenges faced by digital forensic investigators in conforming to the law with respect to presenting legally admissible evidence?
2. How can data mining techniques aid in the attribution of authorship of e-mails?
3. What are the classification models used and how do they assist in the verification of evidence?

By addressing the three sub-questions, the overall objective would have been addressed as the three sub-questions are derived from the research objective. The research sub-questions were addressed in the literature review. Research sub-question one was addressed by establishing the challenges within the digital forensic domain, thereby creating a consolidated categorisation based on previous work.

Research sub-question two was addressed by determining what e-mail forensics and authorship attribution is and then focusing on the data mining tools used during an investigation. It was demonstrated that utilising an automated machine learning classifier such as SVM increases the accuracy of the authorship attribution results. It was also established that many of the tools available are out-dated and there is a greater need as time passes for newer tools that address current problems already identified.

The classification models were examined in addressing research sub-question three. An in-depth scrutiny of the models was performed and a comparison was made in order to establish the gaps between the different models. From this it was demonstrated that a standardised process must be followed in order to produce legally admissible evidence and therefore classification models provide the basis for investigators. It was also highlighted that a standardised process would result in a stronger yield of evidence. Hence the standardisation of the process will improve the verification processes performed on the evidence. Once these questions were addressed, the result was a Proposed E-Mail Forensic Methodology that was produced through the Design Science methodology.

The main objective of this research project is to produce a methodology that will aid a digital forensic investigator in determining authorship of an e-mail and produce legally admissible evidence in a court of law in the process. This objective has been addressed through collectively addressing the research sub-questions. This was demonstrated through the application of the Proposed E-Mail Forensic Methodology to a hypothetical scenario. The adoption of this e-mail forensic methodology can be explained by the diffusion of innovations theory. This will be summarised in the section below.

7.4 Theoretical Frameworks

In order to establish how digital forensic investigators would react to a standardised forensic process methodology, the research utilised the Diffusion of Innovations theory. This theory postulates that the innovation adoption process is one of information gathering and uncertainty reduction. According to Rogers, Singhal, and Quinlan (2007) diffusion research is distinctive due to the communication messages that individuals perceive as "new". Therefore, this is the reason for the high uncertainty in information gathering that individual's experience.

The distribution of adopters of an innovation can be approximated by a normal distribution of the time of adoption. Therefore using the mean and standard deviation of this distribution, one obtains five adopter categories, namely: innovators, early adopters, early majority, late majority, and laggards. Thus with the call for standardisation in the field of digital forensics, investigators will fall into the different categories of adopters.

Since many of the processes in the Proposed E-Mail Forensic Methodology are derived from existing models and frameworks, the adoption process by digital forensic investigators would not be a difficult transition and therefore the digital forensic investigators adoption would fall into the early majority or late majority category. In addition creating a new process by which e-mail forensic investigations should be conducted reduces the high degree of uncertainty inherent in all e-mail forensic investigations as the process would be generally accepted as the de facto standard.

7.5 Research Methodology

The research methodology was discussed in Chapter five and the techniques used during research were examined, both qualitative and quantitative methods. However the research adopted a purely qualitative approach as expert review was attained in order to refine and validate and the Delphi technique was utilised in the development of the Proposed E-Mail Forensic Methodology. Additionally the research paradigms were discussed and it was determined that the study would be approached from an interpretivistic point of view and this led to the discussion on Design Science.

Design Science essentially allows an effective artefact to be produced through thorough and intensive research. This study creates an artefact derived from previous research and is further refined through expert review. A model for producing and presenting Design Science research was presented and examined. Justification was given as to why Design Science was the adopted approach and the reason for selecting expert review was defended.

This constituted the primary data collection and was facilitated through one on one interviews with the experts. Interviews were selected as the primary data source because in the digital forensic discipline, experts in the domain have implicit knowledge of the processes and techniques applied during an investigation. This type of knowledge is difficult to transfer and therefore a deeper understanding of the problem can be gauged from the interview process, and additionally a new perspective can be used to address the problem. The interviews were approached from a structured interview type; however because the knowledge expected from the experts, additional flexibility was provided for. Therefore a combination of structured (closed ended question) and unstructured (open ended questions) interviews were conducted.

The predominant approach was that of open ended questions as this provided the platform for further queries and information extraction.

7.5.1 Discussion

The Design Science approach was applied and this is reflected in the research methodology applied to the research project. Hence ten experts were presented with the study's findings and asked to reflect and comment on the findings as a step to further refining the methodology. This approach also corresponded to the Delphi technique which follows an iterative cycle of refinement. The Delphi technique is a widely used and accepted method for gathering data from respondents within their domain of expertise. The technique is designed as a group communication process which aims to achieve a convergence of opinion on a specific real-world issue. The Delphi technique applied to the study consisted of three rounds of correspondence.

The first round was validated by three experts who provided a critical review of the Proposed E-Mail Forensic Methodology. A second round of the validation process occurred and more experts provided critical reviews of the Proposed E-Mail Forensic Methodology. Finally the third and last validation step saw final comment and review on the Proposed E-Mail Forensic Methodology. Throughout this process of expert review, all feedback, comments and suggestions were taken into account and applied accordingly where necessary. This process adds to the integrity of the findings.

The secondary data collection included a literature survey of Internet sources, frameworks, methodologies, journal articles, past research projects, reports as well as books. The literature survey was performed initially to determine the problem and the research objectives. This is the most important aspect of the secondary data collection phase as this is where the body of knowledge in the discipline is expanded upon. The combination of these data collection methods and the combination of the Design Science resulted in a Proposed E-Mail Forensic Methodology.

The Proposed E-Mail Forensic Methodology was presented and a brief introduction was given followed by an in-depth discussion of the processes within the methodology. The Proposed E-Mail Forensic Methodology is the result of a Design Science approach to the digital forensic investigation process. The Proposed E-Mail Forensic Methodologys' advantages and disadvantages were then highlighted. The applicability of the methodology was examined in a hypothetical scenario based on actual events. The research findings were then presented for expert review and comments were offered. The feedback was mostly favourable in terms of the standardisation process of the methodology; however there was criticism of the e-mail forensic aspect. This was addressed by incorporating the e-mail forensic aspect within the Proposed E-Mail Forensic Methodology.

The research objective and sub-questions posed in Chapter one of this study has now been met. In order to investigate the main research objective three sub-questions were identified. Sub-question one was addressed in Chapter two and the challenges that digital forensic investigators face were identified; the legal requirement of the investigation was addressed by including an additional challenge that focused solely on the law.

Research sub-question two was addressed in Chapter three and the key data mining techniques were examined and the most common techniques used were identified. It was demonstrated through Chapter three that investigators require automated tools in order to sift through the mountains of data that may be recovered and the issue of utilising stylometry in authorship attribution was addressed. Addtionally when stylometry is combined with automated machine learning techniques, the highest accuracy for authorship identification is achieved.

Chapter four addressed research sub-question three and here the research process models, frameworks and methodologies were examined. Unique models that address the digital forensic process in a different perspective were also taken into account. An in-depth review of these models, frameworks and methodologies was performed. A comparison was performed of the commonly referenced and used process models. These models were evaluated on the aspect of how the validation of evidence was performed. Moreover it was established that there is no best practice or standardisation of the procedures followed in the digital forensic discipline. Thus, many of the models are guidelines developed ad hoc for performing investigations and this therefore highlights the importance for the standardisation of procedures and techniques.

7.6 Future Research

This research project attempts to address the lack of standardisation of the e-mail digital forensic process. The research was positioned in the context of an e-mail forensic investigation which traditionally is no different from any other digital forensic investigation. Future research would include the implementation of this Proposed E-Mail Forensic Methodology in an actual digital forensic case or alternatively, applied retrospectively to an existing case to determine if the correct mapping of the processes was achieved. The objective is to eliminate any redundancies incorporated in the Proposed E-Mail Forensic Methodology and to cover any unforeseen gaps that may develop in the near future. A specific case study should be approached in order to experiment on the authorship identification techniques that should be incorporated into the overall investigation process. Such techniques must be tested and be able to produce credible and repeatable results in order to produce legally acceptable evidence. The data mining tools should be improved and new tools created to address the inefficiencies of the previous tools and this should be incorporated into the Proposed E-Mail Forensic Methodology.

7.7 Conclusion

This study identified the lack of standardisation that exists in the digital forensic process within the realm of e-mail digital forensics. The value of the study is determined by the contribution it makes to the existing literature that is available on the standardisation of the digital forensic investigation process and especially so in the South African context where specific laws must be taken into account. There is very little original research available in the South African context. However a number of researchers have attempted to address the issue of standardisation of the process in other countries as highlighted in the analysis of existing frameworks and models.

8. REFERENCES

FBI. (2006, September 30). Retrieved April 15, 2010, from Regional Computer Forensics Laboratory: www.rcfl.gov/Downloads/Documents/RCFL_Nat_Annual06.pdf

FBI. (2008, September 30). Retrieved April 15, 2010, from Regional Computer Forensics Laboratory: www.rcfl.gov/Downloads/Documents/RCFL_Nat_Annual08.pdf

Agarwal, R., Ahuja, M., Carter, P. E., & Gans, M. (1998). Early and Late Adopters of IT Innovations: Extensions to Innovation Diffusion Theory. *Proceedings of the DIGIT Conference* (pp. 1-18). Florida : Florida State University.

Arthur, K. K., & Venter, H. S. (2004). An Investigation into Computer Forensic Tools. *ISSA.* Pretoria: Information and Computer Security Architectures (ICSA) Research Group.

Ayers, D. (2009). A second generation computer forensic analysis system. *Digital Investigation, 6*(3), 34-42.

Baayen, H., van Halteren, H., Neijt, A., & Tweedie, F. (2002). An experiment in authorship attribution. *6th JADT* (pp. 29–37). St. Malo: Universit´e de Rennes.

Baryamureeba, V., & Tushabe, F. (2004). The Enhanced Digital Investigation Process Model. *Proceeding of Digital Forensic Research Workshop.*, (pp. 1-9). Baltimore, MD.

Berendt, B., & Draheim, M. (2007). The Image of Germany in the World: An Email and Web Mining Approach. *Künstliche Intelligenz*, 30-36.

Brody, R. G., Mulig, E., & Kimball, V. (2007). Phishing, Pharming and Identity Theft. *Academy of Accounting and Financial Studies Journal, 11*(3), 43-56.

Broucek, V., & Turner, P. (2002). E-mail and WWW browsers: A Forensic computing perspective on the need for improved user education for information systems security management. *Information Resources Management Association International Conference* (pp. 931–932). Seattle, Washington, USA : IDEA Group.

Broucek, V., & Turner, P. (2006). Winning the battles, losing the war? Rethinking methodology for forensic computing research. *Comput Virol*, 3-12.

Cardwell, K., Clinton, T., Cohen, T., Collins, E., Cornell, J. J., Cross, M., et al. (2007). *Best Damn Cybercrime and Digital Forensics book period.* United States of America: Syngress Publishing.

Carrier, B. D., & Spafford, E. H. (2003). Getting Physical with the Digital Investigation Process. *International Journal of Digital Evidence, 2*(2).

Carrier, B. D., & Spafford, E. H. (2004). An Event-Based Digital Forensic Investigation Framework. *Proceedings of Digital Forensics Research Workshop.* Baltimore: DFRWS.

Case, A., Cristina, A., Marziale, L., Richard, G. G., & Roussev, V. (2008). FACE: Automated digital evidence discovery and correlation. *Digital Investigation*, 65-75.

Casey, E. (2004). The need for knowledge sharing and standardization. *Digital Investigation*, 1-2.

Casey, E., & Stanley, A. (2004). Tool review - remote forensic preservation and examination tools. *Digital Investigation, 1*(4), 284-297.

Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics, 8*(1), 1350-1771.

Chaski, C. E. (2005). Who's At The Keyboard? Authorship Attribution in Digital Evidence Investigations. *International Journal of Digital Evidence, 4*(1), 1-13.

Ciardhuain, S. O. (2004). An Extended Model of Cybercrime Investigations. *International Journal of Digital Evidence, 3*(1), 1-22.

Collis, J., & Hussey, R. (2003). *Business Research: A practical guide for undergraduate and postgraduate students. Second edition.* New York: Palgrave Macmillan.

Colton, S., & Hatcher, T. (2004). The Web-based Delphi Research Technique as a Method for Content Validation in HRD and Adult Education Research. *Proceedings: Academy of Human Resource Development International Conference* (pp. 183-189). Austin, Texas: AHRD.

Corney, M., Anderson, A., Mohay, G., & De Val, O. (2001). Identifying the Authors of Suspect Email. *Computers Security Journal, 1*, 1-17.

Cothran, M. (1999). *Inductive and Deductive reasoning.* Retrieved October 15, 2010, from Classical-homeschooling: http://www.classical-homeschooling.org/v2/index.php?page=164

Dalkey, N. C., & Helmer, O. (1963). An experimental application of the Delphi method to the use of experts. *Management Science, 9*, 458-467.

De Val, O. (2000). Mining E-mail Authorship. *KDD-2000 Workshop on Text Mining, August 20.* Boston: Defence Science and Technology Organisation.

De Val, O., Anderson, A., Corney, M., & Mohay, G. (2001). Mining Email Content for Author Identification Forensics. *ACM SIGMOD Record, 30*(4), 55-64.

De Vos, A. S., Strydom, H., Fouche, C. B., & Delport, C. S. (2005). *Research at Grass roots: for the Social Sciences and Human Services Professions.* Pretoria: Van Schaik Publishers.

Dörre, J., Gerstl, P., & Seiffert, R. (1999). Text mining: finding nuggets in mountains of textual data. *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 398–401). San Diego, US: ACM Press.

Eloff, J., Kohn, M., & Olivier, M. (2006). Framework for a Digital Forensic Investigation. *ISSA.* University of Pretoria: Information and Computer Security Architectures (ICSA) Research Group.

Garfinkel, S. L. (2010). Digital Forensics Research: The Next 10 Years. *Digital Investigation, 7*(1), 64-73.

Giles, J. (2009, August 14). *Email compliance: email law in South Africa*. Retrieved August 18, 2010, from Michalsons Online Legal: http://www.michalsons.com/email-compliance-email-law-in-south-africa

Goel, V., & Dolan, R. J. (2004). Differential involvement of left prefrontal cortex in inductive and deductive reasoning. *Cognition*, 110-121.

Goodman, R., Hahn, M., Marella, M., Ojar, C., & Wescott, S. (2007, May 4). The Use of Stylometry for Email Author Identification: A Feasibility Study. *Proceedings of Student/Faculty Research Day, CSIS, Pace University*. White Plains, New York, USA: Pace University.

Grobler, M. M., & Von Solms, S. H. (2009a). A Best Practice Approach to Live Forensic Acquisition. *Information Security South Africa.* Johannesburg: ISSA.

Grobler, M. M., & Von Solms, S. H. (2009b). Modelling Live Forensic Acquisition. *4th International Workshop on Digital Forensics & Incident Analysis* (pp. 8- 15). Athens, Greece: CSIR.

Guba, E. G., & Lincoln, Y. S. (1994). *Competing Paradigms in Qualitative Research.* (N. K. Denzin, & Y. S. Lincoln, Eds.) Thousand Oaks, CA: Sage Publications.

Gupta, G., Mazumdar, C., & Rao, M. S. (2004). Digital Forensic Analysis of E-Mails A Trusted E-Mail Protocol. *International Journal of Digital Evidence, 2*(4), 2-11.

Hadjidj, R., Debbabi, M., Lounis, H., Iqbal, F., Szporer, A., & Benredjem, D. (2009). Towards an Integrated E-mail Forensic Analysis Framework. *Digital Investigation, 5*(3-4), 124–137.

Hamm, C. M. (1989). *Philosophical Issues in Education: An Introduction.* New York: The Fralmer Press.

Hauck, R. V., Atabakhsh, H., Ongvasith, P., Gupta, H., & Chen, H. (2002). Using Coplink to Analyze Criminal-Justice Data. *Computer IEEE, 35*(3), 30-37.

Hay, D. C. (1997, June 01). *The Zachman Framework: An Introduction*. Retrieved November 15, 2010, from The data Administration Newsletter: http://www.tdan.com/view-articles/4140/

Hershkop, S. (2006). *Behavior-based Email Analysis with Application to Spam Detection.* Columbia: Columbia University.

Hevner, A. R., & March, S. T. (2003). The Information System Research Cycle. *MIS Quarterly, 28*(1), 111-113.

Hirst, P. H. (1974). *Knowledge and the Curriculum.* Routledge and Kegan Paul.

Hsu, C.-C., & Sandford, B. A. (2007). The Delphi Technique: Making Sense Of Consensus. *Practical Assessment, Research & Evaluation*, 2-8.

Ieong, R. S. (2006). FORZA - Digital forensics investigation framework that incoporate legal issues. *Digital Investigation, 3*(1), 29-36.

Iqbal, F., Hadjidj, R., Fung, B. C., & Debbabi, M. (2008). A Novel Approach of Mining Write-Prints for Authorship Attribution in E-mail Forensics. *Digital Investigation, 5*(1), 42-51.

Jordaan, J. (2010, December 03). Personal Communication. (H. Lalla, Interviewer)

Juola, P., & Sofko, J. (2006). A Prototype for Authorship Attribution Studies. *Literary and Linguistic Computing, 21*(2), 169-178.

Katakis, I., Tsoumakas, G., & Vlahavas, I. (2007). Email Mining: Emerging Techniques for Email Management. In A. Vakali, & P. Pallis, *Web Data Management Practices: Emerging Techniques and Technologies* (pp. 219-240). Greece: Idea Group Publishing.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., & Murthy, K. R. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Computation, 13*(3), 637–649.

Kent, J., & Ghavalas, B. (2005). The unique challenges of collecting corporate evidence. *Digital Investigation, 2*(4), 239-243.

Koppel, M., & Schler, J. (2003). Exploiting Stylistic Idiosyncrasies for Authorship Attribution. *IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis.* Acapulco, Mexico.

Koppel, M., Schler, J., Argamon, S., & Messeri, E. (2006). Authorship Attribution with Thousands of Candidate Authors. *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 659-660). Seattle, WA: ACM.

Lee, H., Palmbach, T., & Miller, M. (2001). *Henry Lee's Crime Scene Handbook.* Academic Press.

Lee, S., Savoldi, A., Lim, K. S., Park, J. H., & Lee, S. (2009). A proposal for automating investigations in live forensics. *Computer Standards and Intefaces, 32*(5), 1-10.

Leigland, R., & Krings, A. W. (2004). A Formalization of Digital Forensics. *International Journal of Digital Evidence, 3*(2), 1-32.

Lim, M. J.-H. (2008). *Computational Intelligence in E-mail Traffic Analysis.* Tasmania: University of Tasmania.

Maat, S. M. (2009, August 25). *Cyber crime: A comparative law analysis.* Retrieved September 23, 2009, from Unisa: http://uir.unisa.ac.za/handle/10500/2056

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Supprt Systems, 15*(4), 251-266.

McCombe, N. (2002). *Methods Of Author Identification.* Ireland: Trinity College Dublin.

Mearian, L. (2010, May 04). *HDD industry to ship 300,000 petabytes of storage in four years*. Retrieved July 07, 2010, from Computer World UK: http://www.computerworlduk.com/ technology/storage/hardware/news/index.cfm?newsId=20114

Meyers, M., & Rogers, M. (2004). Computer Forensics: The need for Standardisation and certification. *International Journal of Digital Evidence, 3*(2), 1-11.

Michalsons, L. (2005a, June 07). *Guide to ECT Act.* Retrieved September 16, 2009, from Michalsons Attorneys: www.infoseclaw.co.za/infoseclaw.htm

Michalsons, L. (2005b, June 07). *Guide to E-mail Management.* Retrieved September 16, 2009, from Michalsons Attorneys: http://www.roylaw.co.za/Uploads/Files/Michalsons%20Infosheet%20-%20Guide%20to%20Email%20Management.pdf

Miller, L. E. (2006). Determining what could/should be: The Delphi technique and its application. *Mid-Western Educational Research Association* (pp. 2-8). Columbus, Ohio: Mid-Western Educational Research Association.

Moody, D. (2002). *Department of Computer Science.* Retrieved November 13, 2010, from Norwegian University of Science and Technology: http://www.idi.ntnu.no/~ekaterip/dif8916/ Empirical%20Research%20Methods%20Outline.pdf

Morgan, G., & Smircich, L. (1980). The Case for Qualitative Research. *The Academy of Management Review, 5*(4), 491-500.

Myers, M. D. (1997, June). *Qualitative Research in Information Systems.* Retrieved June 25, 2010, from MISQ Discovery: http://www.misq.org/discovery/MISQD_isworld

Nagwani, N. K., & Bhansali, A. (2010). An Object Oriented Email Clustering Model Using Weighted Similarities between Emails Attributes. *International Journal of Research and Reviews in Computer Science (IJRRCS)*, 2-6.

Oates, B. J. (2006). *Researching Information Systems and Computing.* London: Sage Publications Ltd.

Okoli, C., & Pawlowski, S. D. (2004). The Delphi method as a research tool: an example, design considerations and applications. *Information & Management*, 15–29.

Palmer, G. (2001, November 6). *DFRWS.* Retrieved November 1, 2009, from Digital Forensic Research Workshop: http://www.dfrws.org/2001/dfrws-rm-final.pdf

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., et al. (2006). The Design Science Research Process: A Model for Producing and Presenting Information Systems Research. *Proceedings of the 1st International Conference on Design Science in Information Systems and Technology* (pp. 84-106). Claremont, CA: Claremont Graduate University.

Pereira, C. M., & Sousa, P. (2004). A Method to Define an Enterprise Architecture using the Zachman Framework. *ACM Symposium on Applied Computing* (pp. 1366-1371). ACM Publishing.

Petroni Jr., N. L., Walters, A., Fraser, T., & Arbaugh, W. A. (2006). FATKit: A Framework for the extraction and analysis of digital forensic data from volatile system memory. *Digital Investigation*, 197-210.

Pollitt, M. M. (1995). Computer Forensics: An Approach to Evidence in Cyberspace. *Proceeding of the National Information Systems Security Conference*, (pp. 487-491). Baltimore, MD.

Pollitt, M. M. (2010, August 03). Personal Communication. (H. Lalla, Interviewer)

Powell, C. (2003). The Delphi technique: myths and realities. *Journal of Advanced Nursing*, 376–382.

Reith, M., Carr, C., & Gunsch, G. (2002). An Examination of Digital Forensic Models. *International Journal of Digital Evidence, 1*(3), 1-12.

Rogers, E. M., Singhal, A., & Quinlan, M. M. (2007, June 19). Diffusion of Innovations. New York, New York, USA: Free Press.

Rogers, M. (2005, February). *Ministry of Citizens Services.* Retrieved June 10, 2010, from British Columbia: www.lcs.gov.bc.ca/privacyaccess/Conferences/Feb2005/ ConfPresentations/Marcus_Rogers.pdf

Rowlingson, R. (2004). A Ten Step Process for Forensic Readiness. *Internation Journal of Digital Evidence, 2*(3), 1-28.

Roydhouse, T. B. (1999). Essay on the extent, if any, to which the emergence of new technologies of electronic communication undermine, or threaten to undermine, the justifications for the sub judice rule. New South Wales, New Zealand.

Rudman, J. (1998). The State of Authorship Attribution Studies: Some problems and solutions. *Computers and the Humanities, 31*, 351–365.

Ryan, D. J., & Shpantzer, G. (2005, April). Legal Aspects of Digital Forensics. Washington, Washington, D. C, United Stated of America.

Selamat, S. R., Yusof, R., & Sahib, S. (2008). Mapping Process of Digital Forensic Investigation Framework. *International Journal of Computer Science and Network Security, 8*(10), 163- 169.

Skulmoski, G. J., Hartman, F. T., & Krahn, J. (2007). The Delphi Method for Graduate Research. *Journal of Information Technology Education, 6*, 1-21.

Smith, J. (2008, April 28). *A review of authorship attribution.* Retrieved February 15, 2010, from McGil University: www.music.mcgill.ca/~jordan/coursework/mumt611/jsmith_final_paper.doc

Smith, P. J. (1998, February 10). *Case Law R. vs. Weir*. Retrieved August 15, 2010, from University of Ottawa: http://aix1.uottawa.ca/~geist/Weir.html

Stephenson, P. (2002). The Forensic Investigation Steps. *Computer and Fraud Security, 10*, 17-19.

Stolfo, S. J., & Hershkop, S. (2005). Email Mining Toolkit Supporting Law Enforcement Forensic Analyses. *Proceedings of the 2005 national conference on Digital government research* (pp. 221-222). New York: Digital Government Society of North America.

Szezynska, M., Huebner, E., Bem, D., & Ruan, C. (2009). Methodology and Tools of IS Audit and Computer Forensics – The Common Denominator. *Springer-Verlag Berlin Heidelberg*, 110-121.

Taylor, M., Haggerty, J., & Gresty, D. (2009). The legal aspects of corporate e-mail investigations. *Computer Law & Security Review, 25*(4), 372–376.

Waddoups, C. (2010, September 2). *Electronic Case Filing System.* Retrieved September 25, 2010, from U.S. District Court: https://ecf.utd.uscourts.gov/cgi-bin/show_public_doc?2006cr0811-400

Walters, A., & Petroni, N. L. (2007). Volatools: integrating volatile memory forensics into the digital investigation process. *Black hat DC*.

Watney, M. (2009). Admissibility of Electronic Evidence in Criminal Proceedings: An Outline of the South African Legal Position. *Journal of Information, Law & Technology, 1*, 1-10.

Witten, I. H., Frank, E., Trigg, L., Hall, M., Holmes, G., & Cunningham, S. J. (1999). Weka: Practical Machine Learning Tools and Techniques with Java Implementations. *Proceedings of the ICONIP/ANZIIS/ANNES'99 Workshop on Emerging Knowledge Engineering and Connectionist-Based Information Systems.* (pp. 192–196). New Zealand: Department of Computer Science, University of Waikato.

Wonglimpiyarat, J., & Yuberk, N. (2005). In support of innovation management and Roger's Innovation Diffusion theory. *Government Information Quarterly, 22*(3), 411–422.

Yasinsac, A., Erbacher, R. F., Marks, D. G., Pollitt, M. M., & Sommer, P. M. (2003). Computer Forensics Education. *IEEE Security & Privacy, 1*(4), 15-23.

Zachman, J. A. (2004, March 26). *Concept of the Framework for Enterprise Architecture.* Retrieved November 15, 2010, from Association of Enterprise Architects: http://www.aeablogs.org/eakd/files/Zachman_ConceptsforFrameworkforEA.pdf

Zachman, J. A. (2008). *The Zachman Framework: The Official Concise Definition.* Retrieved November 15, 2010, from Zachman International: http://www.zachmaninternational.us/index.php/home-article/13#maincol

Zhang, X., Liu, J., Zhang, Y., & Wang, C. (2006). Spam behavior recognition based on session layer data mining. *Prococeedings of the 3rd International Conference on Fuzzy Systems and Knowledge Discovery* (pp. 1289-1298). Xi'an, China: LNAI.

Zheng, R., Li, J., Chen, H., & Huang, Z. (2006). A Framework for Authorship Identification of Online Messages: Writing-Style Features and Classification Techniques. *Journal of the American Society for Information Science and Technology, 57*(3), 378–393.

Acronyms

C#	C Sharp
CA	Chartered Accountant
CD	Compact Disk
CD ROM	Compact Disk Read Only Memory
CISA	Certified Information Systems Auditor
CPA	Chartered Public Accountant
DFRWS	Digital Forensic Research Workshop
DNA	Deoxyribonucleic Acid
ECT Act	Electronic Communications and Transactions Act 25
EMT	E-mail Mining Toolkit
EEE	EnCase Enterprise Edition
EIDIP	Enhanced Digital Investigation Model
FBI	Federal Bureau of Investigation (US Government)
FORZA	FORensics ZAchman framework
FTK	Forensic Toolkit
FY	Fiscal Year
GB	Gigabyte (1024 megabytes)
IDIP	Integrated Digital Investigation Process
IEFAF	Integrated E-mail Forensic Analysis Framework
IDC	International Data Corporation
IS	Information Systems
ISP	Internet Service Provider
MDA	Mail Delivery Agent
MRA	Mail Retrieval Agents
MTA	Mail Transfer Agent
MUA	Mail User Agent
PDA	Personal Digital Assistant (electronic handheld information device)
POP3	Post Office Protocol

POS	Part of Speech
PDIR	ProDiscover Incident Recovery
RAM	Random-Access Memory
RICA	Regulation of Interception of Communications and Provision of Communication-Related Information Act
SMTP	Simple Mail Transfer Protocol
SMS	Short Message Service (cellular phone text messaging)
SVM	Support Vector Machine
SABSA	Systems and Business Security Architecture
USDOJ	U.S. Department of Justice Forensics
US	United States
USA	United States of America
VIN	Vehicle Identification Number
WEKA	Waikato Environment for Knowledge Analysis
WWW	World Wide Web
XML	Extensible Markup Language

Glossary

Term	Meaning
Authorship Categorisation	This is a process of allocating a set of rules to identify the author (De Val et al., 2001)
Data Mining	Is the process of analysing data and extracting patterns from data from different perspectives and summarising it into useful information
Design Science	Design Science is technologically orientated and is essentially a problem solving process that leads to the development of an effective artefact, which is of four types: namely constructs; methods; models and implementations (March and Smith, 1995)
Digital Forensics	*"The use of scientifically derived and proven methods toward the preservation, collection, validation, identification, analysis, interpretation, documentation and presentation of digital evidence derived from digital sources for the purpose of facilitating or furthering the reconstruction of events found to be criminal, or helping to anticipate unauthorized actions shown to be disruptive to planned operations"* (Palmer, 2001)
Discipline	A branch of knowledge or teaching
E-Mail	Electronic mail is the telecommunication of messages from one computer to another. E-mail relies on two basic communications protocol: Simple Mail Transfer Protocol (SMTP), which is used to send messages and Post Office Protocol (POP3), which is used to receive messages (Katakis, Tsoumakas and Vlahavas, 2007)
E-Mail Mining	The process of analysing e-mail is called e-mail mining, which is a process of discovering useful patterns from e-mails (Nagwani and Bhansali, 2010)
Electronic Evidence	Electronic evidence refers to electronic data which is sought, located, secured, and searched with the intent of using it as evidence in a civil or criminal case
Interpretivist	Assume that access to reality (given or socially constructed) is only through social constructions such as language, consciousness and shared meanings (Myers, 1997)
IT Artefact	IT artefacts are broadly defined as constructs (vocabulary and symbols), models (abstractions and representations), methods (algorithms and practices), and instantiations (implemented and prototype systems) (Hevner and March, 2003)
Framework	The underlying structure. A computer forensic framework can be defined as a structure to support a successful forensic investigation (Eloff, Kohn and Olivier, 2006)
Genuine User	User who uses e-mail for its intended purpose with no intent to commit any crime (Gupta, Mazumdar and Rao, 2004)
Methodology	The system of methods followed in a particular discipline
Model	A hypothetical description of a complex entity or process

Term	Meaning
Paradigm	The generally accepted perspective of a particular discipline at a given time
Positivist	Assume that reality is objectively given and can be described by measurable properties which are independent of the observer (researcher) and his or her instruments (Myers, 1997)
Procedure	A set of established forms or methods for conducting the affairs of an organised body such as a business, club, or government
Qualitative	Relating to or involving comparisons based on qualities
Quantitative	Relating to the measurement of quantity
Standardisation	The condition in which a standard has been successfully established
Stylometry	Assumes that an author has distinctive writing habits and are exhibited in features such as vocabulary use, sentence complexity and phraseology. Stylometry is the determination of authorship from writing styles (De Val et al., 2001)
Stylometric Features	Stylistics or the study of stylometric features shows that individuals can be identified by their relatively consistent writing styles. Stylometric features include but are not limited to lexical, syntactic, structural and content-specific features. (Iqbal et al., 2008)
Support Vector Machine	Automated classification technique limited to binary classifications based on the structural risk minimisation principle to determine authorship (De Val et al., 2001)
Techniques	The systematic procedure by which a complex or scientific task is accomplished
Write-Print	Based on the concept of Frequent pattern and is a combination of patterns that are used to determine a true author (Iqbal et al., 2008)