TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Digital forensics is a growing field that is gaining popularity among many computer professionals, law enforcement agencies, investigators and other digital forensic practitioners. For this reason, several investigation process models have been developed to offer direction on how to recognize and preserve potential digital evidence obtained from a crime scene. However, the vast number of existing models and frameworks has added to the complexity of the digital forensic field. This situation has further created an environment replete with semantic disparities in the domain, which need to be resolved. Note that the term 'semantic disparities' is used in this thesis to refer to disagreements about the interpretation, description and representation of the same or related digital forensic data or information and terminologies.

In a world where digital technology keeps changing and the evolution of the digital forensic domain continues, it would be appropriate to develop and standardise dynamic and practical methods that can help to resolve many of the present and future disparities bound to occur in digital forensics. Such methods will further aid in creating uniformity in the interpretation, description and representation of the same or related digital forensic data or information. The interpretation, description and representation of digital forensic data or information are important, especially during the digital forensic investigation process, in order to conform to the uniformity of investigative terminologies so that misunderstandings between investigators and other parties, e.g. judges, does not happen.

In this research study, therefore, the researcher employs a pragmatic approach to research and proposes a sematic reconciliation model for resolving semantic disparities in digital forensics. The study is conducted in two phases where the first phase involves investigating the various challenges that digital forensics have faced to date – in a bid to demonstrate the semantic disparities that exist in digital forensics. In the second phase, a model coined as the Digital Forensic Semantic Reconciliation (DFSR) model is presented in an attempt to provide directions in resolving the semantic disparities that occur in the digital forensic domain. The researcher also demonstrates in this study a prototype implementation of the DFSR model called the DFSR prototype.

Finally, to assess the efficiency of the DFSR prototype, several experiments are conducted and the results discussed. All the experiments conducted to test the feasibility and implementations of the proposed DFSR model in this study have delivered remarkable results. Therefore, the proposed DFSR model in this study can be used as an initial guide towards

resolving semantic disparities in digital forensics. The proposed DFSR model, for example, can also be helpful in facilitating the harmonisation and/or uniformity in the interpretation, description and representation of the same or related digital forensic data or information within the field of digital forensics.

LIST OF ABBREVIATIONS USED

AF	Anti-Forensics
APIs	Application Programming Interfaces
CRM	Customer Relationship Management
DBMS	Database Management Systems
DF	Digital Forensics
DFASSV	Digital Forensic Absolute Semantic Similarity Value
DFRWS	Digital Forensics Research Workshop
DFSR	Digital Forensic Semantic Reconciliation
DIMS	Distributed Information Management System
DLP	Data Loss Prevention
DOME	DERI Ontology Management Environment
ERP	Enterprise Resource Planning
FMA	Forensic Memory Analysis
FTK	Forensic Toolkit
GPS	Global Positioning System
IaaS	Infrastructure-as-a-Service
IDSs	Intrusion Detection Systems
ISP	Internet Service Provider
KIF	Knowledge Interchange Format
NACOSTI	National Commission for Science, Technology and Innovation
NAS	Network Attached Storage
NIST	National Institute of Standards and Technology
OKBC	Open Knowledge Base Connectivity
OMWG	Ontology Management Working Group
OS	Operating System
PaaS	Platform-as-a-Service
PDAs	Personal Digital Assistants
RAID	Redundant Array of Independent Disks
RAM	Random Access Memory
RDF	Resource Description Framework
RFID	Radio-Frequency Identification

SaaS	Software-as-a-Service
SAN	Storage Area Network
SIM	Subscriber Identity Module
SSL	Secure Sockets Layer
ST&I	Science, Technology and Innovation
SUMO	Suggested Upper Merged Ontology
TCP/IP	Transmission Control Protocol/Internet Protocol
TOVE	TOronto Virtual Enterprise
TSK	The Sleuth Kit
VoIP	Voice-over-IP
WOL	Web Ontology Language
XML	Extensible Mark-up Language

ACKNOWLEDGEMENT

First and foremost, I want to thank God for giving me the breath of life, the strength to work tirelessly, the grace to endure and all the talents that have enabled me to complete this PhD research study.

Second is a very splendid, and well-deserved, thank you to the following special people: To Prof. Hein S. Venter – thank you for accepting me as a PhD research student and for the invaluable guidance that has constantly been available to me during my studies. As my supervisor, Venter introduced me to the world of digital forensics and made it easier for me to successfully bring together my research ideas. He also gave me direction on the execution of those ideas as well as insights on how to interpret my research work. He further supported discussions of all my ideas and research endeavours. He was always supportive when it came to financial needs, such as conferences and presentations among others. Finally, as a friend and my mentor, his encouragement and exposition to my research methods and writings aided me significantly to enhance all my undertakings throughout this research.

To my family – thank you and much love to my parents, my brothers Edison Lewa and Samuel Piri, my sisters Claudia Furaha and Anita Saumu. Not forgetting my close friend and Uncle, Alex Shume and Aunt Agnes Shume. My role model Dr Ngalla Jillani. Thank you all for the continuous prayer and encouragement. My precious friend and sister, Faith Uchi Shume – thank you for always standing with me during tough times. Much love also goes to Hope, Mark, Esther and Sharon Chiru "Tabia mbii be kisha", Caleb "mtumia wa mudzi" and Lewis "the clever boy", Wawe chiro and Sangazimi, kudos.

To my friends and colleagues, the list is endless – Enos Mabuto, thanks for the chicken eggs you always brought into the ICSA lab. Emilio, Shayur and Gilbert – thanks for your endless help. Stacy Omeleze – thank you so much for your continued encouragement during tough times, especially when I felt like quitting or even changing my research topic. Friends from across the borders – Victor Kebande, "the Botnet man" who always encouraged me; Ken Sigar, Koila, Zipporah Mwololo, Ruth Oginga, John Chebor, Timothy Sawe, Nelson Masese, Joseph Orori, Cleophas Mochogo and Justine Oguta – thank you so much. Dr Maghanga – thank you for your continued encouragement. Prof. Peter Kibas – thank you for always reminding me and pushing me to start my PhD studies. Fredrick Ogore – thank you for the prayers, you are blessed together with your wife Linnet. And all those whom I did not mention because of a lack of space – thank you.

My sincere gratitude also goes to all the individuals and digital forensic professionals around the world who helped in reviewing my research work and published valuable references for my writing. Your accomplishments and deep understanding of digital forensics helped provide the necessary materials for my work. Much love and thanks are also due to Miriam Sulubu kahindi ("the walking dictionary") for reading my work and working on my grammar. You are A-may-zing. To my darling Winnie Charlotte Ashioya thanks for accepting me even with my busy schedules and agreeing to stick by my side even when others chose to back off. Your continued encouragement continuously brightens my tomorrow. God bless you.

Finally, studying requires money, both for living expenses and for the research effort itself. In this regard I thank the University of Pretoria for giving me the Special International Research Award, Postgraduate Doctoral Research Award and Postgraduate Research Support Bursaries. These bursaries and awards were a blessing to me. I also thank Kabarak University and all my close friends and relatives for their financial assistance. Special thanks also go to the National Commission for Science, Technology and Innovation (**NACOSTI**) for their generous financial support towards my higher education. The Commission awarded me the Science, Technology and Innovation (ST&I) grant towards my PhD research project (**NACOSTI/RCD/ST&I 5TH CALL PhD/122**).

Thank you all. God bless you forever.

Nickson M. Karie

July, 2016

CHAPTER 1 : INTRODUCTION

1.1 INTRODUCTION

The importance of digital forensics (DF) in today's modern digital society cannot be stressed enough. This is because digital forensics often plays an indispensable role in both civil proceedings and criminal cases. For example, in the case of a digital investigation, any potential digital evidence captured during the digital forensic investigation process must ultimately be presented in the form of expert reports, depositions and/or testimony in legal or civil proceedings (Brezinski and Killalea, 2002). If the interpretation, descriptions and representation of the same or related digital forensic data or information are conducted in a uniform and/or standard way, it becomes easy and useful in incarcerating any attacker. In this case, the data or information presented stands a much greater chance of being acceptable in court in the event of a prosecution (Brezinski and Killalea, 2002). Wrongly interpreted, described and represented data or information, on the other hand, may create loopholes for perpetrators to exploit, thus, making it hard to convict and prosecute them.

However, with a uniform and/or standard way for interpreting, describing and representing digital forensic data or information, the digital forensic experts and law enforcement agencies can determine – with less effort – the admissibility of any potential digital evidence presented in court. For the purposes of this research, and the remainder of this thesis, the phrase 'data or information', will be used to refer to 'any potential digital evidence captured during the digital forensic investigation process'.

Knowing that digital forensics is considered a growing field, new digital forensic terminologies are bound to appear at any point in the domain. At times, the new terminologies used to describe existing domain data or information may contradict old terminologies in their intended interpretation, description and representation. This situation, thus, causes variations in the understanding of domain terminologies and moreover creates an environment replete with semantic disparities in the domain, which need to be resolved. Semantics, according to Richmond (2012), is the study of the meaning of linguistic expressions. However, semantics can also be used to refer to the interpretation, description and representation of a word or a phrase in a sentence.

The term “Semantic disparities” is, thus, used in this thesis to refer to disagreements about the interpretation, description and representation of the same or related digital forensic data or information and terminologies. An example of such semantic disparities is shown in Table 1.3 representing semantic conflicts in the digital forensic domain.

For this reason, there is a vital need to develop models in the digital forensic domain that can assist in interpreting, describing or even presenting the most common representations of digital forensic data or information in any court of law or during civil proceedings. Such models would help resolve the semantic disparities that occur in digital forensic domain as well as those outside the domain.

One particular example of a semantic disparity problem in digital forensics that stands out to motivate this research is stated as follows. With the emergence of cloud computing technologies, Virtual machine introspection (VMI) has now become the foundation for many other novel approaches to cloud security (Dolan-Gavitt et al., 2011). VMI is software that runs on the virtual machine host and examines the contents of a virtual machine in real time in order to determine if, for example, malware is hampering the particular virtual machine on that host. Such software may be able to externally isolate such malware running internally on the virtual machine. This isolation technology has made it possible for virtualized environments to provide advanced security features in the cloud. In a research by Dolan-Gavitt et al. (2011) the authors state that any application that leverages on VMI must overcome the semantic disparity problem experienced especially during the reconstruction of high-level state information from low-level data sources such as physical memory.

Dolan-Gavitt et al. (2011) state that the advances made by the digital forensic community in reconstructing high-level state information from physical memory dumps are not generalised, however, most of the work focuses on recovering information about specific operating systems and versions where disparity in terminologies are present. Although support for other operating systems is increasing, forensic memory analysis (FMA) does not provide a complete solution to the sematic disparities problem.

It is for this reason that the digital forensics community has for many years grappled with semantic disparity problems in the field of FMA which seeks to extract forensically relevant information from dumps of physical memory. In the absence of a general solution, the information provided by FMA can enable a variety of new VMI

applications and allow researchers to avoid duplicating effort when implementing new systems. Because VMI examines a live system, it also has access to information beyond what is available in a forensic context. In addition to the static view of physical memory provided by a memory dump, VMI applications can watch the state of the system as it changes over time. This allows the use of techniques that are not possible in offline analysis.

In another research, Jones et al. (2006) presented Antfarm for tracking processes in a virtual machine environment. Antfarm tracks the value of a certain register to determine what processes are running inside the virtual machine. Since a memory image gives a view of the system state at a single point in time, this technique will only work in a live environment. Although much of the same information is available with both VMI and FMA, there are some differences associated with VMI that are not present with offline forensic analysis. Because virtual machines are running during VMI, the CPU and memory state of the virtual machine will constantly change as VMI is performed unless the CPUs are suspended before analysis, which would not be acceptable for the virtual machine owner. By contrast, FMA tools need only be concerned with memory volatility at the time the memory snapshot is taken: although the system continues to run throughout the acquisition process, the results are static and can be examined offline. Despite these differences, FMA tools can be of great value to VMI.

Therefore, in a world where digital technology keeps changing and the evolution of the digital forensic domain continues, it would be appropriate to develop dynamic and practical methods that can help to resolve many of the present and future sematic disparities bound to occur in digital forensics, as was demonstrated in the above example of VMI and FMA. Such methods will further aid in creating uniformity in the interpretation, description and representation of the same or related digital forensic data or information.

According to Lin et al. (2006), the problem of semantic disparity becomes even critical in situations of extensive cooperation and interoperation between distributed systems across different enterprises. In the case of digital forensics, for example, such a situation would make it difficult to manipulate distributed data/information in a centralized manner. This is because; the contextual requirements and the purpose of the information across the different systems may not be homogeneous. This situation

therefore serve to motivate this research in coming up with methodologies and specifications aimed at resolving sematic disparities in the digital forensic domain.

Another example of a court case that stands out to motivate this research, though not in the digital forensic domain, is that of Morse et al. vs. Frederick (2007). The case was argued in March 19, 2007 and decided June 25, 2007 at the supreme court of the United States (Bill, 2007).

The case pertains to an event that took place when high school students were allowed to go out and observe an Olympic Torch Relay. A large number of students gathered, among them Joseph Frederick. The students held up a banner which had the words "BONG HiTS 4 JESUS". The then principal of the high school, Deborah Morse, ordered the student to take down the banner but Joseph Frederick declined to do so. For this reason Joseph Frederick was suspended from school (Bill, 2007; Morse et al. vs. Frederick, 2007). The interpretation, descriptions and representation of the words "BONG HiTS 4 JESUS" caused problems in court. This scenario motivated this research on how semantic disparities can cause problems in court especially after a digital forensic investigation process has been done, hence, the need to resolve it.

Moreover, being a growing field, digital forensics is still gaining popularity among many computer professionals, law enforcement agencies, digital forensic practitioners and other stakeholders who need to cooperate in this profession. However, the cooperation between the stakeholders mentioned presupposes the reconciliation of any disparities that are bound to occur in this domain, such as structural semantic disparity, conceptual disparity as well as semantic conflicts (to name a few). These disparities, and other terminologies used frequently throughout this thesis, are briefly explained in the subsections to follow as a basis for why this research was found to be necessary.

1.1.1 Structural Semantic Disparity

Structural semantics refers to the relationships that exist between the meanings of words or phrases used within a sentence (Colomb (1997). It is, thus, possible to break down structural semantics found in a sentence into atomic semantic features. Atomic semantic features in this case refer distinctive properties of the meaning of a word or phrase used in the sentence. This also means that, the atomic semantic features contribute to the meaning of the word or phrase in the sentence.

The atomic semantic features in any sentence can also be used to refer to the actual and/or distinctive properties, objects or relations in the world (Peter H.M., 2001). As a general example, in linguistics, a plus and minus signs is usually used to show the presence or absence of pre-established atomic semantic feature as shown below:

Man is [+HUMAN], [+MALE], [+ADULT] – this is interpreted as, a man is human, male as well as an adult

Woman is [+HUMAN], [-MALE], [+ADULT] – this is interpreted as, a woman is human, Not male but an adult

Some examples of atomic semantic features associated with different terminologies in the digital forensic domain are shown in Table 1.1. Infer from Table 1.1 below that, a plus and minus signs has been used to show the presence or absence of pre-established atomic semantic features.

Table 1.1 Structural Semantic Features

Digital Forensic Terminologies	**Semantic Features**
Digital evidence	[+Digital in nature], [-tangible], [+found within digital devices].
Initial response	[+the first response], [+initial activities after an incident has occurred].
Analysis	[+Identify digital evidence data], [+interpreting data], [+reconstructing data].
Examination	[+in-depth evidence analysis], [+digital forensic tools enactment] [+digital evidence data collection methods]

As shown in Table 1.1 the interpretation, description and representation of the atomic semantic features during a digital investigation process and evidence presentation can result in structural semantic disparities if not done in a uniform and/or standard way. The term “examination” for example with an atomic sematic feature “in-depth evidence analysis”, yet “analysis” is, however, defined separately with its own atomic semantic features can cause structural semantic disparities during a digital investigation process and evidence presentation, hence, the need to resolve it.

1.1.2 Conceptual Disparity

Conceptualization can be defined as a simplified view of some terms containing general notions and other objects that are believed to be of interest for some particular purpose and the relationships between them (Gruber and Thomas, 1993 and Barry, 2003). A clearly and detailed definition of a conceptualization can be referred to as ontology, and at times a conceptualization can be realized by several distinct ontologies (Gruber and Thomas, 1993). Note that ontology is a comprehensive formal definition of how to represent objects that exist in a given domain of interest and the relationships that holds among them (Smith et al., 2006). Ontology can also be defined as a set of well-defined general notions depicting a particular domain of interest (Van Rees, 2003). Grüber, (1993) however, defines ontology as a clearly and detailed definition of a conceptualisation. Conceptual disparity, thus, may occur in digital forensics when the terms used in two different ontologies have meanings that are similar, yet not quite the same (Colomb 1997, Xu and Lee, 2002). For example, the terms "analysis" and "examination" which is meant to have different semantic meaning, may appear in different ontologies to have similar meanings. These two terms, as shown in the Table 1.2, can bring about conceptual disparity if their semantic meanings are not defined properly. This is evident from their conceptualized meaning shown in Table 1.2.

Table 1.2 Different Conceptualization of Terminologies

Digital Forensic Phrase	Conceptualized Meaning
• Analysis	The use of different digital forensic tools and methods to make sense of the gathered digital evidence data (Sibiya et al., 2012).
• Examination	Examination is a comprehensive and thorough ***analysis*** of digital evidence data and the enactment of digital forensic tools and methods used to collect the digital evidence (Lalla and Flowerday, 2010).

1.1.3 Semantic Conflicts

Table 1.3 below shows examples of semantic conflicts that currently exist in the digital forensic domain. This situation in digital forensics adds up to motivate this

research. One can see from Table 1.3 that, essentially, the three terms listed are supposed to mean the same as can be inferred from their descriptions.

Table 1.3 Semantic Conflicts in Digital Forensics

DF Terminology	Semantic Conflicts Descriptions
• First Response	Include the first response to the detected incident (Valjarevic and Venter, 2012).
• Initial Response	Initial response means, performing initial investigations. This includes recording of basic details related to the incident under investigation. It also involves, assembling the incident response team, and informing all the people who need to know about the incident (Mandia et al., 2003).
• Incident Response	Incident response is made up of the process of identifying the presence of concealed evidence and initial, pre-investigation response to a suspected crime such as a breach of computer security. The purpose of Incident response is also to discover, prove the validity, evaluate, and deduce a response action plan for the suspected breach of security (Beebe and Clark, 2005).

It is evident from Table 1.3 that, methodologies and specifications therefore need to be developed in digital forensics so as to assist in resolving any disparities that are bound to occur in this domain. This also includes standardisation of methodologies and specifications for resolving any semantic disparities in digital forensics. Furthermore, the requirement for such methodologies and specifications in digital forensics is of great significance; both for the betterment of the domain and for the effective interpretation, description and representation of digital forensic data or information.

In the research study presented in this thesis, however, the researcher's main focus is on resolving Semantic Disparities (SD) in digital forensics. The term 'Semantic Disparities' as said earlier is used here to refer to disagreements about one or more of the following: the interpretation, description and representation of the same or related digital forensic data, information and terminologies (Xu and Lee, 2002). For the sake of simplicity, the phrase 'digital forensic data, information and terminologies', will be abbreviated to the phrase 'digital forensic terminologies' in the remainder of this thesis.

Unless semantic disparities are detected and resolved in digital forensics, it may lead to misunderstandings especially during a digital forensic investigation process. Even worse, since computer professionals, law enforcement agencies and digital forensic

practitioners may not always have the same background knowledge or may not be from the same jurisdiction, they may not be aware of the existence of such semantic disparities during investigations. This situation therefore presents a problem that needs to be resolved as stated in the problem statement section (Section 1.2). This is then followed by the motivation for this study in Section 1.3. Section 1.4 presents the objectives of the study while Section 1.5 explains the thesis layout. Section 1.6 concludes by presenting the chapter conclusions.

1.2 PROBLEM STATEMENT

This research study recognises the importance of uniformity in the interpretation, description and representation of digital forensic terminologies. Uniformity of terminology use among computer professionals, law enforcement agencies and other digital forensic practitioners is vital, especially during a digital forensic investigation process. Therefore, the main problem tackled in this research study is the lack of methods and/or specifications specifically designed for resolving semantic disparities in the digital forensic domain. The study also examines the consequences that the problem statement have in the digital forensic domain and in particular, the consequences to computer professionals, law enforcement agencies and digital forensic practitioners.

Semantic disparities may occur in digital forensics when the different communicating parties apply different interpretations, descriptions and representations to the same or related digital forensic terminologies. This situation may further cause misapprehension and confusion when attempts are made to harmonise data or information emanating from different sources (Piasecki, 2008). Misapprehension can lead to other problems such as errors in evidence analysis during a digital forensic investigation process.

It is therefore critical that any identified semantic disparities be resolved to ensure uniformity in the interpretation, description and representation of digital forensic terminologies. The problem area identified in this research study can be broken down further into the following research questions:

1. Besides the semantic disparities that occur in digital forensics, what other challenges does digital forensics face?
2. When do semantic disparities occur in digital forensics and what are the current efforts undertaken to resolve them?

3. Of what significance is the whole process of resolving semantic disparities to computer professionals, law enforcement agencies and practitioners in digital forensics?
4. Can ontologies be used in resolving semantic disparities in digital forensics as well as help achieve a unified formal representation of digital forensic domain terminologies?
5. What types of practical methods currently exist that can help to resolve semantic disparities in digital forensics?

Note that the study presented in this research thesis and as reflected in the research questions above is motivated by the need to develop methods that can assist in resolving semantic disparities in digital forensics. This also includes methods that can be used to help achieve a unified formal representation of digital forensic domain terminologies. For this reason, the next section presents more details of what motivated this study.

1.3 MOTIVATION FOR THIS STUDY

The research undertaken for this study was primarily motivated by the realisation that, as the evolution in digital technology goes on, computers and other digital systems are becoming better connected through all kinds of technology and computer networks (Afcea, 2014). Computer crime techniques are also becoming more sophisticated and better coordinated. This evolution in digital technology as a result complicates the crime techniques as well (Afcea, 2014). In addition, the amount of digital information generated and handled by different digital systems – on a daily basis – is enormous. New terminologies emerge to describe different scenarios, as well as particular data or information in digital forensics. For instance, the term ‘digital forensics’, derived as a more encompassing synonym of computer forensics, has broadened over time to incorporate the investigation of all devices with the ability to store digital data. The technical aspects of digital forensic investigations now cover different sub-disciplines – among others, computer forensics, network forensics, (mobile) device forensics, database forensics and software forensics.

Besides, as of the time of writing this thesis the researcher was actively involved in contributing to the process of creating an international standard (ISO/IEC 27043, 2015) for the digital forensic investigation process where the need arised to carefully define,

describe and reason about specific digital forensic terminologies. The ISO/IEC 27043 was later published as an international standard in March 2015.

With the continued evolution in digital technology and the increase in the amount of digital information generated and handled by different computers and digital systems every day, it is clear that in the future, access will have to be provided to more digital information and new terminologies than can reasonably be predicted. Hence, it is crucial at this point to develop dynamic and practical methods that can help create uniformity in the interpretation, description and representation of the digital forensic terminologies. Such methods in digital forensics can assist in resolving many of the present and future semantic disparities that are bound to occur.

Unfortunately, even at the time of writing this thesis, digital forensics lacks internationally standardised methods that have been designed specifically to assist in resolving semantic disparities. The lack of standardisation in many areas of digital forensics as noted by Chaikin (2006) may also contribute to the semantic disparities that now exist in the domain. This situation also supports the motivation of this research.

That being the case, this study seeks to point out the semantic disparities that occur in digital forensics using a pragmatic (mixed methods) approach to research. This approach as used in this study involved using several methods which were deemed best suited to the research problem at hand. Additionally, this research study also proposes the use of digital forensic ontologies as a way towards resolving semantic disparities as well as creating a unified formal representation of the digital forensic domain knowledge and information. This is backed up by a research carried out by Hoss and Carver, (2009) showing that there is currently no such representation of the digital forensic domain knowledge or standardised procedures for gathering and analysing digital forensic knowledge (Hoss and Carver, 2009).

The lack of a unified formal representation of domain knowledge results in inevitable disparities among digital forensic analysis tools, let alone the digital forensic domain terminologies (Hoss and Carver, 2009). Needless to say, errors in interpretation, description and representation of digital forensic terminologies, in the case of a digital forensic investigation process, are more likely to occur where there are no standardised procedures or formal representation of the digital forensic domain data or information (Chaikin, 2006). The specific actions taken by the researcher to achieve the research goals are described in the objectives section to follow.

1.4 OBJECTIVES

Knowing that there were no methods or specifications specifically designed for resolving semantic disparities at the time of this study, the primary objective of this study was, thus, to develop a method for resolving semantic disparities in digital forensics. However, the other contributions of this research study include the enhancement of uniformity in the interpretation, description and representation of digital forensic terminologies. This study also considers how to use ontologies in resolving semantic disparities. Ontologies are also introduced in this research study as a way to create a unified formal representation of the digital forensic domain knowledge and information. The presentation in this thesis is further meant to help build a foundation for future undertakings on how to use ontologies in resolving semantic disparities in digital forensics.

The objectives of this study are shown more specifically in the bulleted list below. However, note that each of the specific objectives listed are rendered from the research questions mentioned earlier. Moreover, each of the specific objectives below is also represented in specific chapters of this research thesis.

- To present, besides the semantic disparities, a literature review on digital forensics and the different challenges currently faced by digital forensics.
- To discuss semantic disparities in digital forensics as well as establishing the current efforts undertaken towards resolving them when they occur in the domain.
- To establish the significance of resolving semantic disparities to computer professionals, law enforcement agencies and digital forensic practitioners.
- To present a literature review on ontologies as well as develop ontologies within the digital forensic domain that can assist in resolving semantic disparities and building a body of unified formal representation of digital forensic domain knowledge.
- To propose a model supported by a prototype developed from dynamic, practical methods and specifications for resolving semantic disparities in digital forensics.

Note that the proposed model in this thesis is intended to help in producing uniform results, i.e. results that are similar in interpretation, description and representation of the

different digital forensic terminologies. The specific chapters based on the outlined objectives in this research thesis are expounded with the help of Figure 1.1 below.

1.5 THESIS LAYOUT

This thesis consists of six parts and nine chapters. Part One comprises the introduction chapter (Chapter 1), while Part Two contains three background chapters (Chapters 2, 3 and 4). Chapters 5 constitute Part Three and explore the semantic disparities in digital forensics. Part Four comprises Chapters 6 and 7, which concentrate on establishing a solution to the semantic disparity problem in digital forensics. Chapters 8 make up Part Five and focus on testing the feasibility of the proposed solutions in Chapters 7. Finally, Part Six consists of Chapters 9, which conclude the thesis.

The current chapter (Chapter 1) explains and introduces the research problem, motivation and objectives. The rest of the thesis is, however, organised as shown in Figure 1.1, followed by an outlook for each of the chapters. Some of the work presented in this thesis has already been published in conference proceedings as well as in scientific journals as shown in Appendix A. Mostly, the chapters are organised in a manner that corresponds to the research papers listed in Appendix A as well as the research objectives listed earlier.

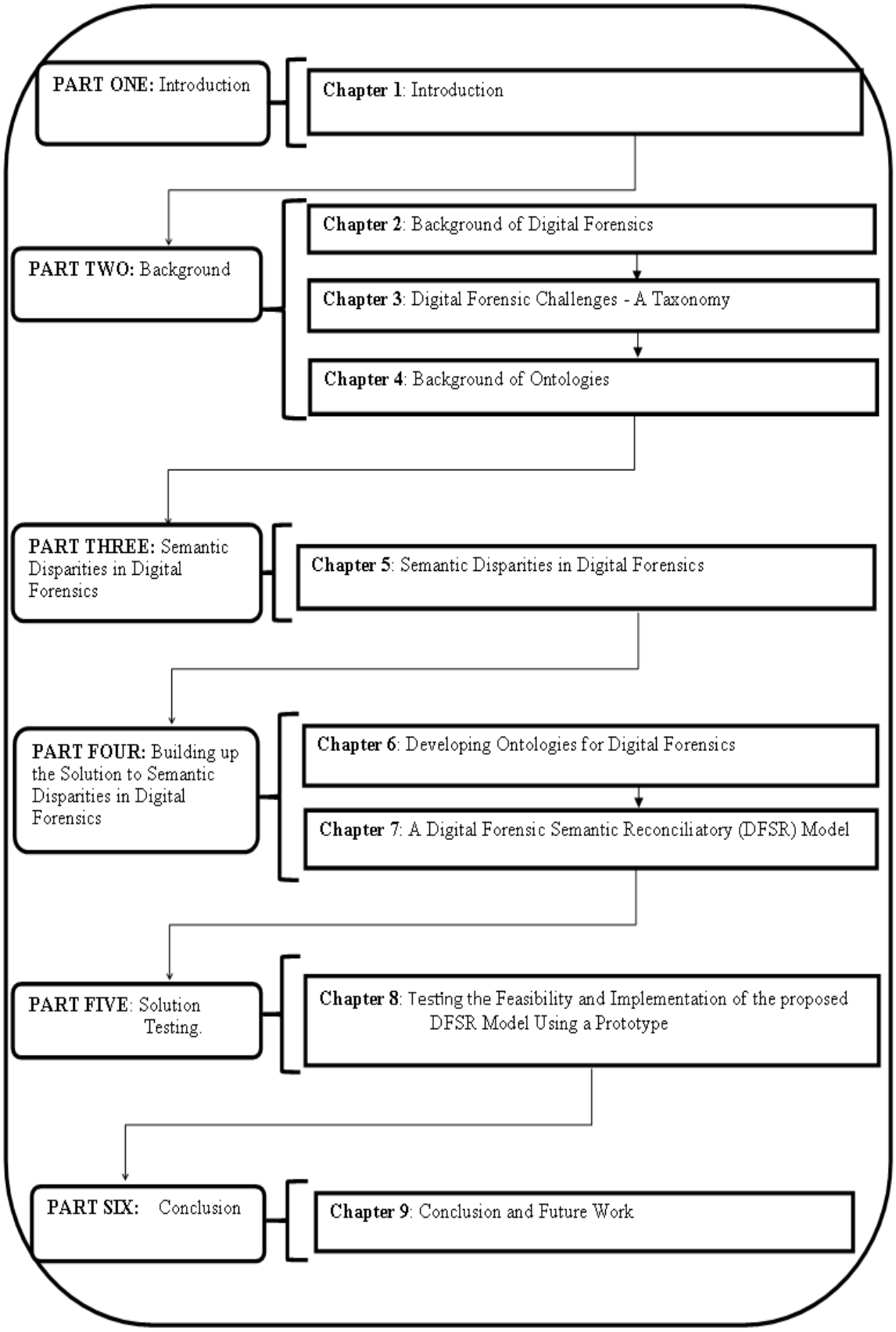


Figure 1.1 Thesis Layout

Chapters 2, 3 and 4 make up Part Two of this thesis and cover the background to the study. In Chapter 2, for example, the reader is provided with a comprehensive background of digital forensics. Nonetheless, special reference is also made here to the main focus of the study, namely the semantic disparities that occur in the digital forensic domain.

In line with the problem statement, various challenges faced by digital forensics (among others, the lack of a unified formal representation of domain knowledge and the lack of standardised methods specifically designed to deal with semantic disparities in digital forensics) are discussed in Chapter 3. A taxonomy of the various digital forensic challenges is then proposed in Chapter 3 based on a comprehensive digital forensic literature survey. Note that the term taxonomy as used in this thesis is coined from Adam, (2015) who defined taxonomy as the practice and science of classifying things according to shared qualities or concepts, including the fundamental truth that underlie such classification. The taxonomy in this thesis therefore classifies the digital forensic challenges according to shared qualities and the fundamental truth that underlie this classification. The taxonomy then summarises the large number of digital forensic challenges into a few well-defined and easily understood categories (of which semantic disparities is among the identified challenges).

Chapter 4 presents some background details of ontologies. Despite the widespread ontology-related research activities and applications in different disciplines, the development of ontologies and ontology research activities is still wanting in digital forensics. Thus, in order to help establish a unified formal representation of the digital forensic domain knowledge and information, the background of ontologies is presented in Chapter 4. This is done as a way to introduce the reader to the fundamental concepts of ontologies, as well as to present ontologies as computational design that permit some kind of reasoning and management of domain knowledge and information.

Chapter 5 explains semantic disparities in digital forensics and make up Part Three of this thesis. This chapter also start the main contribution of the study. Although from the reader's point of view they might look like a literature review they are not. In order to manage the identified semantic disparities in digital forensics, Chapter 5 begins by discussing semantic disparities in the digital forensic domain and further explains the potential causes of semantic disparities with specific examples given where applicable. In addition chapter 5 presents a discussion on how to manage semantic disparities in the

digital forensic domain and elaborates on the different approaches identified to help manage semantic disparities. This chapter also includes the significance of resolving semantic disparities for computer professionals, law enforcement agencies and other digital forensic practitioners.

One of the specific objectives of this research is to establish and propose methods and specifications for resolving semantic disparities in digital forensics. For this reason, Chapters 6 and 7 in Part four of this study focuses on explaining such specifications and methods. Chapter 6 for example, addresses the problem as stated in the problem statement by developing ontologies for digital forensics. The ontologies developed in this chapter are part of the specifications that are needed for creating a unified formal representation of digital forensic domain knowledge and information. Chapter 6 further presents the case for establishing an ontology for the digital forensic disciplines, as well as an ontology for a cloud forensic environment. Such ontologies would enable the better categorisation and representation of knowledge and information, while also helping with the development of future methods and specifications that can offer direction in different areas of digital forensics.

The ontologies presented in Chapter 6 can also be used to better organise digital forensic domain knowledge and explicitly describe the discipline's semantics in a common way. For example, the digital forensic disciplines ontology in Chapter 6 depicts the various distinctions in the different digital forensic disciplines and sub-disciplines identified. The ontology further explains in detail the terminologies used to describe the individual disciplines and sub-disciplines, with specific reference to addressing the problem statement in this research study.

A model coined as the Digital Forensic Semantic Reconciliation (DFSR) model, intended to help resolve semantic disparities in digital forensics, is proposed and explained in Chapter 7 of this thesis. The model is also meant to provide direction towards resolving the semantic disparities that occur in the domain. Such a model can inter alia be used to develop new techniques for detecting and managing semantic disparities in digital forensics.

To assess the feasibility and implementation of the DFSR model, a prototype known as the DFSR prototype is developed and discussed in Chapter 8. Note that Chapter 8 constitutes Part Five of this thesis. The DFSR prototype elaborates on the extent to which the DFSR model discussed in Chapter 7 is developed and implemented.

In addition, Chapter 8 explains how the DFSR prototype can be used as a quick guide towards resolving semantic disparities in the digital forensic domain. A new method coined as the Digital Forensic Absolute Semantic Similarity Value (DFASSV) (Karie and Venter, 2012) is also introduced and explained in Chapter 8. DFASSV is meant to assist in computing the semantic similarity between different digital forensic terminologies. The experiments conducted to test the implementation and accuracy of the DFSR prototype is also explained in Chapter 8. However, the experimental results are based on the individual methods used to develop the DFSR prototype in this study, which include the experiments and the results based on the DFASSV method.

The last chapter (Chapter 9) make up Part Six and conclude this thesis with an explanation of the extent to which the research problem has been addressed as well as the accomplishments. Chapter 9 also points out possible areas for future research based on the current study. A list of the resources consulted in the course of this research is then given, followed by Appendices.

1.6 CHAPTER CONCLUSION

In this chapter the researcher introduced and explained the primary area of focus in this study, as well as identified the problem statement and research questions. As is evident from this chapter there is a need to develop methodologies and specification to address the semantic disparity problem in digital forensics. The motivation for the study and the objectives were also highlighted as a way to show the necessity for this study. The layout of the thesis is explained with the help of Figure 1.1. The whole of Chapter 1 is generally meant to show the scope of the study covered in this thesis.

The next chapter (Chapter 2) presents the background of digital forensics. Chapter 2 is also meant to introduce the reader to some of the basic facts about digital forensics, as well as the different digital investigation process models.

CHAPTER 2 : BACKGROUND OF DIGITAL FORENSICS

2.1 INTRODUCTION

In our daily lives, studies, offices, libraries (or wherever we are using our computers or digital devices), it may seem that we are alone with no one monitoring us. However, every document we create using computers or digital devices, and every step we take on the internet leave behind some digital trace. Digital traces may include deleted files and registry entries, internet history cache and automatic application software backup files (Kessler, 2005).

E-mail headers and instant messaging logs can also be used to give clues as to the intermediate servers through which information traversed. Server logs, on the other hand, can provide information about every computer system accessing a web site (Kessler, 2005). This also implies that every action taken using our computers or digital devices has a number of implications, both advantageous and detrimental. As a result, whatever we do using our computers or digital systems becomes the subject of digital forensic investigations.

The increasing use of the internet – which represents the fastest growing technology tools used by criminals (Kessler, 2005) – has rendered digital forensics inevitable. Figure 2.1, for example, from the internet world starts shows the growing internet users in the world by regions. These statistics makes the internet such a rich field for all manner of criminal activities.

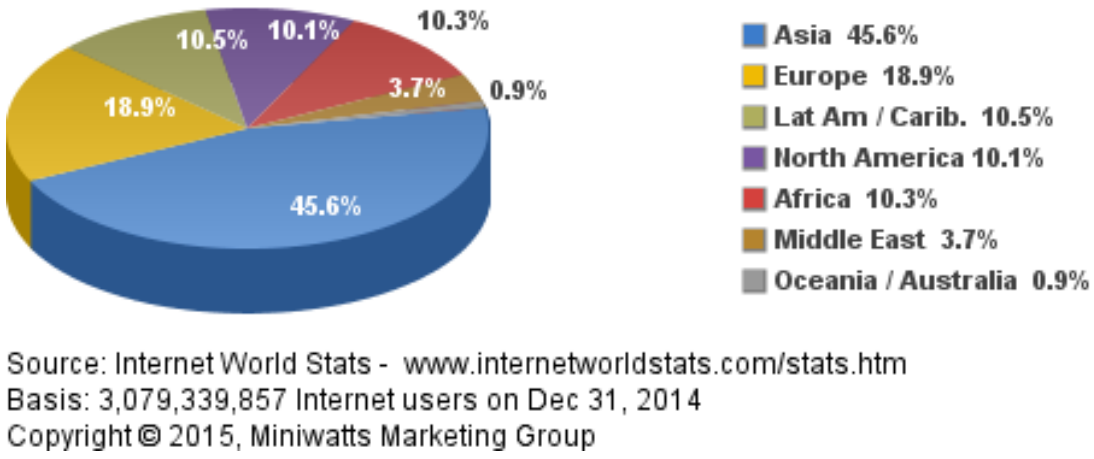
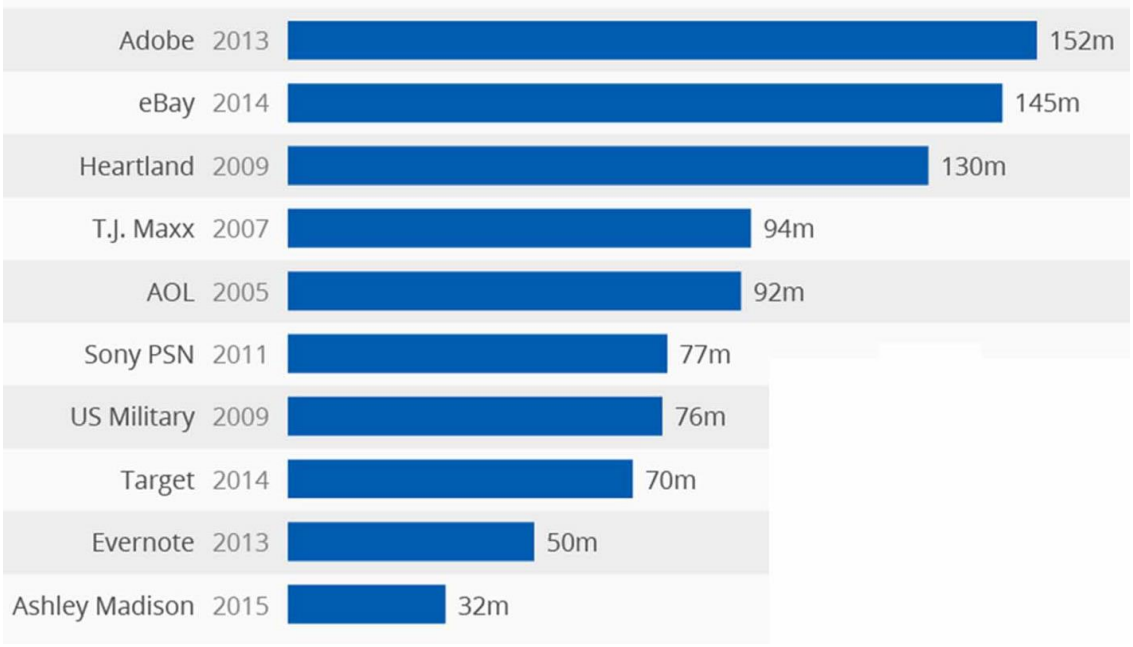


Figure 2.1 Internet users in the World Distributed by World Regions – 2014 Q4

The growing use of the internet has also changed the crime scene investigation and is becoming an increasingly valuable source of digital forensic evidence. A site on the internet used to perpetrate a crime today may for instance be different or absent tomorrow.

Due to this dynamic nature of the internet, the importance of digital forensics for the law enforcement community is growing daily. The situation is aggravated by the fact that digital crime techniques are also becoming more predominant, sophisticated and better coordinated by day. According to Fei (2007) and CERT (2009), the number of incidents involving digital systems keeps rising each year. Cummings (2008) agrees and adds that digital crimes are here to stay and increasing rapidly. This is evident from Figure 2.2 showing the number of compromised data records in selected data breaches as of August 2015.



Amount of Stole Records

Source: http://www.statista.com/statistics/290525/cyber-crime-biggest-online-data-breaches-worldwide/

Figure 2.2 Number of Compromised Data Records in Selected Data Breaches as of August 2015

As a way to restrain the growth of digital crimes, digital forensics is now evolving as the discipline that tries to get answers for the when, what, who, where, how and why questions regarding digital crimes committed around the world (Beebe and Clark, 2005a). For instance, in the case of a digital forensic investigation process, the '*when*' is

used to refer to the time interval of all the actions that took place; the '*what*' is concerned with the actual actions that were performed on the digital system; the '*who*' is concerned with the person or process responsible for the actions; the '*where*' refers to the place where the evidence is located in the digital system; the '*how*' addresses the manner and sequence in which the actions were performed in the system; and the '*why*' seeks to understand the motives behind any unauthorised actions shown to be unruly to planned activities (Beebe and Clark, 2005a).

Chapter 2 is therefore dedicated to introducing the background concepts of digital forensics. Section 2.2 provides an overview of digital forensics while Section 2.3 deals with the digital forensic investigation process models. However, this chapter is also meant to present the reader with some of the disparities that currently exist even in the interpretation, description and representation of digital forensics. This is backed up by the fact that, even at the time of conducting this research, there was no single (common) accepted definition or interpretation of digital forensics. Finally, a chapter conclusion is provided in Section 2.4.

2.2 DIGITAL FORENSICS

Although digital forensics has become prevalent among computer professionals and law enforcement agencies, it is still considered as a somewhat new field of forensic science adopted as a synonym for computer forensics. According to Mark et al. (2002), digital forensics has gained increased popularity in the society since its inception. This is because law enforcement recognises that today's life embraces different types of digital devices that can be exploited for criminal interests, not just computer systems.

While computer forensics tends to pay particular attention on specific techniques for obtaining digital evidence from particular platforms, digital forensics on the contrary must be modelled in such a way that it can shelter a variety of digital devices and systems, including any future digital technologies (Mark et al., 2002). The definition of digital forensics has therefore expanded with time to include the forensics of all digital devices, whereas computer forensics remains "a collection of methods and forensic tools used to search for digital evidence in computer systems" (Caloyannides, 2002).

Unfortunately, at the time of this study, disparities existed even in the definition of digital forensics. Currently there isn't a single (common) and accepted definition of digital forensics. A number of different definitions from different individuals and research organisations exist in literature.

The Digital Forensics Research Workshop (DFRWS) (Palmer, 2001), for example, defined digital forensics as follows:

> "The use of scientifically derived and proven methods toward the preservation, 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).

This definition of digital forensics from DFRWS is a comprehensive one and captures a lot; however, it does not make mention, in its definition, the digital evidence transportation bit of the digital forensics process. Other definitions of digital forensics by different researchers and research organisations are given in the bulleted list that follows:

- According to Carrier, (2008); Reith et al., (2002) and Alazab et al., (2009) digital forensics is defined as "the science of identifying, extracting, analysing and presenting of digital evidence that has been stored in the digital electronic storage devices to be used in a court of law".
- Mohay (2005), however states that, "digital forensics is concerned with the investigation of any suspected crime or misbehaviour that may be manifested by digital evidence". The digital evidence may be evident in various forms, such as digital electronic devices or computers that are simply passive repositories of evidence that document activity, or it may consist of information or meta-information resident on the devices or computers that have been used to actually facilitate the activity or that have been targeted by the activity.
- Van den Bos and Van der Storm (2011) defined digital forensics as "the branch of forensic science where information stored on digital devices is recovered and analysed to answer legal questions".
- According to Gladyshev (2004), "digital forensics is concerned with the use of digital information (produced, stored and transmitted using digital systems) as the source of evidence after a computer security incidence has occurred and also in legal proceedings".

From the above definition it is clear that there are disparities in the interpretation and description of digital forensics. There exist no similarities in the definition of digital forensics except for the fact that the end result of digital forensics is to help legal matters. The phases involved were only captured by the DFRWS (Palmer, 2001) as well as Gladyshev (2004). This situation adds to the disparities in the interpretation and description of digital forensics as a domain

Knowing that, DF is also considered a synonym for computer forensics as mentioned earlier, different researchers have also defined computer forensics as follows:

- Francia (2006) defines computer forensics as "the identification, preservation, and analysis of information stored, transmitted or produced by a computer system or computer network. Its main purpose is to establish the validity of the hypotheses used in an attempt to explain the circumstances or cause of an activity under investigation".
- Yang et al. (2007) defined computer forensic science as "the science of acquiring, preserving, presenting data and analysing information collected on networks".
- According to Abrams and Weis (2003) computer forensics is "the science of obtaining, preserving and documenting evidence from digital electronic storage devices, such as computers, PDAs, digital cameras, mobile phones, and various memory storage devices. All must be done in a manner designed to preserve the probative value of the evidence and to assure its admissibility in legal proceedings".
- Kortsarts and Harver (2007) describe computer forensics as "the scientific examination and analysis of data held on or retrieved from computer storage media in such a way that the information can be used as evidence in a court of law".

Considering all the selected interpretations and descriptions of digital forensics and computer forensics respectively from the above list, it is clear that there is currently no single commonly accepted definition for digital forensics as well as computer forensics. Besides, it is also evident that, from the different definitions used to describe or explain digital forensics and computer forensics, some are in conflict with one another in their intended interpretation and description. Most of the definitions presented above do not specify the steps involved in the investigation process while others do. This creates a

disparity on how to interpret digital forensics. However, the only similarity that cuts across the definitions between computer forensics and digital forensics is in the acquisition and preserving of evidence for use in legal proceedings.

This lack of a unified, common definition for digital forensics and computer forensics has also contributed to the disparities currently being experienced in the domain. Another good example of these disparities is also evident in the different digital forensic investigation process models and the investigation phases involved that have been proposed so far. Standardisation of a digital forensic investigation process model is therefore not an option if the currently experienced disparities and/or any other future disparities are to be resolved. The next section explains the different digital forensic investigation process models that are available.

2.3 DIGITAL FORENSIC INVESTIGATION PROCESS MODELS

Computer security breaches date back to the early 1970s when a group of students found out how to obtain unauthorised entry to large time-shared computers (Noblett et al., 2000). In 1978, the Florida Computer Crime Act was created in USA, making it one of the first laws to help deal with computer fraud and intrusions. In the late 1980s and early 1990s the law enforcement agencies in the United States began working together to restrict the thriving of electronic crimes (Noblett et al., 2000).

In spite of these efforts, the evolution in digital technology caused communication technology that had previously been unavailable to now become easily accessible and to be used to coordinate digital crimes around the world. In fact, digital devices such as computers, mobile phones and other digital systems have become part of modern society and a favourite target of digital crime.

Since many of the daily personal and business transactions are being managed by means of digital systems, a technologically developed world without digital forensics can be devastating, which emphasises the need for standardising digital forensic investigation processes. Standardisation in digital forensics is a positive step towards resolving the disparities that are now experienced in the digital forensic domain. Standardisation will also assist in creating uniformity in the interpretation, description and representation of digital evidence data or information after conducting a digital forensic investigation. Note that, an international standard for digital forensic investigation process was published first in March 2015 (ISO/IEC 27043, 2015).

According to Carrier (2006a), a digital forensic investigation process is one special case of an investigation where the procedures and methods used will allow the outcome to stand up in any court of law. However, to convince the court that the digital evidence presented is worthy of inclusion into the criminal process, the methods and procedures used during investigation must possess scientific validity grounded in scientific methods and procedures (Karie and Venter, 2013b). This implies that the investigation processes should be compatible with the relevant policies and/or laws in various jurisdictions, since evidence may not be admissible in court if it was not properly or legally acquired.

In a bid to help collect potential digital evidence in a forensically sound manner, numerous models, frameworks and methodologies have been proposed to help gather or specify different phases in the digital forensic investigation process (Perumal, 2009). However, according to Kohn et al. (2006), this vast number of proposed models and frameworks has added to the complexity of the field, which further augments the current disparities in the digital forensic domain. This situation led to a call for the standardisation of the digital forensic investigation process (ISO/IEC 27043, 2015).

The ISO/IEC 27043 International Standard has made available recommendations that harmonize idealised models for prevalent investigation processes across different investigation settings. These recommendations include processes beginning with digital forensic readiness up to closure of the investigation, as well as general advice and limitations on processes and appropriate identification, collection, acquisition, preservation, analysis, interpretation and presentation of evidence (ISO/IEC 27043, 2015). In addition, the ISO/IEC 27043 International Standard is intended to complement other standards and documents that provide guidance on the preparations for and actual investigation of information security incidents.

Recent developments in digital forensics have also stressed on the demand for new digital forensic methods and tools that will facilitate successfully investigation of anti-forensics techniques (Alharbi et al., 2011). Table 2.1 below shows a list of some of the proposed digital forensic investigation process models and frameworks currently available. Note that some of the models and frameworks were extracted from Perumal (2009).

Table 2.1 Digital Forensic Investigation Process Models (Source: Perumal, 2009***)***

Model / Framework Name	Author Names	Year	Phases/ processes
1. NIJ Electronic Crime Scene Model	Carrier and Ashcroft	2008	5
2. The DFRWS Investigation Model (Generic Investigation Process)	Palmer	2001	7
3. Abstract Model of Digital Forensic Procedures	Reith, Carr and Gunsh	2002	9
4. Casey's Digital Forensic Framework	Casey	2004	4
5. Computer Forensic Process	M. Pollitt	1995	4
6. An Integrated Digital Investigation Process	Carrier and Spafford	2003	17
7. End-to-End Digital Investigation	Stephenson	2003	9
8. Enhanced Integrated Digital Investigation Process	Baryamureeba and Tushabe	2004	21
9. Extended Model of Cyber Crime Investigation	Ciardhuain	2004	13
10. Hierarchical, Objective-Based Framework	Beebe and Clark	2004	6
11. Event-Based Digital Forensic Investigation Framework	Carrier and Spafford	2004	16
12. Forensic Process	Kent, Chevalier, Grance and Dang	2006	4
13. Investigation Framework	Kohn , Eloff and Oliver	2006	3
14. Computer Forensic Field Triage Process Model	Rogers, Goldman, Mislan, Wedge and Debrota	2006	4
15. Investigation Process Model	Freiling and Schwittay	2007	4
16. The Seamus O Ciardhuáin Extended Model of Cybercrime Investigation	Séamus Ó Ciardhuáin	2004	13
17. Carrier and Spafford's framework	Carrier and Spafford	2003	17
18. The Scientific Crime Scene Investigation Model	Lee et al.	2001	5
19. The Kruse and Heiser Digital Forensic Investigation Model	Kruse and Heiser	2002	3
20. The Harmonised Digital Forensic Investigation Process Model	Valjarevic and Venter	2012	12
21. Integrated digital forensic process model	M.D. Kohn, M.M. Eloff and J.H.P. Eloff	2013	6
22. An Analytical Crime Scene Procedure Model (ACSPM)	Bulbul, Yavuzcan and Ozel, (2013).	2013	10

Based on Table 2.1, it is evident that there were several clear disparities among the different investigation process models at the time of writing this research thesis. The

disparities as captured in the ISO/IEC 27043 International Standard include the following: different numbers of investigation processes; different scope of process models; different scope of the processes with the same names within different process models; different hierarchy levels; and even different concepts applied to the construction of the process models. Note also that, as stated in the ISO/IEC 27043 International Standard, most of the proposed models prior to harmonisation and standardisation were based on physical crime investigation processes (ISO/IEC 27043, 2015).

The digital investigation processes described in the International Standard were, therefore, purposely designed at an abstract level. This means the processes can be used for dissimilar investigations as well as varying types of digital evidence. This methodology was adapted to aid the design and development of high-level processes with the intent to subsequently decompose them into atomic processes (ISO/IEC 27043, 2015). The processes in the standard, aim to be comprehensive in that they represent a harmonization of all previously published digital investigation processes. The investigation processes are also organized in a succinct fashion and describes how to follow these processes.

In order to abstract digital investigation processes at a higher level in the International Standard, they were categorized into the following digital investigation process classes (ISO/IEC 27043, 2015):

- **Readiness Processes**

This class contains processes that deal with pre-incident investigation processes. It is also meant to deal with defining strategies which can be employed to ensure systems are in place, and that the staff involved in the investigative process are proficiently trained prior to dealing with an incident occurring. The readiness processes are non-compulsory to the rest of the investigation processes. This is explained in the published standard (ISO/IEC 27043, 2015) and includes the following:

- Scenario definition;
- Identification of potential digital evidence sources;
- Planning pre-incident gathering;
- Storage and handling of data representing potential digital evidence;

- Planning pre-incident analysis of data representing potential digital evidence;
- Planning incident detection;
- Defining system architecture;
- Implementing system architecture;
- Implementing pre-incident gathering, storage, and handling of data representing potential digital evidence;
- Implementing pre-incident analysis of data representing potential digital evidence;
- Implementing incident detection;
- Assessment of implementation;
- Implementation of assessment results.

- **Initialization Processes**

This is the class of processes dealing with the initial commencement of the digital investigation (ISO/IEC 27043, 2015). The processes involved include:

- Incident detection;
- First response;
- Planning;
- Preparation.

- **Acquisitive Processes**

This class of processes deals with the physical investigation of a case where potential digital evidence is identified and handled (ISO/IEC 27043, 2015) and includes:

- Potential digital evidence identification;
- Potential digital evidence acquisition;
- Potential digital evidence transportation;
- Potential digital evidence storage.

- **Investigative Processes**

This class of processes deals with uncovering the potential digital evidence (ISO/IEC 27043, 2015) and includes:

- Potential digital evidence examination and analysis;

- Digital evidence interpretation;
- Reporting;
- Presentation;
- Investigation closure.

- **Concurrent Processes**

This class of processes continues concurrently alongside the other processes. It differs from the previous classes in the sense that they happen in tandem with the other processes instead of linear. In addition, the particular orders in which the concurrent processes execute is irrelevant as opposed to the other non-concurrent processes (ISO/IEC 27043, 2015) and includes:

- Obtaining authorization;
- Documentation;
- Managing information flow;
- Preserving chain of custody;
- Preserving digital evidence;
- Interaction with the physical investigation.

Figure 2.3 shows the relationships between the various classes of digital investigation processes. Note that the dotted lines in the figure indicate that the particular process is optional. For a detailed explanation of all the processes the reader is advised to refer to the International Standard (ISO/IEC 27043, 2015).

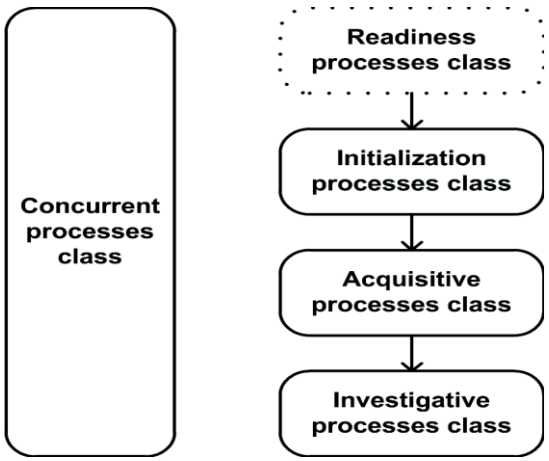


Figure 2.3 Various Classes of Digital Investigation Processes (ISO/IEC 27043, 2015***)***

From the above explanation, it is evident that, standardisation in digital forensics can help to resolve the various disparities that occur during the interpretation, description and representation of any acquired digital forensic data or information. This is so because a standardised investigation process model will ensure uniformity across all scenarios of digital forensic investigations. This will further create an intuitive uniformity in the digital investigation process, irrespective of the investigators involved. Some of the existing ISO/IEC standards about digital forensics are briefly described in Table 2.2 below.

Table 2.2 Existing Digital Forensics Standards

ISO/IEC Standard	Description
1. ISO/IEC 27037:2012	The standard provides detailed guidance on the identification, collection and/or acquisition, marking, storage, transport and preservation of electronic evidence, particularly to maintain its integrity. It defines and describes the processes through which evidence is recognized and identified, recording all the information that serves as evidence at the crime scene, gathering and preserving of the digital evidence, and the packaging and shipping of evidence.
2. ISO/IEC 27041:2015	This standard "provides guidance on mechanisms for ensuring that methods and processes used in the investigation of Information Security Incidents are 'fit for purpose'. It encapsulates best practice on defining requirements, describing methods and providing evidence that implementations of methods can be shown to satisfy requirements. It includes consideration of how vendor and third-party testing can be used to assist this assurance process.
3. ISO/IEC 27042:2015	The standard offers guidance on the process of analysing and interpreting digital evidence, which is of course just a part of the forensics process. It lays out a generic framework encapsulating good practices in this area. The standard emphasizes the integrity of the analytical and interpretational processes such that different investigators working on the same digital evidence ought to come up with essentially the same results - or at least any differences should be traceable to choices they made along the way.
4. ISO/IEC 27043:2015	The standard covers the fundamental concepts behind, and the digital forensic processes involved in, investigating incidents. The standard "makes available recommendations that summarize idealized models for prevalent incident investigation processes across different incident investigation settings involving digital evidence. This includes processes

	that begin with pre-incident preparation and including returning evidence for storage or circulation as well as any general advice and limitations on such processes.
5. ISO/IEC 27050	This 4-part standard concerns the discovery phase, specifically the discovery of Electronically Stored Information (ESI), a legal term-of-art meaning (in essence) forensic evidence in the form of computer data.

Source: http://www.iso27001security.com/html/27037.html [Accessed September 15, 2015].

Among the most commonly used investigation process models in digital forensics prior to the development of the international standard (ISO/IEC 27043, 2015) as shown in Table 2.1 above are:

- The NIJ Electronic Crime Scene Model
- The DFRWS Investigation Model
- The Abstract Model of Digital Forensic Procedures by (Reith, Carr and Gunsch)
- Casey's Digital Forensic Framework

These digital forensic investigation process models are expounded in the sub-sections to follow.

2.3.1 The NIJ Electronic Crime Scene Model

In its digital crime scene investigation blueprint, the National Institute of Justice (NIJ) categorises the digital forensic investigation process into five main phases: preparation, collection, examination, analysis and reporting (Carrier, 2006b, p. 7; Ashcroft, 2001). The different investigation phases as shown in Figure 2.4 are explained in the bulleted list below.

Figure 2.4 NIJ Electronic Crime Scene Model (Carrier, 2006, p. 7; Ashcroft, 2001***)***

- **Preparation**: Preparations should be made to acquire the equipment required to collect any potential digital evidence during the course of an investigation.
- **Collection**: Prior to the collection of evidence, it is assumed that locating and documenting have been done. Potential digital evidence must be gathered carefully so as to preserves its probity.

- **Examination**: In this phase, investigators examine the data that has been acquired using accepted forensic procedures. Wherever possible, the examination should not be conducted on original acquired evidence.
- **Analysis**: Evidence from the examination phase is analysed in this phase to determine its significance and probative value. Analysis activities include, transformation of large amounts of data into suitable analysis sizes; surveying of data to identify obvious digital evidence; and the use of data extraction techniques to mention a few.
- **Reporting**: In this phase, all documentation and the resulting records should be written and presented to the intended stakeholders.

These phases are undertaken in the correct sequential order during digital investigations so as to achieve the correct and desired result. Moreover, these phases are applicable in most of the different disciplines of digital forensics including device forensics and network forensics, software forensics, database forensics and computer forensics. According to Casey (2009), the phases are universally recognised by practitioners in digital forensics because they have proved to be successful when used during digital investigations. However, the NIJ Electronic Crime Scene Model, as observed by the researcher does not mention how evidence is to be transported from a crime scene to the point of examination and analysis which is a critical phase during an investigation process.

2.3.2 The DFRWS Investigation Model

The first Digital Forensics Research Workshop (DFRWS) produced an investigation model (Palmer, 2001) which comprises the following steps: identification, preservation, collection, examination, analysis, presentation and decisions shown in Figure 2.5. The DFRWS model is not meant to be final, but rather a foundation for future work which is to come up with a full investigation model for future research studies (Ciardhuáin, 2004). Although the DFRWS model is presented as sequential, it is possible to give feedback from one particular step to a previous step. Note also in this model that, the transportation phase is not included. Besides, new phases like the identification and decision are introduced in the DFRWS model which was not there in NIJ Electronic Crime Scene Model. This situation in digital forensics confirms the disparities that currently exist in the different proposed investigation process models.

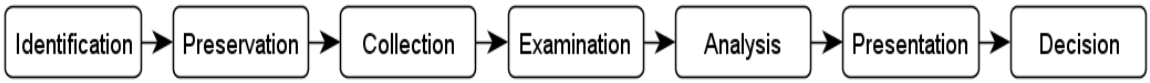


Figure 2.5 Digital Forensic Research Workshop Model (Palmer, 2001***).***

2.3.3 Abstract Model of Digital Forensic Procedures (Reith, Carr and Gunsch, 2002)

Reith, Carr and Gunsch (2002) put forward an investigation model for the digital forensic investigation process which is to some degree very similar to the DFRWS model. The model has nine phases, namely identification, preparation, approach strategy, preservation, collection, examination, analysis, presentation and returning evidence. Their model also supports iterations of individual activity classes. Each of the different phases of the Reith, Carr and Gunsch (2002) model is briefly explained in the bulleted list below.

- **Identification:** This phase recognises a criminal event from some symptoms and decides its type. This is not precisely and clearly expressed within the domain of digital forensics, but it is very important because it has an impact on the other steps.
- **Preparation:** This phase includes the preparing of tools, methods, search warrants, monitoring of authorisations and extensive support from management.
- **Approach strategy:** This phase involves the dynamic formulation of a strategy based on potential impact on onlookers and the specific technology in question. The primary aim of the strategy should be to maximise the acquisition of polluted evidence while minimising the impact on the victim.
- **Preservation:** This phase involves isolating, securing and preserving both the physical and potential digital evidence. This is done by prohibiting people from using the digital device in question or allowing other electromagnetic devices to be used within an affected radius.
- **Collection:** This involves recording the crime scene as well as duplicating the digital evidence by using standardised and accepted methods.
- **Examination:** This comprises a comprehensive and thorough systematic search of digital evidence relating to the crime in question. The examination phase focal points is usually on pinpointing and unearth potential evidence, possibly in

uncommon locations, as well as on constructing a detailed documentation for analysis.

- **Analysis:** This refers to the determination of the significance of evidence, reconstruction of snippet of data and coming up with conclusions, based on the evidence collected. It may sometime take a number of iterations of examination and analysis to endorse a crime belief.
- **Presentation:** This is the summary and clarification of any conclusions made. The presentation should be written in a layperson's language using abstract terminologies. The abstract terminologies should reference the specific details.
- **Returning Evidence:** This phase involves ensuring that both the physical and digital items are returned to the owners, and determine how as well as what criminal evidence should undergo elimination. This is also not necessarily a digital forensics step; but, most of the existing models that seize evidence rarely address this aspect.

Like the other previous investigation process model, the Reith, Carr and Gunsch (2002) model also introduced new phases like; approach strategy and returning evidence, which were absent in other previous proposed models. This scenario in digital forensics adds up to the number of disparities experienced and thus builds up to motivate this research study.

2.3.4 Casey's Digital Forensic Framework

Casey (2004a) proposed a model that consists of four key steps, namely recognition, preservation, classification and reconstruction. Figure 2.4 shows the digital forensic framework proposed by Casey (2004a).

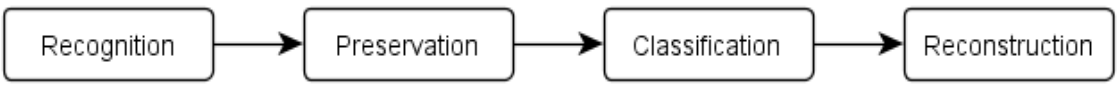


Figure 2.6 Casey's Digital Forensic Framework (Casey, 2004a***)***

The primary focus of Casey model (Casey, 2004a) is on two main areas: first is the forensic process, followed by the investigation itself in the second place. The classification and reconstruction steps which are all new in this model handle the evidence analysis. The model proposed by Casey is basically a general one, and can be

applied to both stand-alone systems as well as networked systems (Ciardhuáin, 2004). This model is also different from the previous proposed model hence adding up to the disparity challenge in digital forensics.

2.3.5 The Integrated Digital Forensic Process Model (IDFPM)

Besides the different investigation process models and frameworks discussed above, Kohn et al., (2013) also presented an Integrated Digital Forensic Process Model (IDFPM). The IDFPM consists of the following processes: preparation, incident, incident response, physical investigation, digital forensic investigation and presentation. As is common with all previous proposed models, some of the process introduced in the IDFPM model are also new (not in the previous models) hence adding up to the disparities in digital forensic investigation process. Each of the different phases of Kohn et al., (2013) model is briefly explained in the bulleted list below.

- **Preparation**

According to Kohn et al., (2013), preparation is the single most critical process in the IDFPM. This is where the organization enables itself to deal effectively with various types of incidents.

- **Incident**

In this phase Kohn et al., (2013) states that the incident scope will have to be determined by the type of investigation conducted. An incident may be detected by an automated incident detection system, or a similar set of event sequences is recognized by an investigator, based on possible previous experience. An investigator must then assess the incident anomaly detected. The detected incident should be confirmed by some other source before action is taken towards an incident response. Once an incident is confirmed, the investigators should be notified to initiate an incident response.

- **Incident Response**

In this phase Kohn et al., (2013) adds that, the first responder is the first custodian to maintain the chain of evidence and custody of potential digital evidence. Therefore, depending on the type of investigation, witnesses need to be safeguarded; suspects need to be detained as soon as possible after arrival and potential evidence must be secured. In addition, the first responder must be able to accurately describe the scene in the initial drafting of documentation; these include photographs, video and sketches (Carrier and Spafford, 2003).

- **Physical Investigation**

In this phase, the physical investigation process occurs in parallel with the digital investigation if the crime is not isolated to the digital space. The focus of the physical investigation is to analyse DNA, fingerprints and other possible physical evidence obtained from the incident scene (Kohn et al., 2013).

- **Digital Forensic Investigation**

At the heart of the IDFPM lies the digital forensic investigation. The processes used in this phase will determine the success of the investigator's findings, which will ultimately be presented in court. The digital evidence and investigator findings are finally communicated to the relevant interested parties which in most instances this will be the authority that authorized the investigation (Kohn et al., 2013).

- **Presentation**

The presentation phase occurs when the hypothesis is presented to people other than the investigators, such as a jury or management. A decision will then be made based on the findings (Kohn et al., 2013).

2.3.6 An Analytical Crime Scene Procedure Model (ACSPM)

Bulbul, Yavuzcan and Ozel, (2013) also proposed an analytical crime scene procedure model (ACSPM) for digital investigations at a crime scene with main focus on crime scene digital forensic procedures, other than that of whole digital investigation process and phases that ends up in a court. After analysing the needs of law enforcement organizations and realizing the absence of crime scene digital investigation procedure model for crime scene activities the authors decided to inspect the relevant literature in an analytical way. The outcome of their inspection was a model, which is supposed to provide guidance for thorough and secure implementation of digital forensic procedures at a crime scene. Bulbul, Yavuzcan and Ozel, (2013) has the following phases:

i. Managerial activities
ii. Crime Scene Examination
iii. System Assurance
iv. Evidence Search
v. Evidence Acquisition
vi. Hypothesis and Validation
vii. Organisation of Potential Evidence

viii. Physical Management of Evidence

ix. System Service Restoration

x. Provide Chain of Custody (CoC)

According to Bulbul, Yavuzcan and Ozel, (2013) in digital forensic investigations each case is unique and needs special examination, it is not possible to cover every aspect of crime scene digital forensics, but the proposed ACSPM model is supposed to be a general guideline for practitioners. For a comprehensive and thorough clarification on the ACSPM phases, the reader is advised to explore the original manuscript by Bulbul, Yavuzcan and Ozel, (2013).

From the different investigation process models and frameworks discussed in this section, it is evident that lack of uniformity (disparities) in the digital forensic investigation phases is one major challenge faced by digital forensics. Therefore, the authors still contend that, besides the ISO/IEC 27043 International Standard, more standardised methods and specifications need to be developed in digital forensics to help resolve the current disparities in the different investigation process models as well as in the digital forensic domain as a whole.

2.4 CHAPTER CONCLUSION

In this chapter the researcher examined and explained the fundamental concepts of digital forensics. Several selected definitions of digital forensics were also presented. Since other terminologies exist as synonyms of digital forensics, the researcher presented the definitions of the terms ‘digital forensics’ as well as ‘computer forensics’. Several digital forensic investigation process models were also discussed in this chapter. This was done to show the extent to which disparities (lack of uniformity) exist even in the definition of digital forensics, as well as in the proposed investigation process models.

The next chapter (Chapter 3) explores the different challenges faced by digital forensics. Some of these challenges contribute to the identified disparities in digital forensics. For this reason, Chapter 3 is meant to spark discussion about the development of methods and specifications for resolving the identified challenges in digital forensics, even more as a way towards resolving any identified semantic disparities in the domain.

CHAPTER 3 : DIGITAL FORENSIC CHALLENGES – A TAXONOMY

3.1 INTRODUCTION

The evolution in digital technology has greatly influenced the way we conduct our daily lives and our business. The use of computers and other digital devices has grown exponentially to the point where almost one and all have their own personal data device that they carry with them continuously. However, as this evolution in the use of computers and other digital devices continues, numerous challenges emerge that are to be faced by the digital forensic domain.

This chapter therefore aims at reviewing existing digital forensics literature and highlighting the different challenges that digital forensics have encountered to date. This chapter also forms the background section of this research study. A taxonomy of the various digital forensic challenges is, however, proposed as a contribution in this field. Note that the term taxonomy is used in this thesis to mean, the practice and science of classifying things according to shared qualities or concepts, including the fundamental truth that underlie such classification (Adam, 2015). From this definition, therefore, the taxonomy proposed in this chapter classifies the large number of digital forensic challenges into a few well-defined and easily understood categories that cover a large number of digital forensic challenges. In fact, the taxonomy was accepted and published by *the Journal of Forensic Sciences (Vol. 60, No.4, pp.885-893)* after undergoing a peer review process. Note also that Semantic disparity is, however, exclusively selected among the many challenges discussed in this chapter and forms the primary focus of this research study.

The discussion of the taxonomy presented in this chapter can thus be useful in future developments of automated digital forensic tools, as well as in explicitly describing processes and procedures that focus on addressing the individual digital forensic challenges identified in this study. Institutions of higher learning should find the proposed taxonomy in this study constructive, especially when they develop curricula and educational material for different undergraduate courses, as well as research projects for postgraduate studies.

Furthermore, the presentation of the taxonomy in this chapter offers a comprehendible categorisation that may shed more light on existing digital forensic challenges. The taxonomy has been designed in a way to accommodate new categories of digital forensic challenges that may crop up as a result of technological change or domain evolution.

Finally, this chapter is meant to show that semantic disparity is a challenge among many other challenges in digital forensics and, hence, the motivation for this study to resolve it. Chapter 3 is also meant to spark discussion in the development of methodologies and specifications for resolving the other identified challenges in digital forensics. This implies that the contributions in this chapter can be used as a stepping stone towards resolving any other identified disparities in digital forensics.

Section 3.2 of this chapter provides a brief overview of challenges faced by digital forensics, while Section 3.3 explains the scope of the taxonomy proposed in this chapter. The taxonomy of challenges for digital forensics is discussed in Section 3.4 and the chapter is concluded in Section 3.5.

3.2 CHALLENGES FACED BY DIGITAL FORENSICS

Since its establishment over a decade ago the digital forensic domain has encountered several challenges. These include challenges such as the vast volumes of data (Kara et al., 2009), education and certification, lack of unified formal representation of domain knowledge, legal system challenges, semantic disparities, etc.

Despite different stakeholders having examined and analysed several existing digital forensic challenges, there is still a need for a formal classification of such challenges. This section of the chapter therefore evaluates existing digital forensic literature and points out the different challenges that digital forensics has encountered over the past decade or so. A taxonomy of challenges faced by digital forensics is then proposed and explained. The taxonomy feeds into this research study by highlighting semantic disparities as a challenge in the field of digital forensics, thus, forming the primary focus of this study, which is to resolve semantic disparities in digital forensics.

Note also that some of the challenges currently experienced in the digital forensic domain are as a result of unresolved disparities and the lack of standardised methods and procedures in the domain. For example, after an investigation process has been conducted, based on a particular investigation process model, there may still remain disparities in the evidence interpretation, description and representation of the data or

information. As an example to support this study, the researcher considered a court case between Smith vs. Groover, 468 F.Supp. 105 (N.D.Ill.1979) in the United States (Smith vs. Groover, 1979). Although this case was not purely based on digital crimes, it is however in line with the problem of semantic disparity addressed in this thesis. After an investigation was conducted, the district court noted an important semantic disparity over the meaning of the terminology "*regulatory umbrella*". The court, noting its proximity in the Committee report to the admitted concerns over private actions against markets, understood the term "*regulatory umbrella*" to mean that the Congress was replacing private actions with a powerful, pervasive new regulatory agency that, unlike those in the past, had all of the tools required to enforce the exchanges' obligations. The district court, on the other hand, felt that the term "*regulatory umbrella*" was meant to signify the new agency's ability to impose duties on the markets notwithstanding the resultant legal exposure for the exchanges in private suits.

Considering the Smith vs. Groover (1979) court case, this problem is not unique to digital forensics, hence, developing practical methods that can aid in resolving the different challenges and disparities in digital forensics is inevitable and as important as the research itself. For digital forensics to remain effective and relevant to the law enforcement agencies, the academic field as well as the private sectors, the domain experts must constantly endeavour to address existing challenges and disparities in the domain. The scope of the taxonomy proposed in this thesis is explained first in the section to follow.

3.3 SCOPE OF THE PROPOSED TAXONOMY

There are many different challenges in digital forensics. In addition, several attempts to address specific or individual challenges in the domain were made by different researchers in the past. However, the presentation in this chapter is an effort by the researcher to propose a taxonomy of digital forensic challenges, based on the review of existing literature in the field of digital forensics.

The boundaries of the taxonomy are restricted to the extent of the literature set for review by the researcher (not more than ten years old at the time of writing this thesis). The researcher also acknowledges that the various challenges presented in this chapter shown in Table 3.1 do not purport to be an exhaustive list due to the limits set on the literature surveyed. An exhaustive list is in most cases also hard to create and, even if created, it would not be easy to handle or manage because of its size. This also implies

that the bigger the size of the list, the more difficult it becomes to manage it effectively. For this reason the sub-categories of the challenges listed in column two of Table 3.1 were merely selected as common examples to facilitate this study and not to serve as an exhaustive list. More specific sub-categories of the challenges in each named category can and should be added as the need arises in future.

The taxonomy has also been designed taking into consideration only the major challenges that digital forensics has faced over the past decade as identified in the literature surveyed. The researcher did not draw a precise distinction between the old and the most recent digital forensic challenges in this chapter, because some of the challenges captured in the taxonomy are inherent to digital forensics, e.g. the vast volumes of data. Future research will, however, consider the possibility of developing an extensive taxonomy with distinctions between the old and the most recent challenges. The next section explains in detail the proposed taxonomy of challenges for digital forensics.

3.4 THE TAXONOMY OF CHALLENGES FOR DIGITAL FORENSICS

In this section, the researcher presents a detailed explanation of the taxonomy of challenges for digital forensics. Table 3.1 shows the structure of the proposed taxonomy.

The taxonomy consists of four rows arranged from top to bottom with the first row depicting the technical challenges faced by digital forensics. This is followed by the legal systems or law enforcement challenges in the second row, the personnel-related challenges in the third row and finally the operational challenges faced by digital forensics in the fourth row.

However, the various sub-categories of the challenges presented in each of the different rows of the taxonomy shown in Table 3.1 focus more on areas that can be considered when developing for instance new curricula and education materials for different undergraduate programmes as well as research projects for postgraduate studies.

The sub-categories can also be useful when developing dynamic digital forensic tools that focus on addressing specific identified digital forensic challenges. Organising the taxonomy into categories and sub-categories was necessary to simplify the understanding of the taxonomy as well as to present specific finer details of the taxonomy.

Table 3.1 The Taxonomy of Challenges for Digital Forensics

Digital Forensics Challenges	Identified Sub-Categories
1. Technical Challenges	i. Difficulties in conducting cryptanalysis ii. Difficulties in managing vast volumes of data iii. Incompatibility among heterogeneous forensic tools iv. Difficulties in managing volatility of digital evidence v. Bandwidth restrictions vi. Limited lifespan of digital media vii. Sophistication of digital crimes viii. Difficulties in managing emerging technologies ix. Limited window of opportunity to collect potential digital evidence x. Difficulties in managing anti-forensics xi. Difficulties in acquiring information from small-scale technological devices xii. Emerging cloud computing or cloud forensic challenges
2. Legal Systems or Law Enforcement Challenges	i. Difficulties in managing jurisdiction ii. Difficulties in prosecuting digital crimes (legal process) iii. Admissibility of digital forensic tools and techniques iv. Insufficient support for criminal or civil prosecution v. Difficulties in managing ethical issues vi. Difficulties in managing privacy
3. Personnel-related Challenges	i. Lack of qualified digital forensic personnel (training, education and certification) ii. Difficulties in managing semantic disparities in digital forensics iii. Lack of unified formal representation of digital forensic domain knowledge iv. Lack of forensic knowledge reuse among personnel v. Challenges pertaining to forensic investigator licensing requirements
4. Operational Challenges	i. Difficulties in incidence detection, response and prevention ii. Lack of standardised processes and procedures iii. Significant manual intervention and analysis iv. Digital forensic readiness challenges in organisations v. Trust of audit trail challenges

The major categories of the various digital forensics challenges as found in various surveyed literatures (with their details and sub-categories as shown in Table 3.1) include the following: technical challenges; legal systems or law enforcement challenges; personnel-related challenges; and operational challenges. The taxonomy shown in Table 3.1 was developed, based on the literature survey, as a way to show the different challenges faced by digital forensics of which difficulties in managing semantic disparities is one of the challenges listed under Personnel-related challenges (Karie and Venter, 2015). Note also from the literature survey carried out during the time of this study, it was evident that, many of the challenges listed in Table 3.1 have been addressed by different researchers. However, no attempt was made to resolve the semantic disparity problems that occur in digital forensics. For this reason, resolving semantic disparities became the primary focus of the study presented in this research thesis.

In the solution approach, however, there are different ways to resolve the semantic disparities in digital forensics including:

(i) Resolving semantic disparities through the use of ontologies, this is discussed in chapter 5 and 6 of this research thesis.

(ii) The second approach to resolving semantic disparities is by using a sematic reconciliation model which is discussed in chapter 7 and tested in chapter 8 of this research thesis.

The ontology and the model therefore are the two different ways discussed in this thesis which can be used to resolve semantic disparities in Digital forensics. In the sub-sections to follow, the various categories and sub-categories of the challenges faced by digital forensics as identified in Table 3.1 are explained in more detail.

3.4.1 Technical Challenges

Technical challenges can be described as those challenges that can be addressed with existing expertise, protocols and operations. Implementing solutions to address any of the identified technical challenges often falls to someone with the authority to do so. Hence, digital forensics needs a good mixture of both technical skills as well as ethical conduct. Some of the identified technical challenges faced by digital forensics are explained in the sub-sections to follow.

3.4.1.1 Difficulties in Conducting Cryptanalysis

With the advances in communication technologies such as the internet, complex encryption products are now widely and easily accessible, presenting the digital forensic examiner with a significant challenge. Moreover, as the levels of quality of encryption go up and encryption algorithms become even more sophisticated, it will be more complex and time-consuming for individuals to conduct cryptanalysis and then put together encrypted files into useful information (Gallegos, 2005). Cryptanalysis is described as the science of 'code breaking' in which an individual reconstructs the original plaintext message from an encrypted version (Thinkquest, 2013) without having a valid decryption key.

There is currently no proven or fully known direct or standardised formula for conducting cryptanalysis. For this reason, the encrypted data in most cases is out of reach without the decryption key. If the victim denies handing over the key or pleads plausible deniability, the investigators will have to try other techniques to get the decryption key (Lowman, 2013). Although it is now the law in the UK that any encryption key must be given to the police, this is not the case in other jurisdictions, and punishment for not surrendering such keys may be far less severe than the potential punishment for any crime committed (Lowman, 2013).

3.4.1.2 Difficulties in Managing Vast Volumes of Data

There has been tremendous growth in the volume of persistent storage – disk storage – used in both personal and corporate systems (Mohay, 2005). With the incredibly large volumes of data existing within application programs such as the Enterprise Resource Planning (ERP) systems, and as mail systems become larger, the volume and amounts of material being generated are by far not humanly readable in a lifetime – let alone in the scope of a trial or litigation (Libby, 2013). This has implications not only for the procedures and techniques used by digital forensic investigators for data acquisition and imaging, but also (and more importantly) for the way in which digital forensic data is analysed.

3.4.1.3 Incompatibility among Heterogeneous Forensic Analysis Tools

Digital forensic tools generally differ in functionality, complexity and cost. Some tools are designed to serve a single purpose or provide unique information to examiners, while others offer a suite of functions (Arthur and Venter, 2004). All the same, most of the

existing forensic analysis tools consist of dissimilar elements or parts (design and algorithms) and are consequently unable to work together harmoniously. Besides, some of the tools are unable to deal effectively with the ever growing storage size of the target devices. This implies that huge targets pose a challenge as they require more sophisticated analysis techniques that allow digital forensic investigators to perform forensic investigations much more efficiently (Richard and Roussev, 2006), thus facilitating digital investigations.

3.4.1.4 Difficulties in Managing Volatility of Digital Evidence

Digital forensic evidence is, by its nature, delicate and vulnerable. Almost any activity performed on a device, whether inadvertently or intentionally (e.g. powering up or shutting down), can alter or destroy potential evidence (DOJ, 2013). In addition, loss of battery power in portable devices, changes in magnetic fields, exposure to light, extremes in temperature and even rough handling can cause loss of data. Collecting volatile data therefore presents a serious challenge to digital forensic investigators, because doing so can change the condition of any system as well as any contents inside the memory itself.

3.4.1.5 Bandwidth Restrictions

According to Taute et al. (2009), bandwidth restrictions in networks can limit or slow down the digital evidence acquisition process. Since the suspect machine in any network is live and active, digital forensic investigators need to connect to the forensic agent installed on the machine via a network. Copying the data as potential digital evidence from the suspect machine to the forensic workstation might slow down the bandwidth, especially if there are many users utilising the bandwidth at that particular time. Large remote evidence acquisitions may also have to be done after hours to accommodate smaller bandwidth capacities, thus posing a challenge to investigators.

3.4.1.6 Limited Lifespan of Digital Media

While digital storage media facilitate the storage of and easy access to electronic data, they do not provide long-term archival storage (Conserve, 2010). This is because ‘bit preservation’ and the ability to monitor for ‘bit loss’ lie at the core of every digital storage medium, and any bit deterioration can compromise digital data (Reed, 2013). The life span of some digital storage media is typically short and also well enough

known for all to be aware of the risks of using them for preservation purposes (Harvey, 2013). This poses a serious storage challenge. Fortunately, with the emerging cloud computing, the cloud servers' leverage on redundant digital storage media which ensures that in the event of a hardware failure, the data continues to be accessible from another part of the cloud where it is stored safely.

3.4.1.7 Sophistication of Digital Crimes

The ever growing complexity of digital crimes presents crucial challenges to investigations and digital forensic investigators. According to a report by the Association of Chief Police Officers (ACPO, 2013), digital forensic investigators are regularly confronted with the truth of complicated encryption schemes, sophisticated hacking tools and mischievous software that may exist simply within memory. Culprits now use anti-forensic methods that may need a multitude of digital investigations in the case of an incident (Eroraha, 2010), thus making it even harder for investigators to get the much needed evidence.

3.4.1.8 Difficulties in Managing Emerging Technologies

According to Sheward (2013), latest and emerging technologies generate new challenges for digital forensic investigators. Dealing with new file systems, for example, or just a new file type, may need a change in strategy or maybe coming up with new methods. While such changes may need small amendments to clearly stated approaches, it is particularly infrequent to have to handle a technology that gives a sheer transition.

3.4.1.9 Limited Window of Opportunity to Collect Potential Digital Evidence

During the collection of potential digital evidence it is crucial for investigators to prioritise which data must be collected first. This becomes a challenge to investigators especially when there are time constraints and the window of opportunity to collect the data is small (Elancheran, 2013), or when the time to image a system is too short. Investigators must take the necessary steps to ensure that they are able to collect and preserve critical information during this window of opportunity and analyse the data in a way that maintains its integrity.

3.4.1.10 Difficulties in Managing Anti-forensics

According to Garfinkel (2009), anti-forensics (AF) is an ever-increasing group of tools and methods that thwart digital forensic tools, investigations and digital forensic

investigators as well. Different individuals make use of anti-forensics to show how weak and untrustworthy computer data is. In order to present computer evidence in any court of law, the prosecutors must also prove that such evidence is genuine. This also means that the prosecutors should be in a position to substantiate that any information furnished as potential digital evidence in court is in fact came from the culprit's computer system and that it has also remained unchanged. Anti-forensics makes it hard for examiners to detect that some form of incidence occurred and it obstruct evidence gathering, thus escalate the time needed by examiners to spend on a particular case and causing suspicion on the forensic report or deposition (Liu and Brown, 2006).

3.4.1.11 Difficulties in Acquiring Information from Small-scale Technological Devices

According to Bennett (2011), unlike conventional desktop or laptop computer forensics, the process to extract information from small-scale technological devices is much more complicated. With a desktop computer, the investigator simply removes the hard drive, connects it to a write blocker (thereby allowing the collection of evidence without giving rise to the possibility of accidentally causing damage to the contents of the computer drive) and images the hard drive so as to analyse the data fully.

Moreover, with the continued growth of the mobile device market, the potential use of such devices in criminal activity will continue to increase (Yates, 2010). There are currently numerous manufacturers and models of mobile devices on the market, which results in creating a huge diversity of potential problems and challenges to investigators. For this reason, it becomes extremely difficult for an investigator to choose the proper forensics tools for seizing internal data from mobile devices (Yates, 2010).

3.4.1.12 Emerging Cloud Computing or Cloud Forensic Challenges

Cloud computing has emerged as an important solution offering organisations a potentially cost effective model to address their computing needs and accomplish business objectives. However, mixed in with the cloud cost effective opportunities, there are numerous challenges that need to be considered prior to committing to a cloud service such as jurisdiction and cloud heterogeneity (Ferguson, 2013). According to Leslie et al. (2011), other challenges faced by the cloud include safeguarding data security, managing the contractual relationship, dealing with lock-in and managing the cloud. Numerous security challenges such as data protection, user authentication and

data breach contingency planning also need to be addressed. The next section explains the challenges pertaining to legal systems or law enforcement.

3.4.2 Legal Systems or Law Enforcement Challenges

There is a growing knowledge in the legal fraternity about the need for digital forensics to get successful prosecutions in court. Unsatisfactory equipment, procedures or inadequate presentation in court could easily cause forensic investigations to fail (Bassett et al., 2006). Therefore, in the sub-sections to follow, we examine some of the legal systems or law enforcement challenges faced by digital forensics.

3.4.2.1. Difficulties in Managing Jurisdiction

The increasing popularity of cloud computing has rendered conventional crime detection even more difficult. The very strengths of cloud computing, which allows anyone anywhere in the world to use publicly accessible software to process data stored in a virtual cyber-space location, could be put to devious use by criminals to store incriminating data on a server located beyond the jurisdiction of the courts of their country of residence, preferably in a state that has not signed a judicial cooperation treaty with that country (Vaciago, 2012). This makes court jurisdiction a serious challenge during prosecution.

3.4.2.2 Difficulties in Prosecuting Digital Crimes (Legal Process)

According to Lauren (2013), prosecuting cyber-crime is no easy task, due to disparate laws. Even with modern forensic competence, legal deficiency in different jurisdictions (besides inconsistent law enforcement and legal processes) makes prosecution a very challenging venture. This has created the need for new legislation that allows for digital evidence to be presented in any court of law or civil proceedings (Khan, 2010), as well as for the prosecution of digital crimes.

Current digital forensic investigations are based on the existing legal system or the legal processes and supporting statutes present. The basic structures and facilities to investigate digital crimes is based on the current existing cyber laws, which makes it hard to embrace specific digital forensic models to conduct digital investigations and prepare reports that are acceptable in court (Khan, 2010). A large number of digital forensic practitioners simply use available technical methods and do not remember about

the actual motive and most important concepts of digital forensic investigations (Ieong, 2006).

3.4.2.3 Admissibility of Digital Forensic Tools and Techniques

Given the enormous volumes of data currently handled by digital forensic investigators, the admissibility of different digital forensic techniques and tools employed during the acquisition and analysis of data is becoming a challenge. As with all other forensic disciplines, digital forensic techniques and tools must also meet fundamental evidential and scientific standards so as to be acceptable as evidence in civil proceedings and courts (Craiger et al., 2006). This also means that it should be possible to prove through empirical testing that the processes, tools, techniques and procedures are correct. In the context of digital forensics, this implies that the processes, tools, techniques and procedures used in the collection and analysis of digital evidence data must be validated and proven to meet scientific standards. Otherwise, the outcome from such tools will not be admissible as potential evidence in court.

3.4.2.4 Insufficient Support for Criminal or Civil Prosecution

According to Mercuri (2009), digital forensic techniques may at times be applied in a way not conforming to approved standards in an attempt to tip the scales of justice in the direction of prosecution. Burgess (2013) also states that in the digital forensic domain (compared to other fields like law), the methods used in civil cases vary to a moderate extent from those used in criminal cases. The data acquisition and evidence presentation may be held to distinct standards, the data collection and imaging processes can be different as well, and the effects of the case may also have varying impacts.

3.4.2.5 Difficulties in Managing Ethical Issues

Bassett et al. (2006) argue that there are many difficult ethical situations that investigators must be ready to encounter when conducting a digital investigation. One of the most prevalent ethical concerns is how investigators should manage the discovery of information intended to be kept secret that is also not relevant to the case being investigated. The question arises of what to do with such information. The general code of ethics to follow in such a situation is that the information must be disregarded because it is irrelevant to the case at hand. However, it is not always easy to pay no attention to such information and any secrets that may be revealed can weigh heavily on the mind of

the investigator. Other ethical concerns may include acknowledgement of errors by investigators on evidence data; bias during an investigation; maintaining control of and responsibility for forensics equipment (Bassett et al., 2006).

3.4.2.6 Difficulties in Managing Privacy

Privacy issues usually arise in the case of an investigation. Privacy is very important to any organisation or victim. In special cases, however, the investigator may be required to share the data or compromise the client's privacy to get to the truth, provided that the necessary documentation (such as a warrant) has been acquired. It is possible that the victim organisation may lose trust in the forensic team if for instance private information is exposed (Anon, 2013a). In addition, disclosure of any of the client's information to the public by direct or indirect means can be a violation of privacy policies as well as of the ethical code of conduct. Any type of electronic transaction that leads to disclosure of private information can also be considered a violation of privacy policies and the code of behaviour/ethics (Anon, 2013a). Confidential information should at all cost be kept private by the forensic investigator. The next section elaborates on the personnel-related challenges faced by digital forensics.

3.4.3 Personnel-related Challenges

As with any potential forensic evidence, testimony that clearly establishes that the potential digital evidence has been under the control of responsible personnel and well-trained digital forensic investigators is required to assure the court of the fact that the evidence is complete and has not been tampered with in any way (Ryan and Shpantzer, 2005). As mentioned earlier, digital forensics therefore needs a stable mixture of both technical skills and ethical behaviour from all personnel involved. In the sub-sections to follow, some of the identified personnel-related challenges faced by digital forensics are explained in more detail.

3.4.3.1 Lack of Qualified Digital Forensic Personnel (Training, Education and Certification)

According to Desai et al. (2009), digital forensics has become an important field of research because of the increased number of cyber-crime cases. Due to the general shortage of trained digital forensic personnel, there is fierce competition for employing digital forensic specialists in law enforcement. Qualified digital forensic experts are a challenge to find, both in the private and public sector. Even if technically proficient

specialists are available, very few are trained or certified to deliver convincing, scientifically valid and expert witness testimony in a court of law or civil proceedings.

3.4.3.2 Difficulties in Managing Semantic Disparities in Digital forensics

Digital forensics as a growing field is gaining popularity among computer professionals, law enforcement agencies, forensic practitioners and other stakeholders. Unfortunately, their divergent backgrounds have created an environment challenged with semantic disparities (Karie and Venter, 2013), which must be resolved. Besides, cooperation between computer professionals, law enforcement agencies and other forensic practitioners presupposes the reconciliation of any semantic disparities that are bound to occur in the domain, which is also a huge challenge.

3.4.3.3 Lack of Unified Formal Representation of Digital Forensic Domain Knowledge

According to Hoss and Carver (2009), there is (at the time of writing this thesis) no unified formal representation of digital forensic knowledge or standardised procedures for gathering and analysing knowledge. This lack of a unified representation inevitably results in incompatibility among digital forensic analysis tools. Mistakes in the interpretation of potential digital evidence are more expected where there exist no formalised or standardised methods for gathering, preserving and analysing digital evidence (Chaikin, 2006). This creates another big challenge in the digital forensic domain.

3.4.3.4 Lack of Forensic Knowledge Reuse among Personnel

According to Bruschi et al. (2004), when investigators conduct an investigation and handle a massive amount of information, they usually use specialised expertise and analyse an extensive knowledge base of digital evidence. Most of the work done is not clearly documented and this hinders external assessments and training. Previous experiences may and should be used to instruct new workforce, to promote knowledge sharing and reuse among investigators, and to expose gathered information to quality evaluation by third parties. Hoss and Carver (2009) add that the putting together of potential digital evidence may many a times be insufficient to endorse legal actions in court or during civil prosecutions. This is because the potential evidence and methods employed to extract the digital evidence did not comply with acceptable legal believes, thus posing a challenge.

3.4.3.5 Challenges Pertaining to Forensic Investigator Licensing Requirements

Schwerha (2008) reports on a push in the United States to require digital forensic professionals to become licensed as private investigators. However, there are many reasons why digital forensic professionals should not be required to license as private investigators. The requirement of licensure will limit the field unnecessarily, as there are too many potential jurisdictions worldwide to allow the average practitioner to be licensed in every jurisdiction (Schwerha, 2008). Moreover, requiring digital forensic professionals to become licensed private investigators will create a big challenge to most average investigators worldwide. The requirement to be a licensed private investigator has little or no connection to the skill set that is necessary to be a high-quality digital forensics professional (Schwerha, 2008).

In the next section, the operational challenges faced by digital forensics are discussed.

3.4.4 Operational Challenges

According to Whitehead (2013), digital crimes (perhaps more than any other type of crime) can be international in their operational scope. Basic recommendations for evidence acquisition need to be set globally. These recommendations range from extensive principles that apply virtually to every investigation through organisational practices. Guidelines will ensure that a minimum standard of planning, performance, monitoring, documenting and reporting is maintained to recommended processes, methods and software and hardware solutions.

In this sub-section of the paper, some of the identified operational challenges faced by digital forensics are explained in detail.

3.4.4.1 Difficulties in Incidence Detection, Response and Prevention

Traditional IT environs with on-premises data processing largely depend on internal security incident management process that utilize monitoring, log file analyses, intrusion detection systems (IDSs), and data loss prevention (DLP) to discover trespassers, attacks and data loss. According to Beham (2012), discovering security incidents is many times a challenge especially for cloud users. Moreover, incident response is needed because attacks regularly compromise personal and business data. It is critically important to respond quickly and efficiently when security violations occur, so as to minimise the loss

or theft of information and disruption of services caused by such incidents (Cichonski et al., 2012).

3.4.4.2 Lack of Standardised Processes and Procedures

The lack of standardisation in digital forensics seriously hinders the investigation process (Leigland and Krings, 2004) and makes it difficult to produce legally admissible digital evidence. As of the time of writing this research thesis, there existed no standardised digital forensic investigation process model for recovering potential digital evidence. According to Köhn et al. (2006), the number of digital forensic process models that have previously been proposed has added more problems to the digital forensic field. This has, therefore, led to a call for standardisation (ISO/IEC 27043, 2015) so as to facilitate the digital forensic investigation process. Recent research has also advocated for new forensic methods and tools that will be able to successfully investigate anti-forensics techniques (Alharbi et al., 2011).

3.4.4.3 Significant Manual Intervention and Analysis

In most cases a physical hard drive image will have to be manually scrutinized and analysed. This process can be simple in a single drive, single partition, as well as a completely allocated disk drive. However, the same process becomes difficult and poses a challenge with multi-volume Redundant Array of Independent Disks (RAID) configurations (King, 2006). According to Ayers (2009), digital forensic analysis is a very complex undertaking. Thus, whenever the process is under manual control, mistakes will be made and bias could be introduced, even inadvertently, thus posing a big challenge to investigators.

3.4.4.4 Digital Forensic Readiness Challenge in Organisations

According to Mohay (2005), forensic readiness is the extent to which computer systems or computer networks record activities and data in such a manner that the records are sufficient in their extent for subsequent forensic purposes, and the records are acceptable in terms of their perceived authenticity as evidence in subsequent forensic investigations. However, Cobb (2013) states that digital forensic readiness sounds like a demoralizing challenge to quite a good number of organisations.

With the advances in cloud computing, organisations have been forced to change the way they plan, develop and enact their IT strategies. According to Reilly et al.

(2011), cloud computing has not been thoroughly considered in terms of its forensic readiness. Hence, there exists a clear need to consider current best practices to include for example certain features of digital forensic readiness in the existing practices to deal with the challenges brought about by a lack of forensics readiness in organisations.

Barske et al. (2010) adds that, although the need for digital forensics and digital evidence in organisations has been explored (as also the need for digital forensic readiness within organisations); decision makers still need to understand what is needed within their organisations to ensure digital forensic readiness.

3.4.4.5 Trust and Audit Trail Challenges

The aim of digital forensics is to examine digital media in a forensically sound fashion, but with additional recommendations and trusted procedures developed to generate legal audit trails. The proof of clear and original audit trails plays a key role in user accountability and digital forensics. However, it is possible that an attacker may edit or delete the audit trail on a computer, particularly in the case of weakly protected personal computers (Yong, 2013). Modern rootkits that dynamically change kernels of running systems to hide what is happening or even to produce false outcome are also on the increase, hence posing a challenge to digital forensic investigators. In the next section, the chapter conclusion is presented.

3.5 CHAPTER CONCLUSION

In this chapter the researcher explained the different challenges faced by digital forensics. This was followed by the scope of the proposed taxonomy in this study. This was done to show the impact and disparities caused by different identified challenges to the digital forensic domain. A taxonomy of the various challenges faced by digital forensics was subsequently presented in this chapter, based on a survey of the existing digital forensic literature. The taxonomy classified the large number of digital forensic challenges into a few well-defined and easily understood categories.

The reader is again reminded that the purpose of this chapter is to serve as a survey of the status quo of the research area. For this reason, more specific categories and sub-categories of the challenges can and should be added to the taxonomy as the need arises in future.

In the next chapter (Chapter 4), the background of ontologies is introduced in a bid to establish a foundation for creating a unified formal representation of digital forensic

knowledge and information. However, ontologies can also be useful in resolving semantic disparities in digital forensics. The development of ontologies for digital forensics is explained in more detail later in Chapter 6 of this research thesis.

CHAPTER 4 : BACKGROUND OF ONTOLOGIES

4.1 INTRODUCTION

The term ontology finds its primary source in the domain of philosophy where it is used to refer to the subject of existence. Still in philosophy, ontology can also be referred to as the area of study that deals with the nature of reality. In computer science, however, ontologies are used in many different contexts and for many different purposes.

For this reason, different definitions of ontology by different researchers are introduced in Section 4.2. Section 4.3 goes on to explain the different methodologies for ontology development, while the types of ontologies are discussed in Section 4.4. Ontology development tools are dealt with in Section 4.5 and the chapter is concluded in Section 4.6.

Chapter 4 is meant to introduce the reader to the concepts of ontology as computing models that can help put together domain information and bring forth a harmonised comprehension of the domain facts that can be used, reused and shared among different groups of people.

Note that, the presentation in this chapter is based on information gathered from existing literature, which has offered useful insights into the research presented in this thesis. Chapter 6 will, however, elaborate on how the ontology concepts discussed in Chapter 4 can be used in digital forensics as one way to resolve semantic disparities as well as present the benefits of developing ontologies for the digital forensic domain.

4.2 ONTOLOGY DEFINITION

To begin with, Smith et al. (2006) define ontology as an exhaustive formal specification of how to represent entities that exist in a given domain and the different relationships that exist among the entities. According to Van Rees (2003), ontology is a collection of well-defined ideas explaining a specific domain of interest. Grüber (1993), on the other hand, defines ontology as an exhaustive specification of a conceptualisation. Staab et al. (2001) adds that ontologies are meant to capture domain information in a generalised way as well as provide a commonly agreed-upon comprehension of the domain, which may be reused and shared across different applications and groups of people.

According to Gokhale et al. (2011), ontologies represent a domain of knowledge and allow relationships such as the definition of classes, relations and functions.

Regardless of their high-level specifications, ontologies also permit flexibility. Nevertheless, for any ontology to be useful it must represent a shared, agreed-upon conceptualisation (Castañeda et al., 2010), in other words it should be accepted by a group of people or a community.

Different groups of people or communities build ontologies for different reasons. However, Noy and McGuinness (2001) made the following summary of some of the reasons why people build ontologies:

- To share a common understanding of the structure of information among people or software agents
- To enable reuse of domain knowledge
- To make domain assumptions explicit
- To separate domain knowledge from the operational knowledge
- To analyse domain knowledge

The different methodologies used for ontology development proposed by different researchers and research organisations are explained in the next section, followed by a discussion of the different types of ontologies.

4.3 ONTOLOGY DEVELOPMENT METHODOLOGIES

According to Gaevic et al. (2009), an ontology development methodology is normally a set of established principles, processes, practices, methods and activities used to design, construct, evaluate and deploy ontologies. To build any high quality ontology, Leung et al. (2012) state that ontology developers need to select and follow an appropriate development methodology consisting of a sequence of steps, activities and guiding principles that are put together in an organised and methodical way. Gaevic et al. (2009) remark that a single, best-known ontology development methodology does not exist yet, because there is still no consensus about a single 'correct' way to model a domain. Furthermore, while the ontology development process is inevitably an iterative process, the available literature shows that constructing ontologies by reusing available ontologies is cheaper than constructing from scratch.

The ontology development process as discussed by Brusa et al. (2006) can be categorised into two main steps: a specification step and a conceptualisation step. The primary aim of the specification step is to get informal understanding about the domain,

whereas the aim of the conceptualisation step is to organise as well as structure the domain information with the help of external representations.

Several ontology development methodologies have been proposed in literature by different researchers and research organisations. Most of these methodologies concentrate on building ontologies from scratch. A few others exist, though, that include methods for merging, re-engineering, maintaining and evolving ontologies (Gaevic et al., 2009). Other ontology development methodologies also exist that exploit the idea of reusing existing ontological knowledge in building new ontologies. In the sub-sections to follow, some of the ontology development methodologies are explained. The focus falls, however, on the methodologies that could be useful in developing domain ontologies that help create a unified formal representation of knowledge for digital forensics as well as help in resolving semantic disparities.

4.3.1 Brusa, Caliusco and Chiotti Methodology

In their research Brusa et al. (2006) proposed an ontology development process that can be categorised into two main steps: a specification step and a conceptualisation step. The specification step is meant to get informal understanding about the domain of interest, whereas the conceptualisation step is meant to organise and structure the domain information using external representations. The whole ontology development methodology used in this study was, thus, based on (Brusa et al., 2006). The steps as discussed by Brusa et al. (2006) methodology are briefly explained below:

4.3.1.1 Specification

During the specification phase the primary aim is to gain knowledge about the domain, (in this case digital forensic domain) and what needs to be achieved Brusa et al. (2006).

4.3.1.2 Conceptualisation

According to Nagyp´al, (2007) this phase is the most complex tasks in ontology development and aims to produce a model of the research methodology domain in a form that will allow communication with domain experts who may not be fully conversant with ontology languages (Nagyp´al, 2007).

4.3.2 Uschold and King's Methodology

Uschold and King (1995) proposed an ontology development methodology that has four different phases: identifying the purpose, building the ontology, evaluation and

documentation. The Uschold and King (1995) methodology was used to develop the Enterprise Ontology, which supports and enables the exchange of information between different individuals, individuals and computational systems, as well as among dissimilar computational systems (Öhgren, 2009).

In the first phase, it is important for ontology developers to be clear about why are they building the ontology and what is the ontology going to be used for. This phase may also consider who are the ontology users and how will they use the ontology.

The second phase involves coming up with the ontology itself. This phase is further divided into three different parts, namely ontology capture, ontology coding and ontology integration.

- **Ontology Capture**: This includes the singling out of the key ideas as well as relationships in the domain of interest; the production of accurate and unambiguous text descriptions for the ideas and relationships; the identification of terminologies to refer to the ideas and relationships; and finally to agree on all the above.
- **Ontology Coding:** This means the explicit representation of the captured conceptualisation using formal language. Coding may also include choosing a good formal representation language, and creating the code thereafter. Sometimes the capture and coding steps are merged into a single step during the ontology development process. Uschold and King (1995) nevertheless suggested the separation of the two.
- **Ontology Integration:** It involves the question of whether to use already existing ontologies to build new ontologies and, if so, how this can be accomplished.

The third phase in Uschold and King's (1995) methodology is evaluation. In this phase, the ontologies, their associated software environment and documentation are judged technically. This process may also include requirements specifications and competency questions.

The fourth and last phase is documentation. This phase recommends the establishment of guidelines for documenting ontologies. The latter may differ, based on the type and purpose of the ontology developed.

4.3.3 Grüninger and Fox's Methodology

According to Grüninger and Fox (1995), the goal of any ontology is to agree on a shared terminology and set of limitations on the entities found in the ontology. Individuals must concur on the objective and final use of the ontology. However, the development of any ontology should first be inspired by a sequence of events that emerge in the applications. Such events may originate from industry partners as problems that they experience in their organisations. Figure 4.1 below shows the procedure for ontology design and evaluation based on the work of Grüninger and Fox (1995).

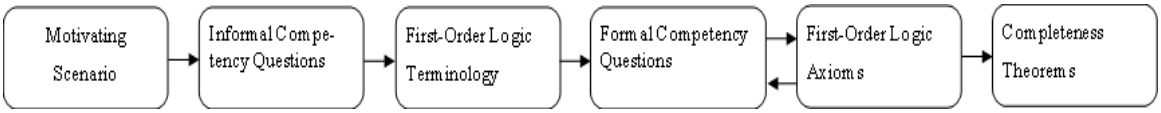


Figure 4.1 Procedures for Ontology Design and Evaluation (Grüninger and Fox, 1995***)***

Infer from Figure 4.1 that any proposal to build a fresh ontology or extend an existing ontology, one must first clearly illustrate the motivating scenario, together with a set of the pre-planned solutions to the problems presented in the scenario. By presenting a scenario, developers can easily comprehend the motivation for the proposed ontology in terms of its application.

Based on the motivation scenario as shown in the first step of Figure 4.1, a set of questions may arise, placing demands on the underlying ontology in the second step. These questions are called informal competency questions, since they are not yet in the conventional ontology language. The informal competency questions can be used to provide an informal justification for the fresh or extended ontology, as well as to examine the ontological commitments that have been made.

The third step in the Grüninger and Fox (1995) methodology involves specifying the terminology of the ontology using the first-order logic. If a new ontology is to be developed, for example, then for every informal competency question, there must be an object, attributes or relations in the proposed ontology or proposed extension to the ontology, which are needed to answer the questions. In stating the terminology of any ontology, identifying the object in the domain of discourse can be represented by constants and variables in the language. Attributes of the objects can be defined using unary predicates, while the relations among objects can be defined using n-nary predicates.

Formal competency questions constitute the fourth step of the proposed Grüninger and Fox's (1995) methodology, where the competency questions are defined as an entailment or consistency problem with respect to the axioms in the ontology. However, formal competency questions place restrictions on the axioms to be included in the ontology. Nonetheless, all terminologies in the statement of the formal competency questions must be included in the terminology of the ontology. Using formal competency questions is a way of evaluating the ontology and its adequacy.

The fifth step as shown in Figure 4.1 involves defining axioms in the first-order logic. According to Grüninger and Fox (1995) this is one of the most difficult aspects of defining ontologies, because the axioms must be necessary and adequate to express the competency questions and characterise their solutions. Without the axioms it is difficult to express the question or its solution.

Finally, the sixth and last step in Figure 4.1 is to generate completeness theorems for the ontology. This step defines the situation under which the solutions to the questions are finalized. It also forms the basis of completeness theorems for the ontology. The Grüninger and Fox (1995) methodology was used to develop the Toronto Virtual Enterprise (TOVE) ontology as part of the TOVE Enterprise Modelling project. The main objective of the project was to produce an enterprise model that could infer answers to many 'common sense' questions about the enterprise (Öhgren, 2009).

4.3.4 Methontology

Developed by Fernandez et al. (1997), methontology is one among the comprehensive ontology engineering methodologies for building ontologies from scratch. The phases involved in this methodology include specification, knowledge acquisition, conceptualisation, integration, implementation, evaluation and documentation.

- **Specification:** The primary objective of the specification phase is to specify the purpose of the ontology, its intended uses, scenario of uses, and end users. Specification may also include the level of formality of the actualized ontology and its scope, which includes the set of terminologies to be constituted, its characteristics and granularity.
- **Knowledge Acquisition**: This is usually treated as an independent activity in the ontology development process. However, this phase can be handled simultaneously with other activities. The knowledge acquisition techniques used

may include brainstorming, interviews, formal and informal analysis of texts, and knowledge acquisition tools.

- **Conceptualisation**: This phase structures the domain knowledge in a conceptual model that describes the problem and its solution in terms of the domain vocabulary identified in the ontology specification activity. In addition, a glossary of terminologies with all possibly useful knowledge in the given domain is constructed in this phase.
- **Integration**: The goal of integration is to help speed up the construction of the ontology by reusing definitions that are already built into other ontologies, instead of starting from scratch.
- **Implementation**: The outcome of the ontology implementation phase is the ontology codified in a formal language like Ontolingua or any other existing language.
- **Evaluation**: The evaluation phase involves technical judgment of the ontology developed, the software environment and documentation with respect to a frame of reference during each phase and between phases of the life cycle. The ontology evaluation also includes verification and validation. Verification takes care of the correctness of the ontology, while validation guarantees that the ontology, the software environment and documentation correspond to the system that they are supposed to represent.
- **Documentation**: The objective of this phase is to ensure that each phase of the above described methodology results in a document that explains the ontology that was built. This includes documents such as a requirements specification document, a knowledge acquisition document, a conceptual model document, a formalisation document, an integration document, an implementation document and an evaluation document.

4.3.5 Karlsruhe and Ontoprise Methodology

Staab et al. (2001) defined a methodology for ontology development that consists of five phases: feasibility study; ontology kick off; ontology refinement; ontology evaluation; and ontology maintenance phase.

- **Feasibility Study**: The main aim of this phase is to help determine the economic and technical feasibility of the project. In this phase the problem and opportunity

areas are identified, followed by selecting the most favourable areas and the best solution to any potential problems. The feasibility study is normally conducted before the ontology is developed, because it forms the basis for the kick-off phase.

- **Ontology Kick-off**: In the kick-off phase the description of what is supported by the ontology and the planned area of the ontology application are produced. It is also in this phase that the ontology requirements specification document is produced. This helps an ontology engineer to decide on what to include, exclude as well as the hierarchical structure of ideas in the ontology.
- **Ontology Refinement**: This phase involves the refinement of the initial draft of the ontology. The main goal here is to produce mature and application-oriented target ontology according to the specifications provided by the kick-off phase.
- **Ontology Evaluation**: The aim of ontology evaluation is to check whether the final ontology satisfies the needs of the ontology specification document and if also the ontology supports or answers of the competency questions analysed in the kick-off phase. A test of the ontology in its targeted application environment is also carried out in this phase to help with gathering valuable feedback from the users and further refinement of the ontology.
- **Ontology Maintenance**: This is the last phase of ontology development and contains the rules for updating, deleting and inserting processes within the ontology. Feedback from users is usually valuable for identifying change to or maintaining ontologies.

4.3.6 Unified Methodology

The unified methodology for the development of ontologies was proposed by Uschold (1996). The unified methodology is obtained from and compatible with both the TOVE and Enterprise methodologies. The phases proposed in the unified methodology include the following: identify purpose; level of formality; identify scope; build the ontology; and formal evaluation or revision cycle.

- **Identify Purpose:** In this phase the purpose of the ontology is defined. The developers should have a clear reason of why they are building the ontology, such as what the ontology will be used for and the possible mechanisms for use. If developers cannot identify the purpose of building the ontology, then they should

consider if it is worth to continue developing the ontology before encountering problems in later stages.

- **Level of Formality:** In this phase the ontology developers must decide how formal the ontology needs to be. This question is determined by the motive and users of the ontology. Besides, the degree of formality needed increases with the degree of automation in the activities supported by the ontology. Sometimes, both an informal and a formal ontology may be needed to fulfil both technical and non-technical users.
- **Identify Scope:** in this phase a set of ideas and terminologies covering the full range of information that the ontology must characterise to meet the requirements identified is produced. The scope of the ontology can be identified by coming up with a detailed scenario that emerges in the applications. This may include problems as well as possible solutions to the identified problems. Brainstorming can also be used instead of or in conjunction with motivating scenarios and competency questions to do a thorough and accurate ontology scoping job.
- **Building the Ontology:** The main goal of this phase is to come up with the definitions. However, some decisions must be taken as to how and whether to arrange the definitions in any specific way to help structure the ontology.
- **Formal Evaluation or Revision Cycle:** This is the last phase of the development methodology described by Uschold (1996). Here the developers compare the competency questions or the user requirements with the developed ontology. The different types of ontologies are explained in the next section.

4.4 TYPES OF ONTOLOGIES

There exist different types of ontologies for different areas of application. According to Davies et al. (2004), ontologies are becoming popular predominantly due to what they guarantee: a shared and common comprehension of a domain that can be communicated between individuals and systems. Vanitha et al. (2011) adds that ontologies vary greatly in size, scope and semantics. Some of the ontologies identified to facilitate the research reported on in this thesis are shown in the bulleted list to follow.

- Generic ontologies
- Specialised ontologies
- Domain ontologies

- Task ontologies
- Domain-independent and Domain-specific ontologies
- Application ontologies
- Terminological ontologies
- Representational ontologies
- Metadata ontologies
- Method ontologies
- Enterprise ontology

The sub-sections to follow elaborate further on the various types of ontologies listed above.

4.4.1 Generic Ontologies

Generic ontologies, also known as common sense ontologies, are developed to define basic notions and concepts that are generic across many different fields such as time and space. This implies that generic ontologies are designed to be shared by a large number of communities, in other words they can be applied to different specialised domains. According to Hadzic et al. (2009), generic ontologies can be accessed by anyone without having to authenticate to a system. In addition, they can be used for searching concepts relating to a domain. However, generic ontologies consist of a minimal number of axioms and it is possible to progress gradually from a generic ontology to a specialised ontology through an incremental process in the number of axioms (Hadzic et al., 2009). Figure 4.2, for example, shows the concept of moving from a digital forensic generic ontology to digital forensic specialised ontologies.

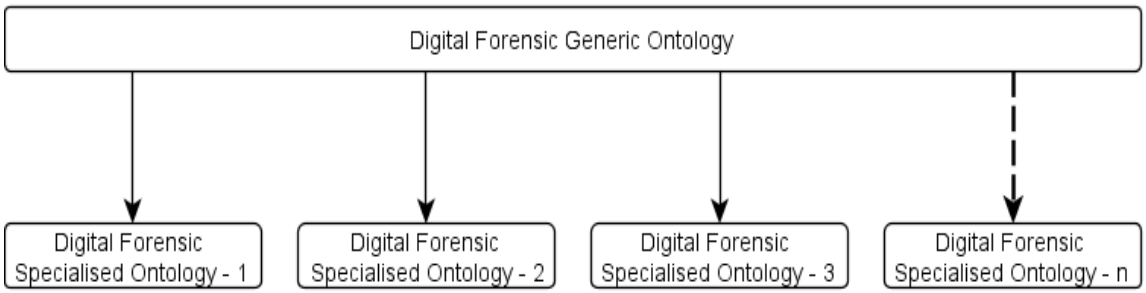


Figure 4.2 Moving from a Generic Ontology to Specialised Ontologies

4.4.2 Specialised Ontologies

Specialised ontologies are usually developed for a smaller group of people within a larger community. They contain more details because they refine and extend the general

descriptions present in generic ontologies (Hadzic et al., 2009). The process used to build a generic ontology based on a number of specialised ontologies is called ontology generalisation. Figure 4.3 shows the concept of moving from digital forensic specialised ontologies to a digital forensic generic ontology.

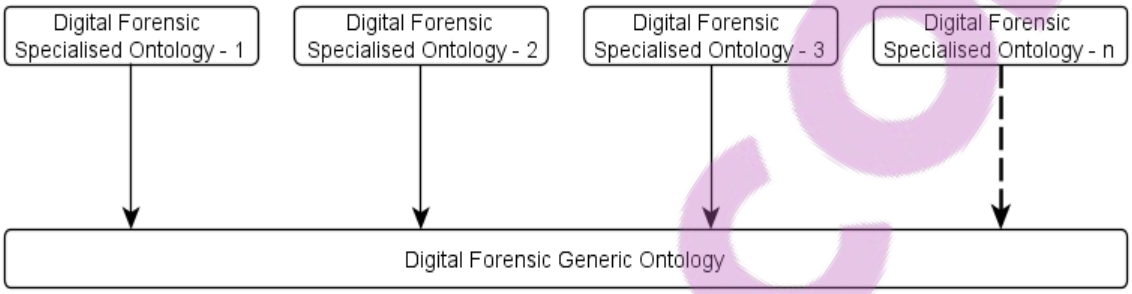


Figure 4.3 Moving from Specialised Ontologies to a Generic Ontology (Ontology Generalisation)

4.4.3 Domain Ontologies

Domain ontologies are designed to represent knowledge relevant to a particular domain, for instance digital forensics. The goal of a domain ontology is to reduce (or eliminate) the conceptual and terminological disparities among the members of a community who need to share information of various kinds (Navigli and Velardi, 2004). Domain ontology has a broad coverage of the domain's terminology, which is achieved by identifying and properly defining a set of relevant concepts that characterise a given application domain. Domain ontology can be categorised as either: task dependent or task independent (Navigli and Velardi, 2004). Note that a task-dependent ontology has some specific domain information that can be used to deal with a particular problem, whereas a task-independent ontology may cover the structure of an object, or the theories and principles that govern the domain.

4.4.4 Task Ontologies

Task ontologies are usually developed with the aim of reusing task knowledge, in other words to provide terminologies specific for particular tasks. Task ontologies generally explains the terminology related to a particular task (Guarino, 1998), for example, the digital forensic investigation process or evidence presentation process. In contrast to domain ontologies, there exist no widely accepted procedures for engineering task ontology or consistency in representing them. However, Martins and Falbo (2008) argue that a task ontology should be able to capture two entwine views: (i) task decomposition

into sub-tasks and control flow, and (ii) knowledge roles to be played by domain concepts in those sub-tasks. This is because the two perspectives are complementary.

4.4.5 Domain-independent and Domain-specific Ontologies

According to Lee et al. (2006), domain-independent ontology provides basic concepts and relations that are adopted to build domain-dependent and domain-specific ontologies. In addition, domain-independent ontologies are intended to be fundamental and universal to ensure generality and expressivity for a wide range of domains, whereas a domain-dependent ontology serves as a bridge between a domain-independent ontology and a domain-specific ontology. A domain-specific ontology, though, specifies concepts particular to a domain of interest, for example, concepts related to digital forensics. Besides, domain-specific ontologies also represent the specified concepts and their relations from a domain-specific perspective (Lee et al., 2006).

4.4.6 Application Ontologies

According to Sacramento et al. (2010), application ontologies enable the identification and the association of semantically corresponding concepts, and thereby assist with information discovery and retrieval, as well as data or information integration. Guarino (1998), on the other hand, argues that application ontologies discuss ideas depending both on a specific domain and task, which are many a times specialisations of both the related ontologies. These ideas often correlate with the responsibilities played by domain entities while carrying out a certain activity.

4.4.7 Terminological Ontologies

According to Madsen and Thomsen (2009), terminological ontologies model concepts and the relations between those concepts, where a concept is described by means of characteristics that denote properties of individual referents belonging to the extension of that concept. Labský (2005), however, states that terminological ontologies are centred on human-language terminologies, without direct reference to the real world. A good example of an existing terminological ontology would be the WordNet, which is used for annotations (Miller, 1995).

4.4.8 Representational Ontologies

According to Iordan and Cicortas (2008), representational ontologies do not devote themselves to any specific domain. This also implies that representational ontologies

provide representational entities but do not state what is to be represented. For example, the Frame Ontology (Gruber, 1993) defines frames, slots, and slot constraints, allowing the communication of knowledge. The Frame Ontology was originally built for capturing knowledge representation conventions under a frame-based approach in Ontolingua, but it was later modified. The reason behind the modification was the creation of the Open Knowledge Base Connectivity (OKBC) Ontology. OKBC is an application programming interface for obtaining knowledge bases stored in knowledge representation systems (SRI, 2013).

Based on the individuals responsible for building an ontology, there are other types of ontologies that can be of value and that are discussed below.

4.4.9 Metadata Ontologies

Metadata ontologies give a vocabulary for depicting the content of on-line information sources. The metadata vocabulary often reveals a set of ideas or terminologies and their associated definitions and connections to each other. The terminologies are usually known as elements, attributes and qualifiers. The definitions usually give semantics that are both human and machine legible (Baker et al., 2001).

4.4.10 Method Ontologies

Method ontologies provide terminologies specific to particular problem-solving methods (Iordan and Cicortas, 2008). In addition, method ontologies are essentially a characterisation of the information type of a method, its primitive semantic categories, their properties, and their logical connections. By browsing through method ontologies, then, an agent can better understand a model (Uschold & Gruninger, 1996). Because method ontologies contain both informal and formal descriptions of the semantic categories of the method, it can also be used in some situations to enforce rules and constraints defined in the method.

4.4.11 Enterprise Ontologies

Enterprise ontology is a formal and clear specification of a shared conceptualisation among a community of people of an enterprise, in other words a group of terminologies and definitions pertinent to the business enterprises (Dietz and Delft, 2006; Uschold et al., 1998). Enterprise ontologies also allow organisations to come to a shared

understanding of the terminologies and concepts that are core to their business processes and applications.

Other ontology types that the reader can also explore further include: workplace ontologies; resource ontologies; personal ontologies; knowledge modelling ontologies; and information ontologies to mention a few. In the researcher's opinion these ontologies were found to be too specific (focused on a single entity), hence the decision not to tackle them in this study.

The next section explains in brief some of the ontology development tools available.

4.5 ONTOLOGY DEVELOPMENT TOOLS

A number of tools are available essentially to assist any individual constructing new ontologies or editing existing ontologies. Such tools can also help individuals in merging multiple existing ontologies. However, according to Duineveld et al. (2000) the usefulness of any ontology development tool is determined by the level of the users and the stage of development of the ontology. In this section, several ontology development tools are explained. The list in this section explores a number of common examples selected to facilitate this study and is not in any way an exhaustive list.

4.5.1 ProtégéWin

ProtégéWin is a Windows-based ontology development tool designed for building ontologies of domain models (Duineveld et al., 2000). It provides a development environment for authoring ontologies and electronic knowledge bases. ProtégéWin assists application developers in creating and maintaining clear domain models, and in incorporating those models straight into their program code. The Protégé methodology, to which ProtégéWin belongs, also allows system builders to build software systems from modular components, including reusable frameworks for assembling domain models and reusable domain-independent problem-solving methods that implement procedural strategies for solving tasks (Eriksson et al., 1995). Based on any particular ontology under construction, ProtégéWin can generate a knowledge acquisition tool for entering the instances of that ontology (Duineveld et al., 2000).

There also exist third-party plugins that extend the ProtégéWin platform's functionality, such as *Web Protégé*, which is an online version of ProtégéWin striving to get all of the native functions. *Collaborative Protégé* is a plug-in extension of the

existing ProtégéWin system that supports collaborative ontology editing, as well as the annotation of both ontology elements and ontology changes.

4.5.2 The NeOn Toolkit

Developed by the NeOn Foundation, the NeOn Toolkit (NeOn, 2013) is an ontology engineering environment founded as part of the NeOn Project. The NeOn toolkit is an open source multi-platform ontology engineering environment that offers comprehensive support for the whole life cycle of ontological engineering. Ontological engineering, according to Pérez et al. (2004), refers to the set of tasks that concern the ontology development process, the ontology life cycle, the methods and methodologies for building ontologies, and the tool suites and languages that support them. The NeOn Toolkit is based on the Eclipse platform. Eclipse is a leading development environment that also offers a considerable set of plug-ins covering different ontology engineering tasks (NeOn, 2013).

4.5.3 Ontolingua

This ontology development environment offers a suite of ontology authoring tools and a library of modular, reusable ontologies (Farquhar et al., 1997). Ontolingua also offers a distributed collaborative environment to browse, create, edit, modify and use ontologies. Ontolingua makes the development of new ontologies easy by incorporating (parts of) existing ontologies from an existing repository. The repository has a huge number of ontologies from varying fields. After finalization, the ontology developed can be included into the repository for possible reuse (Duineveld et al., 2000). The tools in Ontolingua are also aligned towards the authoring of ontologies by assembling and extending ontologies acquired from the repository.

4.5.4 Knoodl

Knoodl is a product of Revelytix Inc. (Revelytix, 2015). Knoodl makes it easy for community-aligned development of Ontology Web Language (OWL)-based ontologies and the Resource Description Framework (RDF) knowledge bases. Knoodl is also a Distributed Information Management System (DIMS) that contains tools that cater for activities like creating, managing, analysing and visualising RDF and OWL descriptions. Knoodl features support collaboration in all stages of these activities, and is normally hosted in the Amazon Elastic Compute Cloud (Amazon EC2) and it can be used for free

(Revelytix, 2015). Note that Amazon EC2 is a web service that offers resizable computing capacity in the cloud. The design of Amazon EC2 makes web-scale computing easier for developers (Amazon, 2015).

4.5.5 DERI Ontology Management Environment (DOME)

DOME is a product of the Ontology Management Working Group (OMWG, 2015) whose mission is to come up with a suite for efficient and effective management of ontologies, which gives an essential solution of the overall problem. It is a programmable Extensible Mark-up Language (XML) editor used in a knowledge extraction role to transform Web pages into Resource Description Framework (RDF). The main inspiring rules of DOME are simplicity, completeness and reuse. The ontology management suite comprises tooling support for editing and browsing, versioning and evolution, as well as mapping and merging offered in the form of freely combinable Eclipse plug-ins (OMWG, 2015).

4.5.6 Sigma

Sigma is an open source knowledge engineering environment for developing, viewing and debugging theories in first-order logic (Pease, 2003). Sigma is considered an appropriate environment for the development of expressive ontologies in first and higher order logic (Pease and Benzmüller, 2012). Sigma also works with the Knowledge Interchange Format (KIF) and is optimised for the Suggested Upper Merged Ontology (SUMO). SUMO and its domain ontologies constitute the largest formal public ontology being used today for research and applications in search, linguistics and reasoning. SUMO is the only formal ontology that has so far been mapped to the entire WordNet lexicon (Pease, 2013). Note that WordNet® is a vast lexical database of English where nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms, each expressing a well-defined idea. The synonyms are connected together by means of conceptual-semantic and lexical relations. The resulting network of meaningfully related words and concepts can be traversed with the help of a browser. The WordNet's structure makes it a very helpful tool for computational linguistics and natural language processing.

Additionally, Sigma includes various important features for knowledge engineering work, including terminology and hierarchy browsing, the ability to load different files of logical theories, a full first-order inference capability with structured

proof results, a natural language paraphrase ability for logical axioms, and support for displaying mappings to the WordNet lexicon and various knowledge base diagnostics (Pease, 2003).

4.6 CHAPTER CONCLUSION

In this chapter the researcher examined and explained the basic concepts of ontologies. The ontology development process, which includes ontology development methodologies, was also explained. Different types of ontologies were identified and explained in this chapter as well. Finally, several ontology development tools were discussed. This was primarily done to help the reader capture the ontology concepts that are later used in Chapter 6 for developing ontologies for digital forensics.

The next chapter (Chapter 5) explores semantic disparities in digital forensics as well as how to manage them. This chapter also form part of the main contribution of this research study.

CHAPTER 5 : SEMANTIC DISPARITIES IN DIGITAL FORENSICS

5.1 INTRODUCTION

According to Aye et al. (2008), communication tools mainly facilitate connections and interactive communication between individuals. The communicating individuals usually share an area of communicative commonality. However, the communication tools used at any point may not contribute any information for a conversation topic. For example, when you send an e-mail to an individual and you forget to include particular details, the e-mail program – being a tool that facilitates the communication with the receiver – may not in any way add such information on behalf of the sender. This also implies that the communication tools may lack the ability to define some of the terminologies (on behalf of the sender or the receiver) used during communications. For this reason, participants may face problems (because of semantic disparities) when they do not understand a particular terminology or topic during a conversation. As a result, people end up having difficulty both in communicating and exchanging information among themselves. This is especially common when the people who are engaging in a conversation have variant backgrounds.

In the context of digital forensics, communication usually involves different stakeholders – investigators, computer professionals, law enforcement agencies – who should ideally always cooperate in this profession. Unfortunately, semantic barriers in communication may become apparent because of the different backgrounds that are bound to characterise participants. This implies that the parties involved may have difficulties comprehending terminologies used during the communication period. For example, during the presentation of digital evidence in court or civil proceedings, the use of terminologies such as initial response, first response, incidence response (among other terms) can lead to misapprehension if not defined according to the context of use. Misunderstandings can therefore occur between individuals because of the difficulties and differences in comprehending the terminologies used.

Note that semantic barriers in communication refer to the misunderstandings that can occur between two individuals who try to communicate while both individuals attach

different meanings, interpretation and description to the terminology used. To avoid misunderstandings, the sender of any information must use terminologies that are similar in meaning and that the receiver will understand in the same manner as intended by the sender. This implies that the connotation of the terminologies used by the sender should be related to the content and the context at hand.

The aim of this chapter is to identify, present and discuss semantic disparities in digital forensics. However, this Chapter also elaborate on how to manage the identified semantic disparities in the digital forensic domain. This is in line with the primary area of focus in this research study, which is to resolve semantic disparities in digital forensics.[1]

The remaining part of this chapter is constructed as follows: Section 5.2 explains what semantic disparity is, while Section 5.3 discusses the advances in semantic disparity research. The potential causes of semantic disparity in the digital forensic domain are identified, presented and discussed in Section 5.4. Managing sematic disparities is handled in section 5.5 while Section 5.6 explains the identified approaches to manage the disparities. The advantages of semantic reconciliation are then discussed in section 5.7. Finally, Section 5.8 presents a conclusion of this chapter.

5.2 DEFINING SEMANTIC DISPARITY IN DIGITAL FORENSICS

Semantic disparity as defined by Xu and Lee (2002) refers to "disagreements about the meaning, interpretation, description and intended use of the same or related data". According to the Oxford Dictionaries (2013), disparity refers to the state of being different (lack of uniformity). For the purpose of this study, though, semantic disparity is used to refer to disagreements about the interpretation, descriptions and representation of the same or related data or information and terminologies in a domain of interest.

Semantic disparity as discussed in this research thesis is sometimes addressed as 'semantic heterogeneity' or 'semantic gap' in other previous research works (Xu and Lee, 2002; Sheth and Larse, 1990; Wang and Liu, 2009). Nevertheless, in this study the researcher adopts the use of the term 'semantic disparity' in place of semantic heterogeneity or semantic gap.

[1] The contents of this chapter were presented at the international conference on digital forensics, security and law (ADFSL-2013), Richmond, Virginia, USA. The paper entitled "Significance of Semantic Reconciliation in Digital Forensics" was later published by the Journal of Digital Forensics, Security and Law.

Semantic disparity is not merely a problem during communication of domain data or information, but generally a hard problem to solve. According to Sheth and Larsen (1990), the problem of semantic disparity is not well understood in many domains and in the case of this research study, semantic disparity in digital forensics is not well understood either. There exists no consensus concerning a clear interpretation of the semantic disparity problem in general (Xu and Lee, 2002; Sheth and Larse, 1990).

Already in 2011, as mentioned earlier, Dolan-Gavitt et al. agreed that the digital forensics community was struggling with semantic disparity. This was evident especially during the reconstruction of human-readable (high-level state) information from low-level data sources such as physical memory.

Unfortunately, as mentioned in the problem statement in Section 1.2, digital forensics lacks comprehensive methods and specifications that can assist in resolving the semantic disparities that hamper communication between the different digital forensic stakeholders. Developing methods in digital forensics that can be used to resolve semantic disparities is therefore a worthwhile aim. The next section elaborates on the advances in semantic disparity research.

5.3 ADVANCES IN SEMANTIC DISPARITY RESEARCH

Research in semantic disparity is generally centred on resolving semantic distractions between different parties who work together to achieve the same goal or arrive at the same end. Semantic distractions occur when domain terminologies are used differently from what is preferred. This also means that a single domain terminology may convey many different meanings. Advances in this research have managed to identify different forms of semantic disparity that are worth presenting in this section. A majority of these disparities, though, focus more on the field of databases, while others focus on distributed systems (Karie and Venter, 2013). (Note that although this section is actually a brief background, it is not included in the background section of this thesis because it is directly related to this chapter and not applicable to the rest of the thesis.)

To begin with, efforts by Colomb (1997) presented the case for structural semantic disparity. Structural semantics defines the relationships between the meanings of terminologies within a sentence. Structuralism, moreover, explains the agreement or harmony in the meaning of certain terminologies and utterances. The major problem as presented by Colomb (1997), however, lies in what can be called the fundamental conceptual disparity. Fundamental conceptual disparity occur when the terms used in

two different ontologies have meanings that are similar, yet not quite the same (Xu and Lee, 2002). For example, the use of the word 'departure city' and 'origin' in airline reservation ontologies can cause misapprehension. This is because the place of origin of an individual may not necessary be related to the departure city. It is possible that an individual can depart from a particular city but the said departure city may not be the origin.

Bishr (1998) in turn elaborated on schematic disparity, a phenomenon that crops up when information that is shown as data in one schema is shown in another as metadata (Bishr, 1998; Miller, 1998).

According to Lin et al. (2006), the problem of semantic disparity is extremely critical in situations of extensive cooperation and interoperation between distributed systems across different enterprises. In the case of digital forensics, for example, such a situation would make it difficult to manipulate distributed data or information from a cloud environment in a centralised manner. This is because the contextual requirements and the purpose of the information from a cloud computing environment, as well as across the different digital systems or different digital forensic tools being used, may not be homogeneous. It also implies that the cloud computing environment, the different digital systems or different digital forensic tools may not be composed of parts or elements that are all similar, hence making it hard to manipulate the data in a centralised manner.

Although the database perspective on semantic disparity is good and offers insights (Xu and Lee, 2002), it limits the understanding of semantic disparity and how to manage it in other domains. The next section identifies and explains in brief the potential causes of semantic disparity in digital forensics.

5.4 Potential Causes of Semantic Disparity in Digital Forensics

As mentioned earlier, semantic disparity may occur in digital forensics when the communicating parties (computer professionals, law enforcement agencies and other digital forensic practitioners) use different interpretations, descriptions and representations of the same or related domain data or information and terminologies. This causes variations (a lack of uniformity) in understanding the domain information and in how such information is specified and structured in different digital forensic application areas.

Semantic disparity can obviously hinder the progress of communication because it renders inexactness of meaning to the information delivered and may cause every different individual involved to reach his or her own different conclusion in the end. This implies that the communication process will not be complete, because the receiver may not understand the sender's message. For example, the use of the words ‘Digital Evidence’ and ‘Electronic Evidence as shown in Table 5.1 may cause ambiguity in the course of a digital forensic investigation process that involves people with varying backgrounds. This is because the descriptions of these terms as shown in Table 5.1 are in conflict with each other and can easily confuse the receiver of such information. Such confusion can further interfere with the intended objectives or intended meaning of these terms when used during a digital forensic investigation process.

Semantic disparity furthermore restricts effective communication between the sender and the receiver of information. Having the ability to identify and resolve semantic disparities in digital forensics can assist stakeholders such as investigators, computer professionals and law enforcement agencies in decision making and reasoning during digital forensic investigations. In the case where these collaborating stakeholders cannot communicate or engage in a conversation due to semantic barriers or conflicts encountered, the introduction of semantic reconciliation to overcome existing barriers or conflicts becomes vital. The various conflicts (including examples where applicable) that can cause semantic disparity in digital forensics are listed below.

- Semantic conflicts
- Descriptive conflicts
- Structural conflicts

These conflicts are discussed in the sub-sections to follow. Note, however, that the list is used only as common examples to facilitate this study and do not try to be an exhaustive list in any way.

5.4.1 Semantic Conflicts

Semantic conflicts occur when different people involved in the same domain do not perceive exactly the same set of real-world objects, but instead they visualise overlapping sets (Bishr, 1998). As a result, disagreement occurs about the interpretation, descriptions and representations of the same or related data or information. Examples of semantic conflicts as identified by Naiman and Ouksel (1995) include structural and

representational differences, mismatched domains, and naming conflicts of the reality being modelled. Table 5.1 shows examples of some of the semantic conflicts (descriptions and interpretation of terminologies) sampled from the digital forensic domain.

Table 5.1 Example of Semantic Conflicts Found in Digital Forensic Terminologies

DF Terminology	Semantic Conflicts
Digital Evidence	"Digital evidence encompasses any and all digital data that can show that a crime has been committed or can provide a link between a crime and its victim or a crime and its perpetrator" (Harley, 2003).
Digital Evidence	"Information of probative value stored or transmitted in digital form (SWGDE and IOCE, 2000) and may be relied upon in court. It can be found on a computer hard drive, a mobile phone, a personal digital assistant (PDA), a CD, and a flash card in a digital camera, among other places" (NIJ, 2015).
Digital Evidence	"Digital evidence includes information on computers, audio files, video recordings, and digital images. This evidence is essential in computer and Internet crimes, but is also valuable for facial recognition, crime scene photos, and surveillance tapes" (NSIT, 2015).
Digital Evidence	"Digital evidence of an incident is any digital data that contains reliable information that supports or refutes a hypothesis about the incident" (Carrier and Spafford, 2004).
Electronic Evidence	"Electronic evidence is any electronically stored information that may be used as evidence in a lawsuit or trial. Electronic evidence includes any documents, emails, or other files that are electronically stored. Additionally, electronic evidence includes records stored by network or

	Internet service providers" (Jessica, 2015).
Electronic Evidence	"Electronic evidence is any data or information stored in electronic format or on electronic media. For example, any recording made on tape (video or audio), computer floppy disk or compact disk is generally regarded as electronic evidence" (Nigel and Tim, 2000).

Infer from Table 5.1 that the terminologies in column one are meant to convey the same meaning to any receiver of such information. However, reading from the second column of the table, the descriptions are in conflict with each other, which can easily cause confusion or misapprehensions. The next section explains in brief the descriptive conflicts.

5.4.2 Descriptive Conflicts

Descriptive conflicts include naming conflicts due to homonyms and synonyms, as well as conflicts in respect of attribute, domain, scale, cardinalities, constraints, operations, etc. (Bishr, 1998; Sheth and Gala, 1989; Larson et al., 1989). In the case of digital forensics, descriptive conflicts can occur when two terminologies representing the same or related ideas of the domain concepts are described using different sets of properties. Table 5.2 presents some of the descriptive conflicts identified in the digital forensic domain. Infer from Table 5.2 that the terminologies in column one (i.e. Analysis) are identical, and yet their descriptions in column two are not the same. The same terminologies in column one are described differently in column two, hence the descriptive conflict which causes misapprehension.

Table 5.2 Examples of Descriptive Conflicts Found in Digital Forensics

DF Terminology	Descriptive Conflicts
• Analysis	"Determines significance, reconstructs fragments of data and draws conclusions based on evidence found. The distinction of analysis is that it may not require high technical skills to perform and thus more people can work on this case" (Reith et al., 2002).

• Analysis	"Analysis involves the use of a large number of techniques to identify digital evidence, reconstruct the evidence if needed and interpret it, in order to formulate an hypothesis on how the incident occurred, what its exact characteristics are and who is to be held responsible" (Valjarevic and Venter, 2012).
• Analysis	"The use of different forensic tools and techniques to make sense of the collected evidence" (Sibiya et al., 2012).
• Examination	"Examination is an in-depth analysis of the digital evidence and the application of digital forensic tools and techniques that are used to gather evidence" (Lalla and Flowerday, 2010).
• Examination	"An in-depth systematic search of evidence relating to the suspected crime. This search focuses on identifying and locating potential evidence, possibly within unconventional locations. Construct detailed documentation for analysis" (Reith et al., 2002).

(Note that the terminologies in Table 5.1 and Table 5.2 are merely selected as common examples to facilitate this study and should by no means be treated as an exhaustive list.)

The researcher found that the terminologies in Tables 5.1 and 5.2 are mostly used by digital forensic investigators and law enforcement agencies during and after a digital forensic investigation process has been conducted, hence the motivation for this study. Note from Table 5.1 that the conflict is between the different digital forensic terminologies (Digital evidence and Electronic evidence) used to mean or express the same thing – hence the semantic conflict – while in Table 5.2 identical digital forensic terminologies are described differently – hence the descriptive conflict.

5.4.3 Structural Conflicts

Structural conflicts occur when two or more people use the same model, but choose different constructs to represent common real-world objects (Lee and Ling, 1995). In the context of digital forensics, structural conflicts can occur when different domain members (investigators) use the same digital forensic investigation process model but choose different constructs to present their results or findings after an investigation

process has been conducted. Note that the term ‘constructs’ is used to mean ideas or theories containing various conceptual elements, and is considered to be subjective but not based on any empirical evidence (Houts and Baldwin, 2004). Structural conflicts problems can be solved by standardisation. Figure 5.1 (a) and (b) below shows an example of how structural conflicts can occur in digital forensics during a digital evidence presentation session in court or civil proceedings.

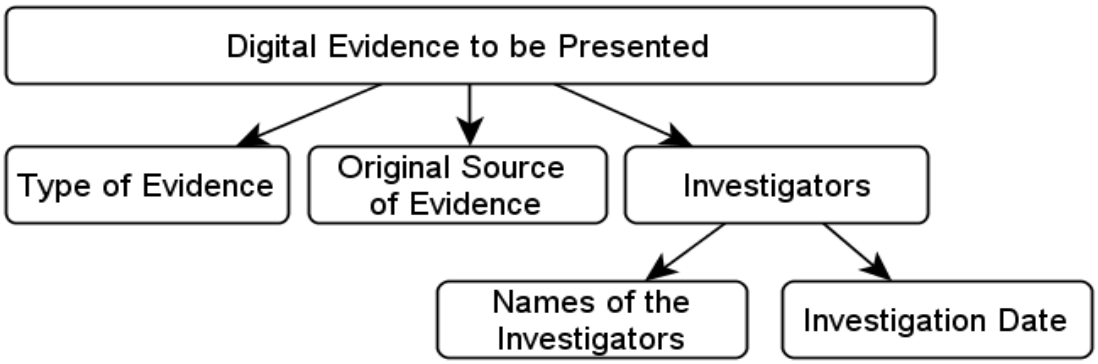


Figure 5.1 (a)

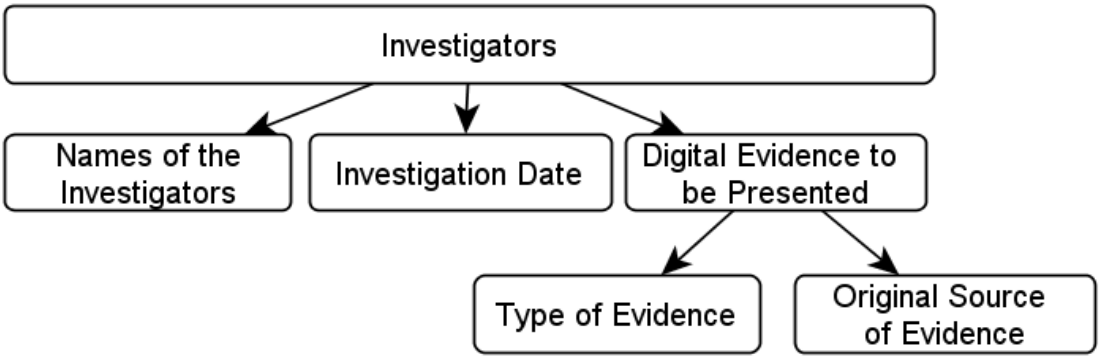


Figure 5.1 (b)

Figure 5.1(a) and (b) Example of Structural Conflicts in Digital Forensics

Infer from Figure 5.1(a) and (b) that the objects to be represented are all the same. However, the structure of presentation is different; this is sometimes because of the differences in the backgrounds of the investigators in charge of the whole process and hence the structural disparity.

After attending several sessions of expert testimony during digital evidence presentation in court and civil proceedings, the researcher found that different constructs were used by different digital forensic experts to convince the court that the potential digital evidence presented is worthy of inclusion in the criminal process.

However, an examination of the constructs used during digital evidence presentation revealed that they were based on experience rather than on standardised guidelines, procedures or digital forensic logics. This discrepancy is attributed to the lack of standardised guidelines or procedures in digital forensics for even making the most common representations of potential digital forensic evidence in court or at civil proceedings (Cohen, 2011). A discussion on how to manage semantic disparities in digital forensics is presented in the next section.

5.5 MANAGING SEMANTIC DISPARITIES IN DIGITAL FORENSICS

This research thesis recognises the importance of uniformity in the interpretation, description and representation of digital forensic data or information and terminologies.

However, managing semantic disparities in a growing field like digital forensics can be a daunting task. The situation is aggravated by the fact that the technological trends in digital forensics are ever-changing; new terminologies are constantly introduced into the domain and new meanings are assigned to existing terms (Karie and Venter, 2012). The technological change and domain evolution in digital forensics raise the need for developing dynamic methods and specifications with the ability to effectively assist in resolving semantic disparities. Such methods will further assist in establishing an efficient semantic reconciliation process in the domain.

Furthermore, the requirement for semantic reconciliation methods and specifications in digital forensics is essential for the betterment of the domain, the effective use of different domain terminologies and the representation of domain information. Having different approaches under which semantic disparity may be managed in digital forensics can be of great significance. The different approaches identified in this study that can help manage semantic disparities in digital forensics are explained in the next section.

5.6 APPROACHES TO MANAGE SEMANTIC DISPARITIES IN DIGITAL FORENSICS

Due to the fact that the digital forensic discipline is still evolving, managing semantic disparities can be a laborious task. However, according to Farshad and Andreas (2001), there exist different approaches that can aid in managing semantic disparities. As with other examples explained earlier, the bulleted list presented below merely contains examples to facilitate this study and should not be treated as an exhaustive list. Nevertheless, a comprehensive approach towards managing semantic disparity is

indispensable in digital forensics if a solution is desired. Semantic disparities can be managed in the following ways:

- By building ontologies and reasoning based on these ontologies
- Through semantics integration
- Through explicit use of common shared semantics
- By using a semantic reconciliation model

The sub-sections to follow briefly explain the above identified approaches for managing semantic disparities.

5.6.1 Managing Semantic Disparities by Building Ontologies and Reasoning Based on these Ontologies

Building ontologies in digital forensics can alleviate the problem of semantic disparity by providing formal, explicit definitions of data and by reasoning over related concepts. Ontologies in most cases capture the conceptualisation of experts in a particular domain of interest (Falbo et al., 1998). Ontology mapping, on the other hand, can also be employed in digital forensics to find semantic correspondences between similar elements of different ontologies, thus allowing people to agree on terms that can be used when communicating (Noy, 2004).

Building proper domain ontology in terms of its explication and in accordance with the conceptualisation of domain experts can also help in managing semantic disparities in digital forensics. According to Kajan (2013), however, considering that anyone can design ontologies according to his/her own conceptual view of the world; care must be observed during the process of designing ontologies because ontological disparity among different parties may easily become an inherent characteristic.

Moreover, using ontology in digital forensics, the domain experts can together share their standard understanding of the digital forensic domain structure. Ontologies can as well enable the reuse of expertise employed during digital forensic investigation processes (Ćosić and Ćosić, 2012). According to Farshad and Andreas (2001), the representation of ontologies and reasoning based on these ontologies makes it possible to capture and represent ontological definitions and important features that can be used in representing ontologies for reasoning. In the case of digital forensics, such an approach would help create clear and agreed-upon definitions of the different terminologies used in the domain.

Besides, such an approach can be of great value in managing semantic disparity in digital forensics because the relationships that hold among the different domain terminologies can be realised and structured. To further explore the representation of ontologies and reasoning based on ontologies, the reader is advised to consult Palmer (2001); Caloyannides (2004) and Crouch (2010) respectively.

More details on the concept of developing ontologies for digital forensics are explained in as part of Chapter 6 in this thesis. The next sub-section gives an explanation of how to manage semantic disparities through semantic integration.

5.6.2 Managing Semantic Disparities through Semantic Integration

Semantic integration deals with the process of interrelating information from diverse sources to create a homogeneous and uniform semantic of use (Noy, 2004). In the case of digital forensics, semantic integration can make communication between computer professionals and the legal community or law enforcement agencies easier by providing precise concepts that can be used to construct domain data or information and knowledge. In addition, semantic integration can facilitate or even automate communication between different systems, thus offering the ability to automatically link different ontologies (Gardner, 2005) in the digital forensic domain. Semantic integration is also explained further in Chapter 7 of this research thesis.

The next section explains how to manage semantic disparities through the explicit use of common shared semantics.

5.6.3 Managing Semantic Disparities through Explicit use of Common Shared Semantics

The explicit and formal definitions of semantics of terms have always guided many researchers to apply formal ontologies (Guarino, 1998) as a potential solution to semantic disparity. A traditional ontology usually is made up of logical axioms that communicate the meaning of terminologies for a certain domain (Bishr et al., 1999; Kottman, 1999). Moreover, traditional ontologies usually takes care of understanding the members of a particular domain and helps to reduce ambiguity in communication (Farshad and Andreas, 2001), interpretation, description and representation of domain information. Therefore, the explicit use of common shared semantics in digital forensics can be a stepping stone towards resolving semantic disparities in the domain.

The next section explains in brief how to manage semantic disparities in digital forensics using a semantic reconciliation model.

5.6.4 Managing Semantic Disparities using a Semantic Reconciliation Model

The concept of a semantic reconciliation model, also called the Digital Forensic Semantic Reconciliation (DFSR) model, is explained in detail in Chapter 7 of this research thesis. The DFSR model is an attempt towards developing a method that can be used for resolving semantic disparities in the digital forensic domain. The DFSR model has been designed to use semantic similarity measures for resolving semantic disparities. To implement the DFSR model, a method for computing the semantic similarity between the different digital forensic terms in question – known as the Digital Forensic Absolute Semantic Similarity Value (DFASSV) – is discussed and explained in Chapter 8.

The next section discusses the advantages of semantic reconciliation in digital forensics.

5.7 Advantages of Semantic Reconciliation in Digital Forensics

While many research activities have taken place in digital forensics, very few of them have been towards semantic reconciliation. Semantic disparity in any domain can alter the context as well as the purpose of any information delivered by an individual, hence the need for semantic reconciliation. Depending on the traditional knowledge-based approach, resolving semantic disparities can be very hard, if not next to impossible. Such knowledge, though, could be useful to define for example new digital forensic terms, especially when attempting to standardise terms in the field of digital forensics (Karie and Venter, 2012).

Methods and specifications therefore need to be developed in digital forensics to effectively assist in semantic reconciliation. Additionally, such methods and specifications should be capable of resolving current and future semantic disparities in the domain. This is because semantic reconciliation, as explained in this study, is a promising conception towards resolving semantic disparities in digital forensics. Semantic reconciliation in the digital forensic domain can yield the following advantages:

- Effective communication
- Common understanding
- Correct interpretation
- High levels of collaboration
- Uniform representation of domain information

- Faster harmonisation of information from different sources
- Less error during analysis of potential digital evidence information

These advantages are further explained in more detail in the sub-sections to follow.

5.7.1 Effective Communication

One of the barriers to effective communication in any domain is semantic disparities. Semantic reconciliation, on the other hand, can be used to bridge the semantic gap between different communicating parties, thus engendering effective communication in the domain (Parsons and Wand, 2003). Effective communication implies that information between the different digital forensic stakeholders (computer professionals, law enforcement agencies and digital forensic practitioners) is interpreted in such a way that the sender's desired effect is achieved. Semantic reconciliation in digital forensics seems essential if effective communication is to be achieved.

5.7.2 Common Understanding

Semantic disparities may arise in digital forensics as a result of inconsistent interpretation, description and representation of domain data or information and terminologies. This may include the use of different alternatives or definitions to describe the same domain data or information. Semantic reconciliation implies that different digital forensic experts can achieve a common understanding by reconciling the meaning of terms and creating a common interpretation, description and representation of domain terminologies (Parsons and Wand, 2003). Moreover, after a semantic reconciliation process, the meaning of information as interpreted by the receiver is aligned with the meaning intended by the sender (Anon, 2013b). In the case of court or civil proceedings, a common understanding can help different stakeholders to treat queries conveniently, while at the same time maintaining consistency in their understanding of the various digital forensic terminologies and data used during such proceedings.

5.7.3 Correct Interpretation

Whenever two or more independent digital forensic practitioners with varying professional backgrounds need to cooperate during a digital forensic investigation process, semantic conflicts may occur. It is therefore critical that semantic disparities be resolved to facilitate the correct interpretation of domain data or information during the investigation process. Semantic reconciliation can enhance correct interpretation through

detecting the semantic similarities between different terminologies and the data used by independent practitioners to describe or represent domain information (Parsons and Wand, 2003). Note that the concept of semantic similarity measures is discussed in more details in Chapter 8 of this research thesis. Semantic similarity measures numerically compute the degree of similarity or relatedness among different terminologies. In the case of this research study, the semantic similarity measures focus primarily on terminologies used in the digital forensic domain.

5.7.4 High Levels of Collaboration

Many organisations are increasingly promoting collaborations as an important feature of organisation management (Tschannen-Moran, 2001). However, effective collaboration demands both reasoning and effective communication. Thus, semantic reconciliation in digital forensics can lead to high levels of collaboration between computer professionals, law enforcement agencies and digital forensic practitioners. Besides, semantic reconciliation can help create uniformity in the use of data or information and terminologies in the digital forensic domain, thus facilitating cooperation and collaboration among experts.

5.7.5 Uniform Representation of Domain Information

In the case of potential evidence presentation in any court of law, information displaying numerous semantic variances can be semantically unreliable. Hence, semantic reconciliation can help create a uniform representation of domain data or information, and render the interpretation, description and representation of the domain information much easier and more accurate (Wang et al., 2005).

5.7.6 Faster Harmonisation of Information from Different Sources

Streamlined information management and processing have increasingly become significant within organizations, especially when they are merging. However, to realize semantic interoperability from one information system to another using different terminology, then the interpretation of the information that is being exchanged has to be harmonised across the systems (Ubbo et al., 2002). Whenever two different contexts do not use a uniform interpretation or description of the same information Semantic disparity may arise. Therefore, the adoption of semantic reconciliation for the explication of implicit and hidden knowledge is a promising method to curb the problem

of semantic disparity in digital forensics and can assist in faster harmonisation of data or information from different sources.

5.7.7 Less Error during Analysis of Potential Digital Evidence Information

In the case of a digital forensic investigation process, errors in the analysis and interpretation of evidence are more likely when semantic disparities occur – even more so when there are no standardised procedures or formal representation of domain information (Chaikin, 2006). Semantic reconciliation, on the other hand, will enable computer professionals, law enforcement agencies and digital forensic practitioners to agree on terminologies or keywords to be used in interpreting and representing certain key information in the case of a digital forensic investigation process. Semantic reconciliation will also help establish keyword structures so that the relationships between different terminologies are easily recognised. Finally, semantic reconciliation in digital forensics can enhance the analysis of potential digital evidence data or information. The next section presents the chapter conclusion.

5.8 CHAPTER CONCLUSION

The discussion in this chapter describes semantic disparities in the digital forensic domain. Advances in semantic disparity research and the potential causes of semantic disparity in digital forensics were also identified, presented and discussed.

In a bid to manage semantic disparities in digital forensics, this chapter also explains how to manage semantic disparity in digital forensics. In addition to the significance of resolving semantic disparities in the digital forensic domain, different approaches towards managing semantic disparities in digital forensics are also discussed. Advantages of semantic reconciliation in the digital forensic domain were also identified and explained.

Chapter 7 of this thesis will elaborate further on the digital forensic semantic reconciliation model as a way towards resolving the semantic disparities that occur in digital forensics. The next chapter (Chapter 6), however, explains the development of ontologies for the digital forensic domain.

CHAPTER 6 : DEVELOPING ONTOLOGIES FOR DIGITAL FORENSICS

6.1 INTRODUCTION

Ontology is a broad term that includes a wide range of activities, complexities and issues of fundamental and significant concern, such as the ontology development process. Nevertheless, ontologies have been widely used in different fields as techniques for representing and reasoning about domain knowledge (Van Rees and Amor, 2003; Shadbolt et al., 2006). In addition, ontologies can be used as a means of specifying and defining descriptions of concepts and their relationships. Such descriptions can further enhance the sharing of a uniform comprehension of the structure of information among domain experts (Brusa et al., 2006). Note that some concepts in this chapter were already explained in Chapter 4; however, they are repeated in the introductory part of this chapter for the purposes of flow and consistence.

According to Boyce and Pahl (2007), ontologies can be developed as a way to share a uniform understanding of the structure of information among entities in a bid to enable the reuse of domain knowledge and to make clear any assumptions about a domain that are usually implicit. If assumptions that underlie an implementation are made explicit in ontologies, then it is becomes easy to change the ontology when knowledge about the domain changes (Boyce and Pahl, 2007). Thus, developing ontologies that describe the uniform entities in which shared knowledge can be represented in digital forensics can help create uniformity and a common understanding in the domain. These characteristics can also enhance or improve cooperation among computer professionals, law enforcement agencies and other digital forensic practitioners in the case of an investigation process.

Developing ontologies in digital forensics can furthermore help to organise the domain knowledge better and to describe the domain information and semantics explicitly and in an ordinary way. Ontologies in digital forensics can be used to resolve some of the semantic disparities that exist in the domain. Moreover, in an expanding field such as digital forensics, developing ontologies that can provide direction in different areas of the domain (such as professional specialisation, certifications,

development digital forensic tools, curricula, as well as educational materials) are truly worthwhile.

In this chapter the researcher explains ontologies as one way to resolve semantic disparities as well as better organise digital forensic domain knowledge and explicitly and plainly describes the domain information, knowledge and semantics[2]. This also implies that ontologies, as presented in this chapter, are used to specify common vocabularies with which to make assertions, as well as analyse digital forensic domain information and knowledge. However, the primary reason for bring ontology into this study, as explained in this chapter , is to show how useful ontologies can be in resolving semantic disparities in digital forensics.

Section 6.2 explains related work on the development of ontologies for digital forensics. Sections 6.3 and 6.4 present examples of ontologies developed for digital forensics in this thesis, while Section 6.5 highlights the benefits of developing ontologies for digital forensics. A conclusion to this chapter is presented in Section 6.6.

6.2 RELATED WORK ON DEVELOPING ONTOLOGIES FOR DIGITAL FORENSICS

Very little literature on issues related to ontology development for the digital forensic domain was available at the time of writing this thesis. As a matter of fact, even what is present in literature seems to be somewhat varied. However, several ontologies have previously been proposed within digital forensics. Such ontologies have offered valuable contributions to the development of ontologies for digital forensics for inclusion in this research thesis.

As an example, Brinson et al. (2006) presented a detailed cyber forensics ontology in an effort to create a new way of studying cyber forensics. This ontology consists of a five-layered hierarchical structure with the final layer being specified areas that can be used for certification and specialisation. There ontology however, concentrates on new ways of studying cyber forensics and do not explore how to use such an ontology for resolving semantic disparities which is the focus of this study.

In a different effort, David and Richard (2007) introduced the concept of Small-Scale Digital Device Forensics (SSDDF) ontology. Their proposed ontology provides law enforcement agencies with the right knowledge concerning the devices found in the

[2] The contents of Section 6.3 of this chapter were published as a research paper by the Journal of Forensic Sciences, September 2014, Vol. 59, No. 5, while Section 6.4 was presented as a research paper at an international conference and later published in the Proceedings of the European Information Security Multi-Conference (EISMC 2013).

Small-Scale Digital Device (SSDD) domain. Additionally, they suggested that the ontology can be used as a way to supplement the development of a set of standards and methods with which to approach SSDD. This ontology still do not in any way present or explain how it can be used to resolve semantic disparities in digital forensics.

Ćosić and Ćosić (2012) points out the problems experienced by investigators in the pursuit of forensic investigations of digital devices, primarily because of misunderstanding or the false understanding of certain important ideas. They further propose an ontology of digital evidence as a possible method acceptable as a solution for this problem. However, in there paper they did not address the problem of semantic disparities and hence there ontology do not provide a solution to the semantic disparity problem experienced in digital forensics.

Allyson and Doris (2009) discussed the concept of 'Weaving Ontologies to Support Digital Forensic Analysis'. They argue that digital forensic analysis face several challenges. Although there are different techniques and tools to help with the analysis of digital evidence, they inadequately address key problems such as the vast volumes of data, lack of unified formal representation, standardised procedures, incompatibility among heterogeneous forensic analysis tools, lack of forensic knowledge reuse, and lack of sufficient support for criminal or civil prosecution. Allyson and Doris (2009) further suggest the applicability and usefulness of weaving ontologies to address some of these problems. They also propose an ontological method that can lead to the future development of automated digital forensic analysis tools. However, Allyson and Doris (2009) did not consider how there ontologies can be used to resolve semantic disparities in digital forensics which is the primary focus of this study.

Although other works on ontology development exist, neither that nor the cited references in this research study have so far presented ontologies as one way to resolve semantic disparities in digital forensics as explained in this chapter. The researcher nonetheless acknowledges the fact that the previous work on ontologies as discussed above has offered useful insights toward the development of ontologies in this thesis.

The next section presents a detailed explanation of the proposed ontology for the different digital forensic disciplines and sub-disciplines in this thesis.

6.3 A General Ontology for Digital Forensic Disciplines

Ontology development as discussed by (Brusa et al., 2006) can be categorised into two phases: first is the specification phase and second is the conceptualisation phase. In this

section the researcher therefore present a detailed explanation of the ontology developed for the different digital forensic disciplines and sub-disciplines based on the specification phase and the conceptualisation phase discussed by Brusa et al. (2006). The primary objective of the specification phase is to gather knowledge about the different digital forensics disciplines and sub-disciplines, while the aim of the conceptualisation phase is to organise and structure the acquired knowledge about the disciplines and sub-disciplines by using external representations.

Note that, different research methods were used during the specification phase to acquire knowledge about the different digital forensics disciplines and sub-disciplines. Some of the methods that were found to be helpful during the specification phase of this research study are explained below:

6.3.1 Literature surveys

A literature survey involves reviewing all readily available materials (StatPac, 2015). Literature survey was thus found to be a good and an inexpensive method of gathering information. For this reason, the researcher took time to go through very many published materials both on scientific journals as well as over the web. This is evident from the referenced materials at the end of this study. Literature surveys helped to acquire the much needed knowledge that was used to develop the ontologies in this study.

6.3.2 Personal Interviews

The researcher used personal interviews as a way to gather in-depth and comprehensive information about the different digital forensics disciplines and sub-disciplines. The information gathered from the interviews was first recorded on paper and later used in the development of the ontology presented in this study (Karie, 2014).

6.3.3 Talking with People

The researcher used this method to talk to different digital forensic practitioners as a way to get information during the initial stages of this research study. Different people contributed differently to the knowledge on the ontology development process as well as the terminology used to build the ontology shown in Figure 6.1. The information gathered was recorded on paper and used to develop the ontologies in this study (Karie, 2014) Talking to people proved to be very helpful and an inexpensive way of gather information.

Figure 6.1 shows the structure of the ontology developed for the different digital forensic disciplines and sub-disciplines. Note that due to the small font size of Figure 6.1, Figures 6.2 to 6.7 show enlarged extracts of the entire ontology as represented in Figure 6.1.

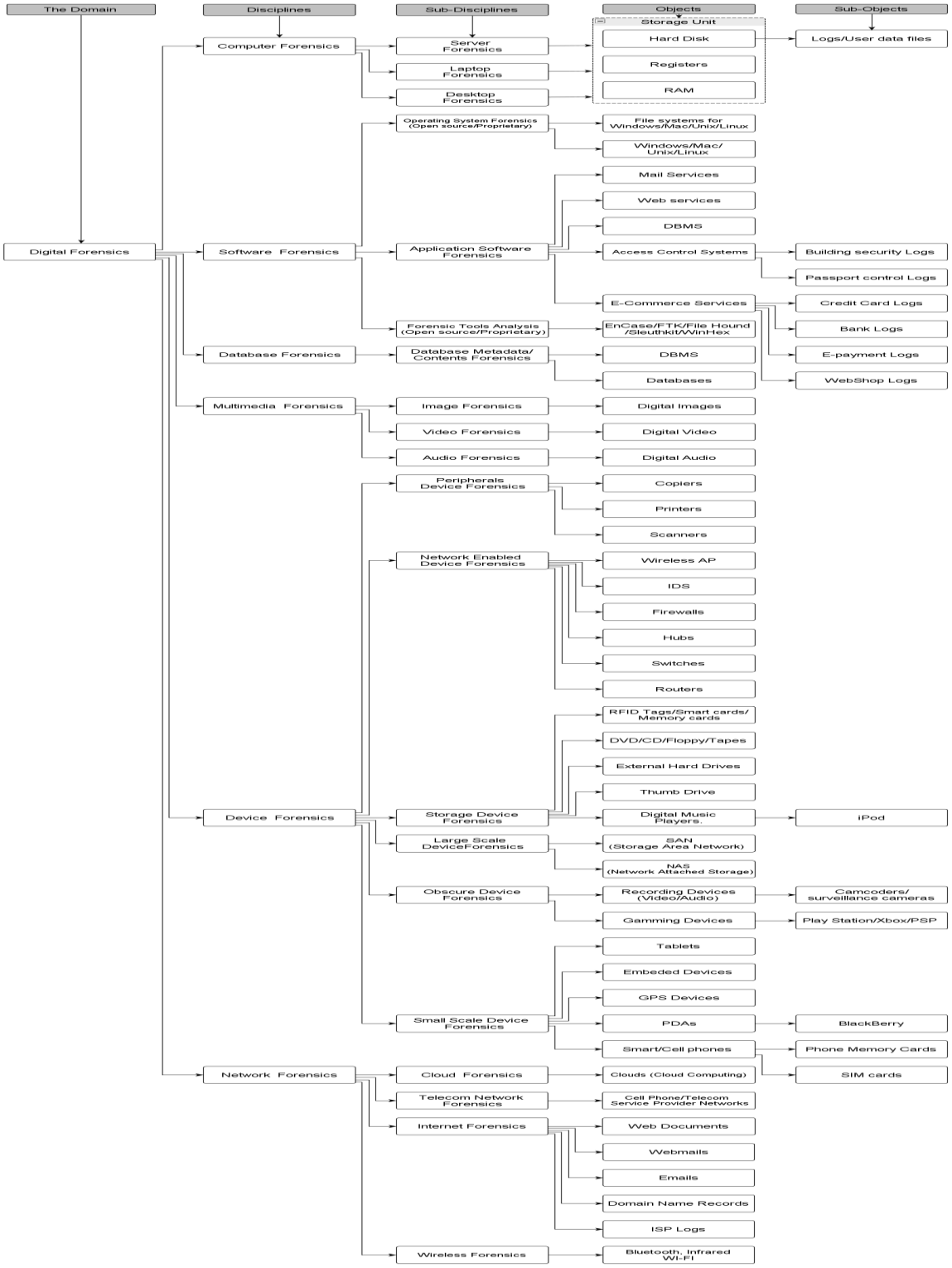


Figure 6.1 Ontology for Different Digital Forensics Disciplines and Sub-disciplines

The ontology consists of five layers arranged from left to right and the first layer depicts the main domain of focus (i.e. digital forensics). This is followed by the digital forensic disciplines in the second layer, and the sub-disciplines within the digital forensic domain in the third layer. Objects and sub-objects are introduced in the fourth and fifth layers of the ontology as a way of representing individual and specific finer details of the sub-disciplines within digital forensics. Organising the ontology into disciplines, sub-disciplines, objects and sub-objects was necessary so as to simplify the understanding of the ontology as well as to present specific finer details of the ontology.

The sub-disciplines, objects and sub-objects presented in the ontology focus more on areas that can be considered for professional specialisation and certification, as well as for the development of digital forensic tools, curricula and educational materials. Due to the limitations of the different research methods used in this study, infer from the ontology in Figure 6.1 that the objects and sub-objects listed were merely selected as common examples to facilitate this study and should not be treated as an exhaustive list. More sub-disciplines, objects and sub-objects can and should be added as the need arises in future.

From the ontology in Figure 6.1 it seems that some of the objects presented do not have sub-objects. In the researcher's opinion, breaking them down to a finer-grained level would be superficial at this stage. However, in future it should be possible to mention sub-objects that can be incorporated under the applicable objects, especially when developing curricula and education materials. The major digital forensic disciplines explored in this study (with their details as shown in Figure 6.1) include computer forensics, software forensics, database forensics, multimedia forensics, device forensics, and network forensics.

For the purpose of this research thesis, computer forensics is divided into server forensics, laptop forensics and desktop forensics, while software forensics focuses on application software forensics; operating system forensics (open source and proprietary) and forensic tools analysis (open source and proprietary).

Database forensics concentrates on database contents and database metadata, while multimedia forensics is divided into digital image forensics, digital video forensics and digital audio forensics.

Device forensics is divided into peripheral device forensics, network-enabled device forensics, storage device forensics, large-scale device forensics, small-scale

device forensics and obscure device forensics. Finally, the ontology concludes with network forensics, which is divided into cloud forensics, telecom network forensics, internet forensics and wireless forensics.

In the sub-sections that follow, the digital forensic disciplines and sub-disciplines as identified in the ontology in Figure 6.1 are explained in more detail.

6.3.4 Computer Forensics

According to Crouch (2010), computer forensics is a branch of digital forensics that uses analysis techniques to gather potential evidence from desktops, laptops and server computers for investigating suspected illegal or unauthorised activities. More precisely, computer forensics pays more attention on uncovering potential digital evidence after a security incident has taken place (Anon, 2012).

Note that we refer to 'potential' evidence in this chapter, since digital artefacts are only considered to be 'evidence' mostly in the final phases of the digital forensic investigation process, namely the reporting phase. This also implies that, for the collected potential evidence to be considered as competent evidence (Ryan and Shpantzer, 2005), it must possess scientific validity grounded in scientific methods and procedures.

The potential evidence gathered in most cases is usually found stored on the computer's internal storage unit as shown in Figure 6.2, which includes the hard disk that also stores operating system data (e.g. log files) and application or user data (e.g. word processing files). Computer forensics also considers the value of data that may be lost by powering down a computer, and thus collection of potential evidence can be conducted while the system is still running (Crouch, 2010), for instance from the Random Access Memory (RAM) or registers.

The primary aim of computer forensics is to conduct an organize investigation while maintaining a documented chain of evidence that can withstand legal scrutiny in a court of law, whether for a criminal or civil proceeding (Crouch, 2010). For the purpose of this thesis, the areas covered under computer forensics include server forensics; laptop forensics and desktop forensics (see Figure 6.2).

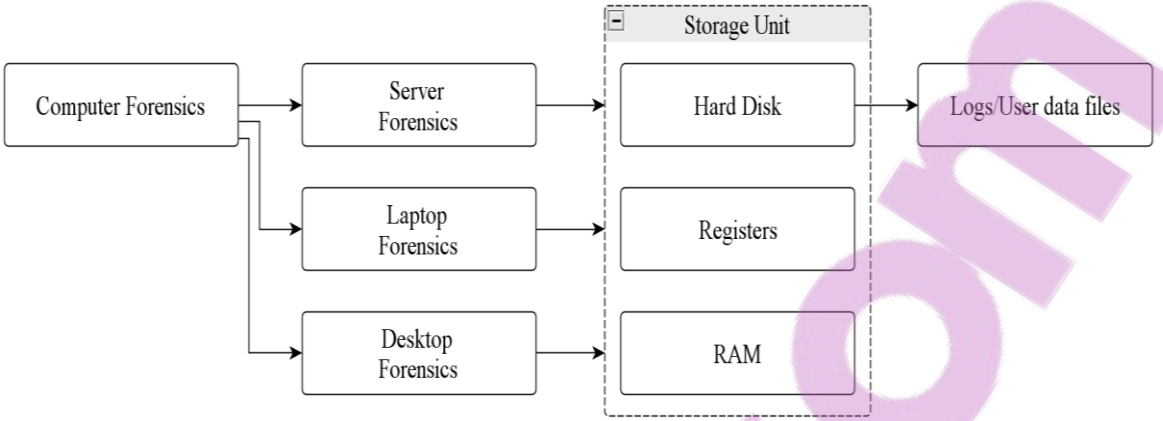


Figure 6.2 Computer Forensics

6.3.4.1 Server Forensics

In a network environment, a server is usually that powerful computer that is dedicated to managing mass system and user resources. Server forensics therefore focuses on finding digital evidence that is stored within the server machine (Obialero, 2003). In essence, server forensics deals with finding potential evidence in the same way that potential evidence is found on a desktop or laptop computer, the only difference being the significantly larger storage and somewhat different access capabilities to be dealt with on a server computer.

6.3.4.2 Laptop Forensics

Laptop forensics is dedicated to finding digital evidence from laptop computers. Laptops are designed to be light and mobile. Because of their mobile nature, laptops are popular computing systems and high contenders for hosting potential evidence. The hardware in a laptop is typically custom built for that particular model. According to Pierce (2003), very few components follow any given industry standard. This issue particularly complicates the process of digital forensic analysis on laptops and therefore laptops should be handled by a specialist who understands its configuration. Nevertheless, laptop forensics still forms part of computer forensics.

6.3.4.3 Desktop Forensics

Desktop forensics is meant to find digital evidence from desktop computers once a security incident has occurred. Since there are so many different ways to classify computers (Brinson et al., 2006), the ones discussed above (server, laptop and desktop) serve as examples to facilitate this study. With the advancements in technology, it should sooner or later be necessary to add other items to this category.

6.3.5 Software Forensics

Software forensics is a discipline concerned with uncovering potential evidence through examining software. According to MacDonell et al. (1999), software forensics is also a research area that strives to conduct investigations into all characteristics of computer program authorship by handling pieces of program source code as linguistically and stylistically analysable objects. Software forensics can be used to detect for example plagiarism in academia where students' assignments can be compared to see if some are 'suspiciously similar' (MacDonell et al., 1999; Whale, 1990).

According to Hanks et al. (2002), incidents and accidents that can be connected with software failure many a time result in disasters and other losses. The demand to learn from these events turns out to be more critical as software systems become more composite and the ways they can fail become less inherent.

Moreover, Johnson (2000) and Johnson (2002) argue that existing software development methods do not provide clear access to retroactive information about the composite and systemic causes of incidents and accidents. What is known from forensic engineering generally, as well as the study of failure, has yet to be applied comprehensively to software (Johnson, 2002). Software forensics (also known as software forensic engineering) can therefore be used to address such deficiencies.

A vast number of computer programs (software) are also available on the software market today. However, for the purpose of this research thesis we considered only a few. The reader is thus advised to consider other software as well, especially when developing curricula and education materials.

The list of software used in this study serves as examples and should not be perceived as an exhaustive list. For the purpose of this study, software forensics covers operating system forensics, application software forensics and digital forensic analysis tools, as shown in Figure 6.3.

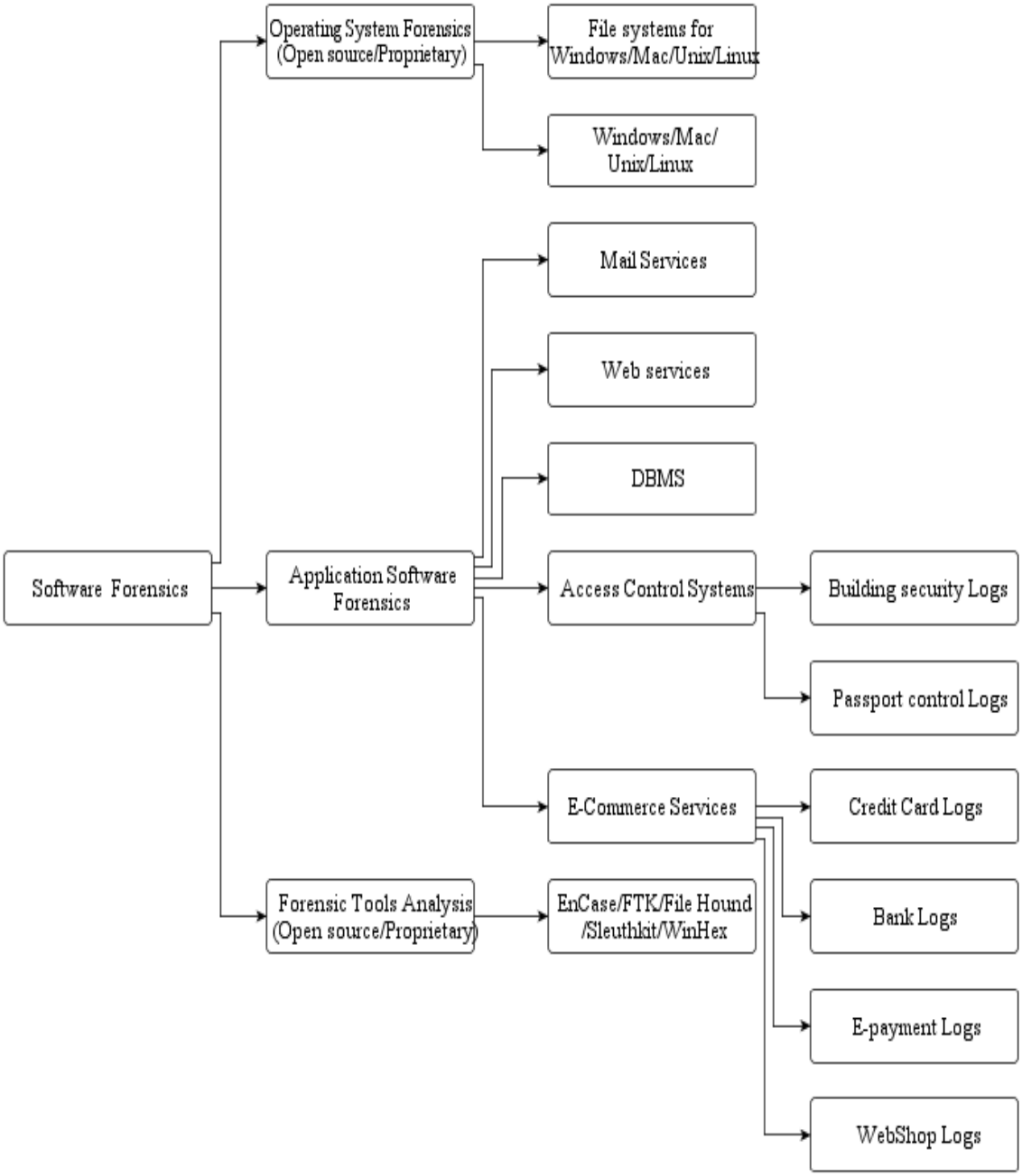


Figure 6.3 Software Forensics

6.3.5.1 Operating System Forensics

The operating system (OS) serves as the primary software installed on any computer system. The OS is also and often perceived as part and parcel of the entire computer system. Thus, in the case of a digital investigation, the investigator should be aware of the fact that many different operating systems exist, and that each has its own associated file structures. If the investigator knows in advance what particular operating system needs to be dealt with, he/she is able to search for and locate any potential digital evidence more efficiently and effectively (Brinson, 2006).

Operating systems may be categorised as open source or proprietary. Among the common and widely known operating systems are Windows, Apple Mac, Unix and Linux, and an investigator should be acquainted with these operating systems and their different file systems in particular.

6.3.5.2 Application Software Forensics

Application software is basically designed to help end users perform specific tasks. They either come bundled together with the computer system or can be purchased separately and installed later on the system. Application software forensics focuses on analysing and retrieving potential evidence from application software such as e-mail services, access control systems (e.g. building security logs and passport control logs), web services, database management systems, and e-commerce services (e.g. credit card logs, bank logs, e-payment logs and web shop logs) (see Figure 6.3).

6.3.5.3 Forensic Tools Analysis

There are many different open source and proprietary digital forensic tools available for use during digital investigations. Some of the commonly known tools include Encase (Guidance Software, 2012), Forensic Toolkit (FTK) (AccessData, 2012) and the Sleuth Kit (TSK) (Sleuth Kit, 2012). These tools are designed to perform a collection of digital forensic investigation functions and would basically include most of the investigation techniques applied during a digital investigation process.

However, there exist other digital forensic investigation tools that perform more elementary investigation functions such as WinHex, which is essentially a universal hexadecimal editor. Such a utility is particularly helpful in viewing any data in its raw form in order to perform low-level data analysis. X-Ways Imager is yet another example of such an elementary tool, which is basically a forensic disk imaging tool (X-Ways, 2012).

6.3.6 Database Forensics

Database forensics as explained by Olivier (2009) and Weippl (2009) focuses on databases and their related content and metadata. Most business's critical and sensitive information, e.g. bank accounts and medical data, is usually recorded and stored in databases. Unlawful disclosure, modification or theft of such data can be harmful to organisations. Therefore, database forensics aims at investigating unlawful disclosure,

modification or theft of data within a database in a bid to track down any perpetrators with such malicious intent (Weippl, 2009). An investigator's understanding of database concepts and how to use database management systems (DBMS) is clearly of crucial importance to database forensics, as shown in Figure 6.4.

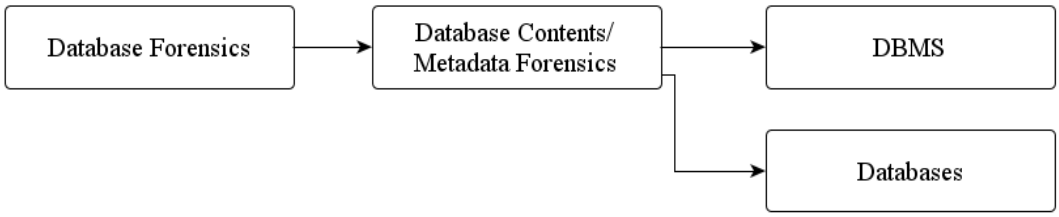


Figure 6.4 Database Forensics

6.3.7 Multimedia Forensics

In modern digital age, the creation and manipulation of digital images, videos and audio have been simplified through digital processing tools that are effortlessly and extensively available (ISIS, 2012). Such tools may include but are not limited to Adobe Photoshop CS6 (Adobe Photoshop, 2012), Adobe Premiere Pro CS6 (Adobe Premiere, 2012) and Pinnacle Studio (Pinnacle Studio, 2012). Adobe Photoshop CS6 is mostly used for picture and photo editing, while Adobe Premiere Pro CS6 and Pinnacle Studio are typically used for video editing. This implies that the authenticity of images, videos and audio can no longer be taken for granted (ISIS, 2012).

According to Böhme et al. (2009), questions regarding media credibility are of increasing importance and of particular interest in court, where consequential decisions might be derive from evidence in the form of digital media. Multimedia forensics can be used to uncover the authenticity information of captured images, videos and audio files. Such information can also serve as potential evidence in any court of law or civil proceedings. The main areas covered by multimedia forensics in this study include digital image forensics, video forensics and audio forensics – as shown in Figure 6.5. Image forensics, video forensics and audio forensics are explained briefly in the sub-sections that follow.

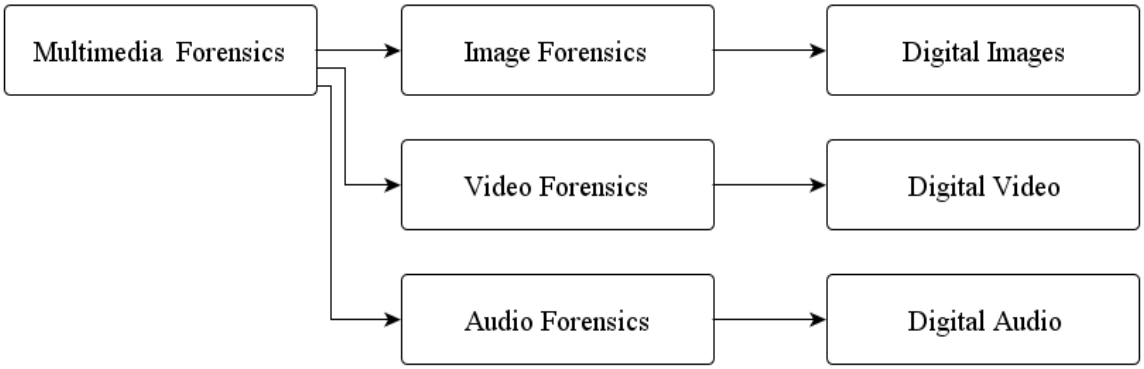


Figure 6.5 Multimedia Forensics

6.3.7.1 Digital Image Forensics

Digital image forensics is concerned with uncovering potential digital evidence found within digital images (ISIS, 2012). This may include digital evidence such as image origin (often referred to as image file type identification), image source identification and image forgery detection (Swaminathan et al., 2006). Digital image forensics can also be used to verify the authenticity of images (Gloe et al., 2007; Swaminathan et al., 2008).

6.3.7.2 Digital Video Forensics

Digital video forensics, like digital image forensics, is concerned with uncovering potential digital evidence found in video files. With the emergence of high-quality digital video cameras and complicated video-editing software, it is becoming increasingly easier to make unauthorized alterations with digital video (Wang and Farid, 2007). Digital video forensics can be used to good effect to detect cloning or duplicating frames, or even parts of a frame when people or objects have been removed from a video (Wang and Farid, 2007; Frederic et al., 2009; Stamm and Liu, 2011).

6.3.7.3 Digital Audio Forensics

Digital audio forensics may be defined as the application of audio science and technology in a bid to conduct an investigation and establish the truth in criminal or civil courts of law. Digital audio forensics is meant to uncover potential digital evidence about audio files. This may include, for example, environment recognition from digital audio files (Muhammad and Alghathbar, 2011). Environment recognition is used to refer to the physical environment under which digital audio samples were recorded. Audio

forensics can also be used to determine what kind of microphones were used (Kraetzer et al., 2007).

6.3.8 Device Forensics

Device forensics deals with the gathering of digital evidence from different types of devices. Devices may range from small-scale devices such as mobile phones, Personal Digital Assistants (PDAs), printers, scanners, cameras, fax machines (CDAC, 2012) among others, to large-scale devices such as the SAN (Storage Area Network) and NAS (Network Attached Storage) systems. The number of devices in this discipline of digital forensics is increasing daily and in the researcher's opinion, this is the motivation for considering device forensics as a separate and vast discipline of the digital forensic domain.

For the purpose of this study, device forensics is divided into peripheral devices, network-enabled devices, storage devices, large-scale devices, small-scale devices, and obscure devices – as shown in Figure 6.6. The list presented should not be considered as exhaustive as most new digital devices could well be categorised within this discipline of digital forensics.

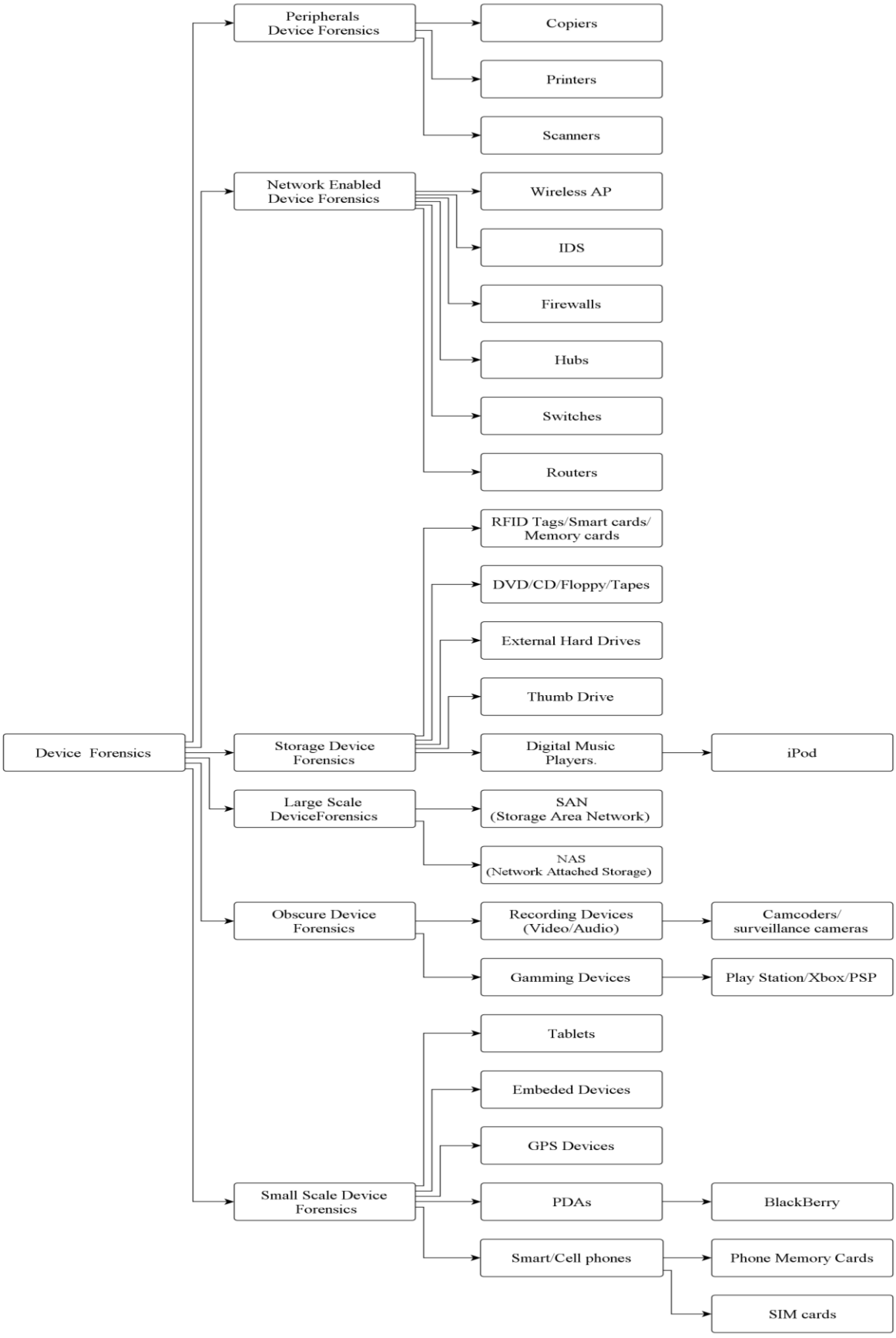


Figure 6.6 Device Forensics

6.3.8.1 Peripheral Devices

Peripheral devices are normally used to expand a system's capabilities. However, they do not actually form part of the core computer architecture. In addition, peripheral devices vary greatly and can range from external to internal peripherals. Examples of external peripherals may include the mouse, keyboard, printer, monitor and scanner, among others. Internal peripheral devices (often referred to as integrated peripherals), on the other hand, may include devices such as a CD-ROM drive and internal modems. A thorough analysis of peripheral devices can reveal much information that is of potential value to a digital forensic investigator.

6.3.8.2 Network-enabled Device Forensics

With the development of network and telecommunication technologies, communication infrastructure has rapidly spread in many sectors of the industry. As a result, various network-enabled devices with Ethernet and Transmission Control Protocol/Internet Protocol (TCP/IP) communication functions can be found in different practical applications (Sena, 2012). Such devices may include Intrusion Detection Systems (IDSs), firewalls, hubs, switches, routers and wireless access points (to mention a few). Some of the network-enabled devices have the ability to store data and information, and therefore such information can serve as potential evidence during an investigation.

6.3.8.3 Storage Device Forensics

A storage device is any hardware device that has been specifically designed to store data or information. Storage devices can be primary to a computer (e.g. the RAM) or they can be secondary (e.g. DVD, CD, tapes, Radio-Frequency Identification (RFID) tags, smart cards, memory cards (flash drives) and external hard drives). Such devices can contain valuable potential evidence in the case of an investigation. Hence, investigators should be aware of the different capabilities supported by different storage devices.

6.3.8.4 Large-scale Device Forensics

Nowadays, investigators and analysts increasingly have to deal with large (terabyte-sized) data sets when conducting digital investigations (Jee et al., 2008). Such large data sets are mostly found stored on large-scale devices such as the SAN (Storage Area Network) and NAS (Network Attached Storage) systems. With the evolution in large-scale storage systems technology, it is possible that petabyte storage will soon replace terabyte-sized devices (Aberdeen, 2012). Petabyte-sized storage is considered the newest

frontier in the ever-growing world of data storage devices (Aberdeen, 2012). Therefore, investigators need to know how these devices operate in order to be able to effectively gather potential digital evidence. Like any other device, large-scale devices can provide potential evidence that can be used in a court of law or civil proceedings.

6.3.8.5 Small-scale Device Forensics

Small-scale devices, as the name suggests, are small and versatile. Moreover, the proliferation of hand-held digital devices has captured the majority of the market and is primed to become the next frontier in technology (David and Richard, 2007). Thus, a clear understanding of how these devices operate is necessary to adequately preserve, identify and extract useful information during a digital forensic investigation (Brinson et al., 2006). Examples of small-scale devices include but are not limited to tablets, embedded devices, Global Positioning System (GPS) devices, Personal Digital Assistants (PDAs) and mobile (smart) phones.

Mobile phones, for example, are becoming a focus of attraction in digital forensic investigations due to the feature-rich versatility of these devices. When dealing with mobile phone device forensics, some of the main artefacts of interest that may contain potential evidence are SIM (Subscriber Identity Module) cards and memory cards, of which the latter may be built in (on-board).

6.3.8.6 Obscure Device Forensics

Obscure devices are those devices that, in the opinion of the researcher, cannot be classified under any of the other sub-disciplines of device forensics. Such devices have the ability to store data or information that may possess evidentiary value in a digital forensic investigation process. Examples of obscure devices may include digital recording devices (video and audio) such as camcorders, surveillance cameras, gaming devices such as Sony's Play Stations, Microsoft's Xboxes and Nintendo's Wii consoles, which can also be analysed for potential digital evidence.

6.3.9 Network Forensics

According to Palmer (2001), network forensics "is a branch of digital forensics that basically uses scientific proven techniques to collect, use, identify, examine, correlate, analyse, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned

intent, or measured success of unauthorised activities meant to disrupt, corrupt, and/or compromise system components as well as providing information to assist in response to or recovery from these activities".

Unlike other branches of digital forensics, network forensics deals with volatile and dynamic information that can easily get lost after transmission in any network environment. An intruder might be able to wipe off all log files on a compromised host and therefore network-based evidence may be the only evidence available for forensic analysis (Hjelmvik, 2012). For the purpose of this study, network forensics is divided into cloud forensics, telecom network forensics, internet forensics and wireless forensics – as shown in Figure 6.7.

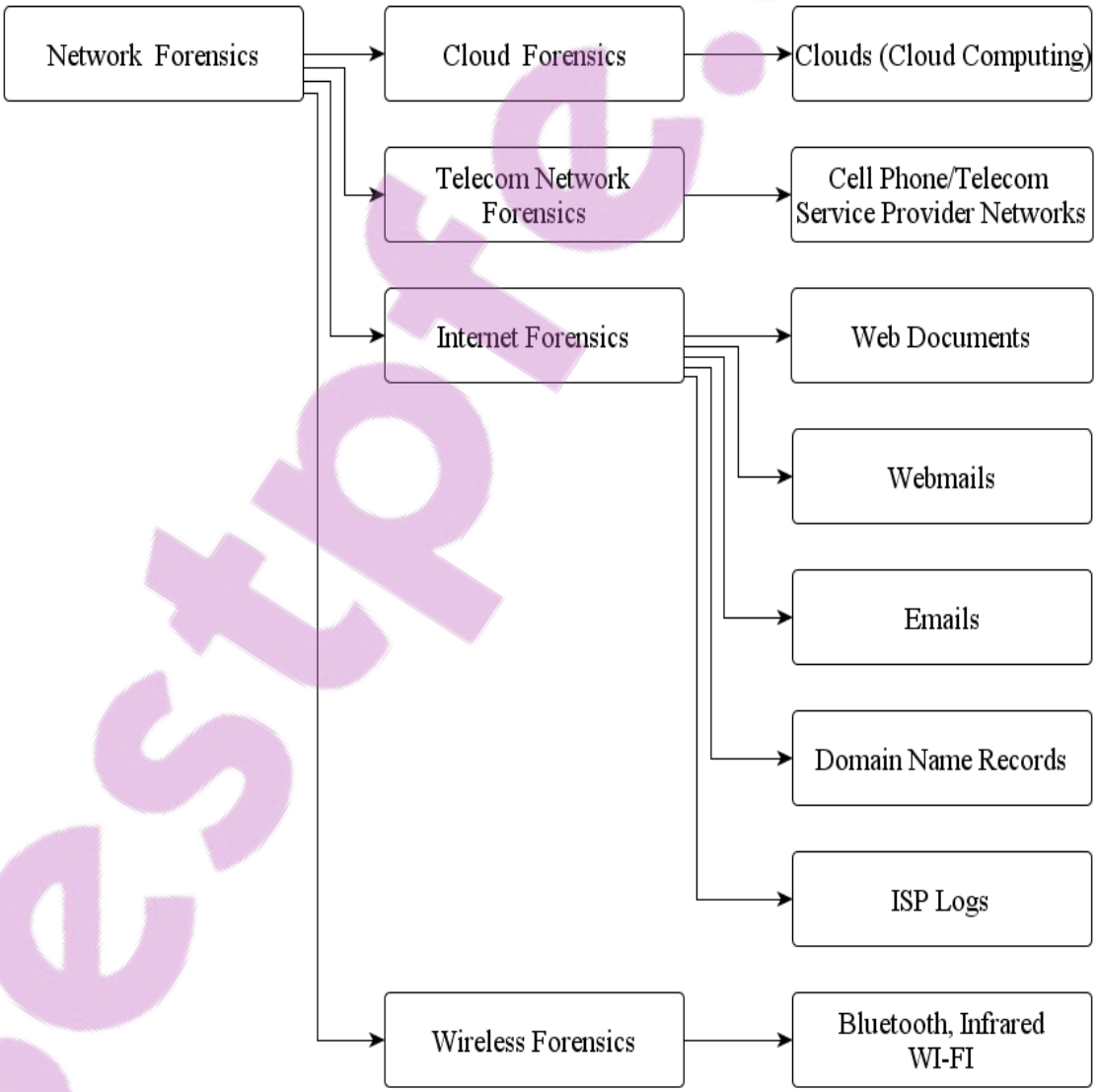


Figure 6.7 Network Forensics

6.3.9.1 Cloud Forensics

Cloud computing is reckoned to be one of the most transformative technologies in the history of computing. This is so because it is radically changing the way in which information technology services are produced, delivered, accessed and managed (Ruan et al., 2011). Cloud forensics as defined by Ruan et al. (2011) is an emerging field that deals with the application of digital forensics techniques in cloud computing environments and it is a sub-set of network forensics. Technically, cloud forensics follows most of the main phases of network forensic processes, with extended or novel techniques tailored for cloud computing environments in each phase. For this reason, the researcher placed cloud forensics as a sub-discipline of network forensics in the proposed ontology as shown in Figure 6.7. However, with the development in technology and the evolution in digital forensics some of the sub-disciplines (like cloud forensics) may in the near future be considered as standalone disciplines. This is because the cloud forensic technologies and the cloud environments are also growing at a faster rate.

6.3.9.2 Telecom Network Forensics

Telephones are often used to facilitate criminal and terrorist acts. The signalling core of public telephone networks generates valuable data about phone calls and calling patterns that may be used in criminal investigations, especially with the widespread uptake in voice-over-IP (VoIP) systems. Unfortunately much of this data is not maintained by service providers and is therefore unavailable to law enforcement agencies (Moore et al., 2005). If such data can be collected and stored, it can be analysed forensically and greatly facilitate the prosecution of criminals in court or in civil proceedings.

6.3.9.3 Internet Forensics

With the evolution in global commerce, many business organisations store vital business information online and carry out business transactions over the internet. Such organisations are under constant threat of falling victim to internet attacks. Moreover, because the internet is huge and unregulated, it has become a fertile ground for all types of cyber-crimes (Jones, 2005). If the internet is to become a safe platform for transacting business, internet forensics has to become very important as well.

Internet forensics is a research field that deals with the analysis of activities that occurred on the internet. It aims to uncover hints about people and computers involved

in internet crime, most notably fraud (e.g. credit card fraud) and identity theft (PCMag, 2012). Note that the terms 'internet crime' and 'cyber-crime' are often used interchangeably (Kowalski, 2002). Cyber-crime is usually used to refer to any criminal activity in which a computer or network is the source, tool, target or place of crime (Prasanna, 2012; Singh, 2012). The Cambridge English Dictionary, however, defines cyber-crimes as crimes committed with the use of computers or relating to computers, especially through the internet (Prasanna, 2012).

As a result, internet forensics tries to uncover the origins, contents, patterns and transmission paths of e-mail and Web pages, as well as browser history and Web servers' scripts and header messages (PCMag, 2012). It can also be used to take out information that lies hidden in every e-mail message, web page and web server. Such information may contain potential digital evidence that can be analysed for forensic purposes. In this research thesis, the researcher listed the following areas under internet forensics as common examples: Web-mail, e-mail, domain name records, Internet Service Provider (ISP) logs and web documents. However, there is much more that can be gathered from the internet as compared to what is listed in this section.

6.3.9.4 Wireless Forensics

The adoption of wireless technologies by different organisations in recent years has created concerns about control and security. Incident handlers and law enforcement have been forced to deal with the complication connected with wireless technologies when managing and responding to security incidents (Siles, 2012). For this reason, wireless forensics that has emerged as a result of wireless technologies focuses on capturing or collecting digital evidence data that propagates over a wireless network medium.

Wireless forensics also tries to make sense of the collected digital evidence in a forensic capacity so that it can be presented as valid digital evidence in court. The evidence collected can correspond to plain data, but can include voice conversations as well (Siles, 2012).

In line with the main focus of this research thesis, the section to follow introduces another example of ontology for a cloud forensic environment. The ontology can be used to specify common vocabularies with which to make assertions, as well as analyse digital forensic domain information and knowledge. In addition, the ontology can be

used to organise the domain knowledge better and to describe the domain information and semantics explicitly and in an ordinary way.

6.4 AN ONTOLOGY FOR A CLOUD FORENSIC ENVIRONMENT

With the emergence of cloud computing technologies, cloud forensics has become essential. Moreover, with the widespread and continued deployment of internet-based applications and network-enabled devices aimed at supporting mechanisms for cloud computing, the cloud environments and components can potentially be rendered incomprehensible.

In this section of the thesis, the researcher presents an ontology for a cloud forensic environment in an attempt to provide a structured depiction of the different cloud environments (cloud deployment models) and cloud components (cloud service models) with which investigators should be well-versed in the case of an investigation process involving the cloud. The ontology in this section is, however, not completely new to cloud forensic experts. Such an ontology was, however, developed as a means to share a common understanding of the structure of information among cloud environment entities in a bid to enable the reuse of domain information and to make explicit those assumptions about cloud forensics that are normally implied. The ontologies presented in Sections 6.3 and 6.4 of this chapter are part of the contributions by the current research towards creating a unified formal representation of digital forensic domain knowledge and information.

The ontology presented in this section shows, for example, the relationships and interactions between the different cloud environments and the cloud components. Such a simplified ontology can help investigators to comprehend the cloud environment and components with less effort. This also makes the interpretation, description and representation of data or information in a cloud environment simple enough for many stakeholders to understand with ease. Figure 6.8 shows the structure of the proposed ontology for a cloud forensic environment showing how ontologies can be used to share common understanding of the structure of information in digital forensics. Due to the small font size of Figure 6.8, Figures 6.9 to 6.11 contains enlarged extracts of the ontology as depicted in Figure 6.8.

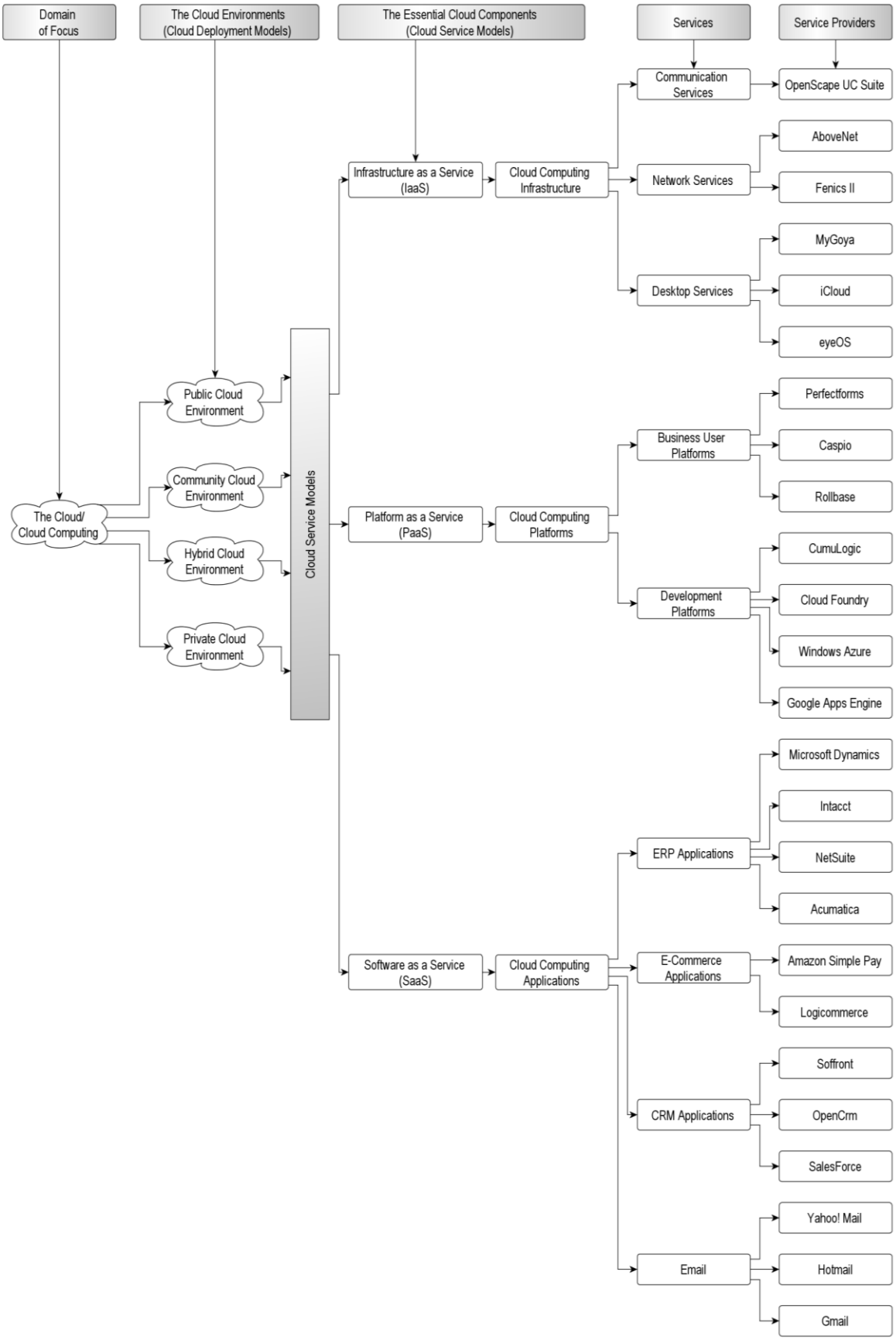


Figure 6.8 Cloud Environments and Essential Cloud Components

The ontology consists of five layers arranged from left to right and with the first layer depicting the main domain of focus (i.e. the cloud or cloud computing). This is followed by the cloud environments in the second layer and the essential cloud components in the third layer. Services and service providers are introduced in the fourth and fifth layers of the ontology as a way of representing individual and finer-grained details of the essential cloud components also referred to as cloud service models.

Note that cloud service models enable software platform and infrastructure to be delivered as services. The term 'service' is used to show the fact that they are given on demand and are paid for, on a usage basis (Czarnecki, 2011).

In the researcher's experience, organising the ontology into the particular cloud environments, essential cloud components, services and service providers, was necessary to simplify the understanding of the ontology. The services and service providers listed in the fourth and fifth layers of Figure 6.8 were selected as common examples to facilitate this study and do not constitute an exhaustive list.

The major areas explored (with their details as shown in Figure 6.8) include the cloud environments, the essential cloud components, services and the service providers. For the purpose of this study, the cloud environments (cloud deployment models) are divided into public cloud environment, private cloud environment, community cloud environment and hybrid cloud environment.

The essential cloud components (cloud service models), on the other hand, are divided into Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) (Czarnecki, 2011). However, infer from Figure 6.8 that IaaS, PaaS and SaaS are accessible through cloud computing infrastructure, cloud computing platforms and cloud computing applications respectively. In the sub-section to follow, the cloud environments as identified in Figure 6.8 are explained in more detail.

6.4.1 The Cloud Environments (Cloud Deployment Models)

The cloud environments as identified in Figure 6.8 are shown in the list below, followed by an explanation of each in the sections that follow:

- Public Cloud Environment
- Private Cloud Environment
- Community Cloud Environment
- Hybrid Cloud Environment

6.4.1.1 Public Cloud Environment

A public cloud is one in which a service provider makes available resources such as applications, platforms and infrastructures to the general public over the internet. Public clouds are owned and operated at data centres belonging to the service providers and are shared by multiple customers (Subramanian, 2011a). This also means that public clouds offer unlimited storage space and increased bandwidth via the internet to any organisation across the globe. Such services on the public cloud may be offered free or on a pay-per-usage model. The degree of visibility and control of public clouds depends on the delivery mode. However, there is less visibility and control in public clouds compared to private clouds, because the underlying infrastructure is owned by the service providers.

6.4.1.2 Private Cloud Environment

A private cloud can be viewed as the implementation of cloud computing services on resources dedicated to an organisation (i.e. the organisation owns the hardware and software), whether they exist on-premises or off-premises. A private cloud offers an organisation the advantage of greater control over the complete stack, from the bare metal up to all the services accessible to users (Ubuntu, 2013).

6.4.1.3 Community Cloud Environment

A community cloud is one that is tailored to the shared needs of a business community. Community clouds are operated specifically for a targeted group. Usually, such groups (communities) have similar cloud requirements and their ultimate goal is to work together to achieve their business objectives. According to Techopedia (2013), community clouds are often designed for businesses and organisations working on joint projects, applications, or research, which requires a central cloud computing facility for building, managing and executing such projects. The infrastructure in a community cloud is shared by a number of organisations with common interest such as security, compliance, jurisdiction, etc., whether managed internally or by a third-party, or hosted internally or externally. The cost is, however, shared by all the participating organisations (Techopedia, 2013).

6.4.1.4 Hybrid Cloud Environment

A hybrid cloud is a combination of both public and private clouds (Subramanian, 2011b). This means that a vendor who owns a private cloud can form a partnership with

a public cloud provider, or a public cloud provider can form a partnership with a vendor that provides private cloud platforms. According to Mell and Grance (2011) of the National Institute of Standards and Technology (NIST), a hybrid cloud is a constitution of two or more public, private, or community cloud infrastructures that remain unique entities but are joined together by either standardised or proprietary technology that enables data and application portability.

Using the hybrid cloud architecture, organisations and individuals are able to gain degrees of fault tolerance combined with local and immediate usability, without dependency on internet connectivity. This is due to some of the resources in a hybrid cloud being managed in-house, while others are provided externally. In the next section the essential cloud components that also form part of the proposed ontology in this study are explained.

6.4.2 The Essential Cloud Components (Cloud Service Models)

Whichever the cloud environment deployed, cloud service providers will always offer their clients (individuals and organisations) the following three categories of cloud service models:

- Infrastructure-as-a-Service (IaaS)
- Platform-as-a-Service (PaaS)
- Software-as-a-Service (SaaS)

These service models are further discussed in the sub-sections to follow.

6.4.2.1 Infrastructure-as-a-Service (IaaS)

IaaS is a cloud computing service model that offers physical and virtual systems (cloud computing infrastructure), including an operating system, hypervisor, raw storage, and networks (Oracle Corporation, 2012). Servers represent the main computing resource in IaaS and are often virtual instances within a physical server. The service providers usually own the computing infrastructure and are responsible for housing, running and maintaining it. On the other hand, organisations pay on a per-use basis. IaaS helps organisations realise cost savings and efficiencies while modernising and expanding their information technology capabilities, without having to spend capital resources on infrastructure (GAS, 2013).

Infer from Figure 6.9 that the cloud computing infrastructure is further divided into communication services, network services and desktop services that form the fourth layer of the ontology shown earlier in Figure 6.8.

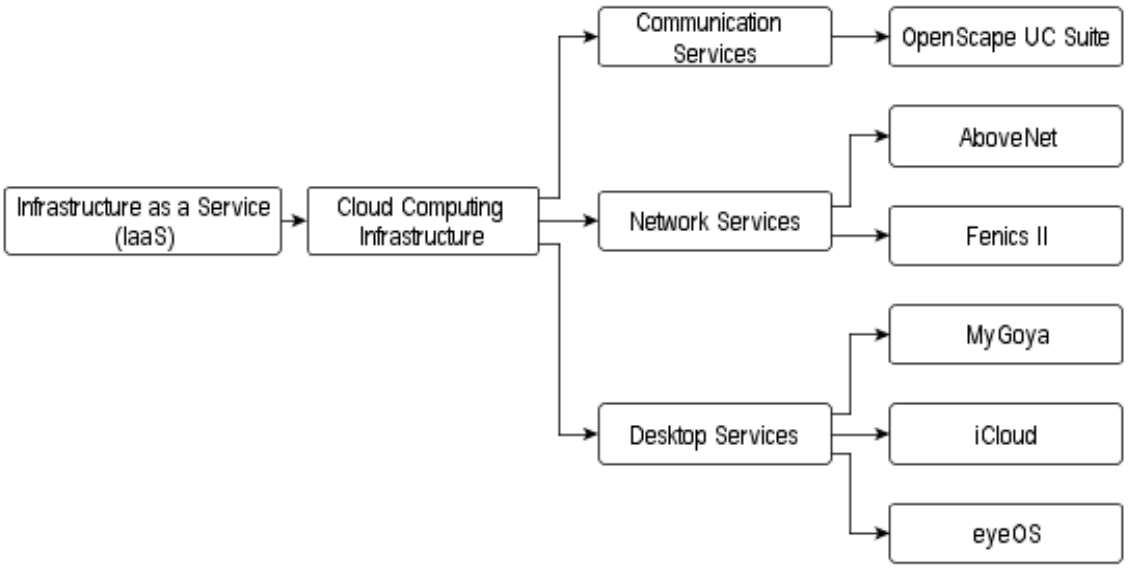


Figure 6.9 Infrastructure-as-a-Service

The communication services show OpenScape UC Suite as one of the service providers, while the network services have AboveNet™ and Fenics II as service providers. Finally, desktop services show MyGoya, iCloud and eyeOS as service providers. The service providers as shown in Figure 6.8 constitute the fifth layer of the ontology. However, the contents of the fourth and fifth layers (services and service providers respectively) in Figure 6.8 were introduced to provide selected examples for the purpose of this study.

6.4.2.2 Platform-as-a-Service (PaaS)

PaaS as explained in an expert group report by the European Commission (2010) provides computational resources (cloud computing platforms) via a platform upon which applications and services can be developed and hosted. PaaS typically uses dedicated application programming interfaces (APIs) to control the behaviour of a server hosting engine that executes and replicates the execution according to user requests. Cloud computing platforms may include the operating system, the programming language execution environment, the database, and the web server. PaaS also allows clients to use the virtualised servers and associated services for running applications or

developing and testing new applications. The cloud computing platforms as shown in Figure 6.10 are divided into business user platforms and development platforms.

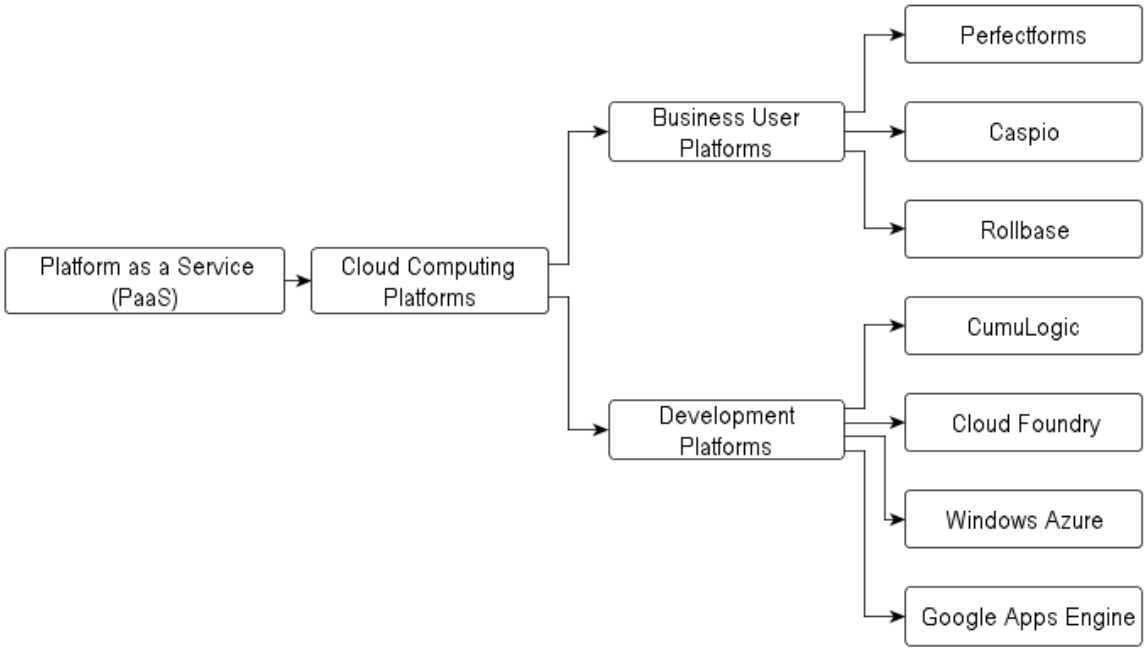


Figure 6.10 Platform-as-a-Service

Business user platforms have PerfectForms, Caspio™ and Rollbase as service providers. The development platforms show CumuLogic, Cloud Foundry™, Windows Azure™, and Google™ Apps Engine as selected service providers. However, as mentioned earlier, these are also mere examples for the purpose of this study and do not form an exhaustive list.

6.4.2.3 Software-as-a-Service (SaaS)

SaaS, also referred to as Service or Application Clouds (European Commission, 2010) offers implementations of specific business functions and business processes that are given with specific cloud capabilities, in other words they provide cloud computing applications or services using a cloud infrastructure or platform, rather than providing cloud features themselves. Moreover, SaaS also provides internet-based access to different software, thus presenting new opportunities for software vendors to explore.

The cloud computing applications as shown in Figure 6.11 are further divided into Enterprise Resource Planning (ERP) applications, e-commerce applications, Customer Relationship Management (CRM) applications and e-mail as selected examples.

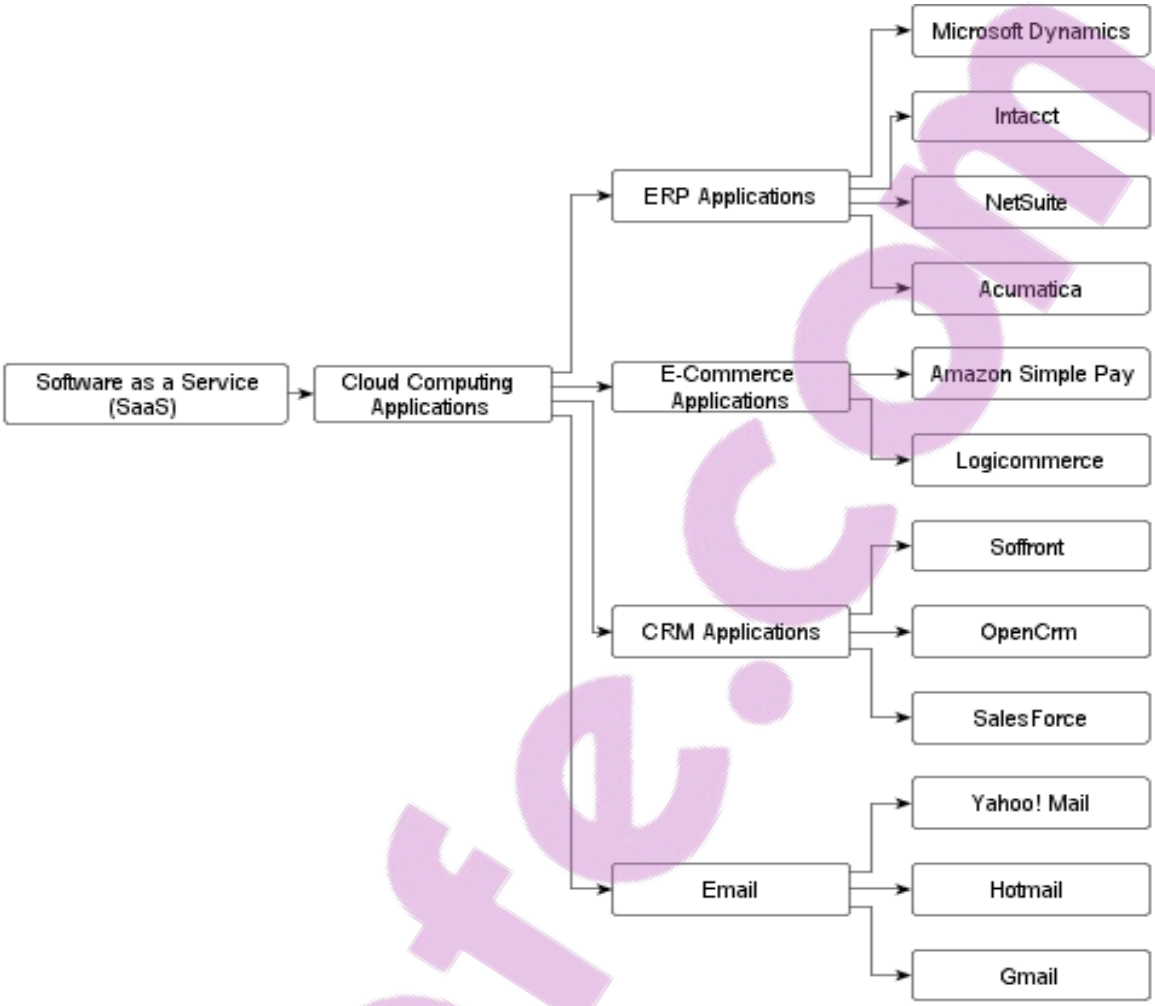


Figure 6.11 Software-as-a-Service

The ERP applications have Microsoft Dynamics™, Intacct®, NetSuite and Acumatica as service providers. E-commerce applications show Amazon Simple Pay and Logicommerce™ as examples of service providers. The CRM applications have Soffront®, OpenCrm and SalesForce® as service providers, while e-mail has Yahoo!®, Hotmail® and Gmail™ as examples of the service providers. The service providers were also selected as common examples for the purpose of this ontology and therefore do not purport to be an exhaustive list. The benefits of developing ontologies such as the ones explained in Sections 6.3 and 6.4 for the digital forensic domain are briefly explained in the next section.

6.5 BENEFITS FOR DEVELOPING ONTOLOGIES FOR DIGITAL FORENSICS

Information and knowledge sharing among people of a particular domain is one of the many benefits of developing ontologies. In the case of digital forensics, ontologies can be used as a way to develop a set of standards and methods by means of which to

approach the digital forensic domain. As an example, the ontologies presented in Sections 6.3 and 6.4 of this chapter can be used in the digital forensic domain to address issues such as professional specialisation and certification, as well as development of digital forensics tools, curricula and education materials.

For the case of professional specialisation, the digital forensic disciplines and sub-disciplines presented in the ontology in Figure 6.1 can be used to give direction to individuals interested in specific areas of specialisation. Such areas can produce specialists in computer forensics, software forensics, database forensics, multimedia forensics, device forensics and network forensics.

Institutions of higher learning can also benefit from the ontologies developed for digital forensics, especially when developing curricula and education materials for different undergraduate and postgraduate studies. Different modules can be developed with the help of ontologies to assist students in comprehending digital forensic concepts with less effort. Prerequisites for modules can, in addition, be designed effectively with the help of digital forensic ontologies so as to avoid conflicts among and redundancy of concepts.

Developers of digital forensic tools can use the developed ontologies both to fine-tune existing digital forensic tools and when considering the development of new digital forensic tools and techniques for specific areas of interest in the domain.

Finally, ontologies in digital forensics can be used to create a unified formal representation of the domain knowledge and information among computer professionals, law enforcement agencies and other digital forensic stakeholders. Besides, uniformity in the digital forensic domain can lead to high levels of collaboration between computer professionals and law enforcement agencies. Uniformity in the use of data or information and terminologies in the digital forensic domain can further help to expedite cooperation and collaboration.

6.6 CHAPTER CONCLUSION

In this chapter the researcher examined and explained the concepts of developing ontologies for digital forensics. A general ontology for the different digital forensic disciplines and sub-disciplines was developed, while an ontology for a cloud forensic environment was also proposed and explained. These ontologies are part of the contributions to the digital forensic domain presented in this research thesis.

The benefits of developing ontologies for digital forensics were also discussed in this chapter. This was done in an effort to find a way to help resolve the disparities that exist in digital forensics, as well as to create a base for the development of a unified formal representation of digital forensic domain knowledge and information. To avoid misunderstandings caused by semantic disparities, digital forensics therefore requires such ontologies to be developed to resolve the semantic disparities that may occur in the domain.

Chapter 7 which is next explains the development of a digital forensic semantic reconciliation model, also as a way towards resolving the semantic disparities in the digital forensic domain.

CHAPTER 7 : A DIGITAL FORENSIC SEMANTIC RECONCILIATION (DFSR) MODEL

7.1 INTRODUCTION

Decades of digital forensic research have been conducted. Nonetheless, it remains a challenge for computer professionals, law enforcement agencies and other digital forensic practitioners to exchange and harmoniously use information from heterogeneous sources when a digital forensic investigation process has to be carried out. This is aggravated by various disparities that are common within the digital forensic domain, of which one notable challenge is the semantic disparities that occur. Therefore methods and specifications need to be developed to assist in resolving semantic disparities in the digital forensic domain.

This chapter, therefore, aims at proposing a systematic Digital Forensic Semantic Reconciliation (**DFSR**) model in an attempt to provide direction towards resolving the semantic disparities that occur in the digital forensic domain. Such a model can for example be used to develop new techniques for detecting and managing semantic disparities. The DFSR model can also be incorporated as part of existing digital forensic tools to help create uniformity in interpreting, describing and representing digital forensic terminologies as well as enhancing a common understanding of domain information.

Section 7.2 of this chapter presents related work while Section 7.3 explains why it is necessary to develop a semantic reconciliation model for the digital forensic domain, followed by a discussion of the proposed DFSR model in Section 7.4. A discussion of the DFSR model is presented in Section 7.5 and finally a chapter conclusion is presented in Section 7.6.

7.2 RELATED WORK

In this section of the study the researcher presents related work concepts on semantic reconciliation models. In a paper by Avigdor et al. (2003) the researchers presented an elaborate model for semantic reconciliation and analyse in an organized way the elements of the process results, especially the basic uncertainty of the matching process and how it reflects on the resulting mappings. A significant component of their research

is the singling out and analysis of different elements that have an effect on the effectiveness of algorithms for automated semantic reconciliation, leading to the development of better algorithms by minimizing the uncertainty of current algorithms.

In another paper by Carlos and Eduardo (2014), the authors propose a way to increase the discovery of Web Services based on the semantic reconciliation of providers and requesters before the discovery process itself. Their system reuses and integrates ontological information extracted from several online pools of ontologies to make it possible to:

1. Add semantics easily to existing non-semantic services; and
2. Perform a semantic keyword based search autonomously of the ontologies used by the provider thus bridging the semantic gap between requesters and providers (Carlos and Eduardo, 2010).

In another effort by Chungoora, and Young (2011) they investigate improved concepts to achieve semantic reconciliation in the context of the Semantic Manufacturing Interoperability Framework (SMIF). Their approach uses a Common Logic-based underpinning for enabling the evaluation and verification of cross-model correspondences. They then teste their approach by applying the relevant logic-based mechanisms in order to show the reconciliation of two individually developed machining hole feature knowledge models. Through this, they then demonstrate that the approach enables semantic reconciliation of important structures within ontology-based models of design and manufacture.

More work done by Kuhanandha and Michael (1999) explains how information spaces stored as ontologies are appropriate for semantic reconciliation. In addition, they make mention of a number of advantages of using ontologies in a distributed and dynamic environment like the Internet. Their paper also analyses the setting up and review of an ontology-based distributed information system developed using the Java language all to help in semantic reconciliation. Based on the concepts discussed on the related works, the next section discusses the need to develop a digital forensic semantic reconciliation model in digital forensics.

7.3 THE NEED TO DEVELOP A DIGITAL FORENSIC SEMANTIC RECONCILIATION MODEL

Ever since its genesis, the digital forensic field has continued to gain importance in society. This is because there is an absolute need for computer professionals, law

enforcement agencies and other digital forensic practitioners to cooperate when conducting a digital forensic investigation. Semantic disparities may cause investigators to face crippling problems when they do not understand a particular domain terminology or keyword used by their counterparts to interpret, describe or represent domain data or information during a digital forensic investigation. For this reason, methods and specifications need to be developed that have some kind of intelligence for detecting and managing semantic disparities that may become apparent in the digital forensic domain.

As mentioned earlier, the digital forensic community has always struggled with semantic disparity, as noted by Dolan-Gavitt et al. (2011). Unfortunately, digital forensics lacks comprehensive or standardised methods and tools that have been specifically designed to resolve semantic disparities. Most of the existing digital forensic investigation tools consist of dissimilar elements or parts and consequently inhibit the ability of stakeholders to work together harmoniously. Despite the advances in digital forensics, computer professionals, legal professionals and researchers are yet to resolve the challenges associated with semantic disparities in the domain.

In the case of an investigation process that leads to a trial, for example, any statement made during the presentation of potential digital evidence should be such that it introduces the court to the necessary terminology and types of digital evidence that may be presented. This is necessary since the presentation of any potential digital evidence may involve different digital forensic domain terminologies, issues and concepts that are complex or unfamiliar to the court (Karie and Venter, 2013a).

Therefore, the DFSR model introduced in this chapter is an attempt towards developing a new way to resolve semantic disparities in the digital forensic domain. It is intended to help create a common way to clearly interpret, describe and represent the different types of data or information in digital forensics – especially data presented in court or civil proceedings. Moreover, it can help the court to understand the terminologies used to present the potential digital evidence data and allow for the successful outcome of the trial. The next section explains the proposed DFSR model in detail.

7.4 THE PROPOSED DIGITAL FORENSIC SEMANTIC RECONCILIATION (DFSR) MODEL

The DFSR model, as demonstrated in this study, was designed using the Incremental model. In the incremental model the requirements as presented by Marciniak, (2001) as

well as Munassar and Govardhan (2010) are usually divided into various segments. Multiple development cycles take place, making the life cycle a “multi-waterfall” cycle, i.e. cycles are divided up into smaller, more easily managed modules. Each module passes through the requirements, design, implementation and testing phases. A working version of software is produced during the first module, so one has a working version early on during the software life cycle. Each subsequent release of the module adds function to the previous release. The process continues till the complete system is achieved. This way made it possible in this study to develop the most essential modules first to test the feasibility of the DFSR model.

The DFSR model presented in this section, hence, is an attempt towards developing methods and specifications for resolving semantic disparities in digital forensics. Such methods and specifications should execute faster, perform better and more accurately and be free from any ambiguities. Besides, effective cooperation among different stakeholders in the digital forensic domain, as mentioned earlier, presupposes that information originating from varying sources should be harmonised to create uniformity and common understanding in the domain.

The harmonisation process, though, can be very costly and close to impossible if it is to be done manually. This is especially true when the people involved have different background and perceptions regarding the interpretation, description and representation of certain digital forensic terminologies (Karie and Venter, 2013b). It is exactly this situation in digital forensics that has motivated the development of the DFSR model as presented in this chapter to help resolve sematic disparities. Note also that the model proposed here is meant to complement many of the existing tools; however, it can also be used to develop completely new tools. Figure 7.1 shows the high-level logical conceptualisation of the DFSR model. This model however borrows some features from existing structures such as the annotation process used with WordNet structure (Miller, 1995). Both the development of the semantic annotations and the creation of the semantic repository employs the WordNet like logics hence the possibility of automatically transferring a list of Words and their characteristics from WordNet into the sematic repository of the proposed model shown in Figure 7.1.

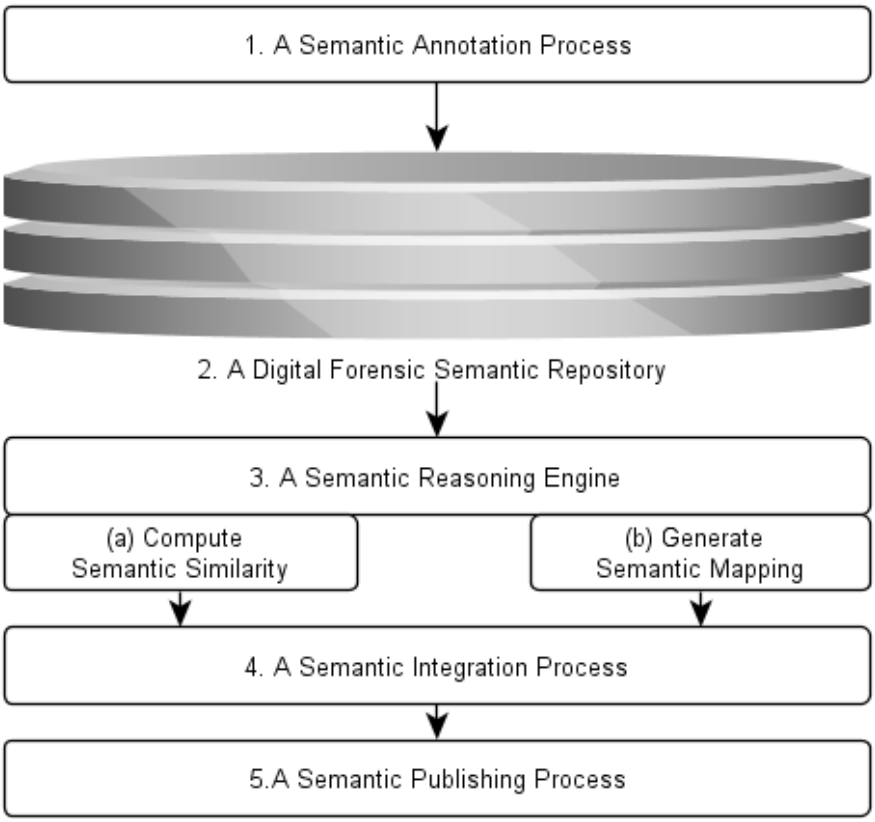


Figure 7.1 High-Level Logical Conceptualisation of the DFSR Model

The DFSR model consists of five phases arranged from top to bottom, where the first phase involves the creation of semantic annotations based on the different digital forensic domain terminologies. The second phase uses the developed semantic annotations from Phase 1 to build a live and active electronic semantic repository. The third phase has a sematic reasoning engine which involves the use of accepted or standardised methods that are able to compute semantic similarities of different domain terminologies as well as generate semantic mapping, based on specific extracted terminology parameters from the semantic repository. Phase 4 of the proposed model handles semantic integration and finally the fifth and last phase deals with semantic publishing. Together the five phases make up the proposed DFSR model. In the sub-sections to follow, Phases 1 to 5 as shown in Figure 7.1 are explained in more detail.

7.4.1 A Semantic Annotation Process

The semantic annotation process is an act of expressing knowledge about a particular resource, terminology or phrase. This process involves attaching names, attributes, comments, descriptions, etc., to specific domain terminologies (Ding, 2006). Semantic

annotation is therefore responsible for providing all the information (including additional metadata) about an existing domain terminology or data.

A standard semantic annotation exercise has three major building blocks: the ontology, a data instance recognition process and lastly an annotation generation process. The ontology gives an account of the domain of interest. In this case the domain of interest is digital forensics. The concept of ontologies was already discussed earlier in chapter 6 of this thesis. The data instance recognition process comes next and is meant to discover all the instances of interest in a target document or terminology based on the defined ontology. Finally follows an annotation generation process that is meant to create a semantic meaning disclosure file for each annotated document or terminology (Ding, 2006). Through the semantic meaning disclosure file, any ontology-aware machine agent can understand the target document or terminology (Ding, 2006). The three components are shown in Phase 1 of Figure 7.2 as 1.(a), 1.(b), 1.(c).

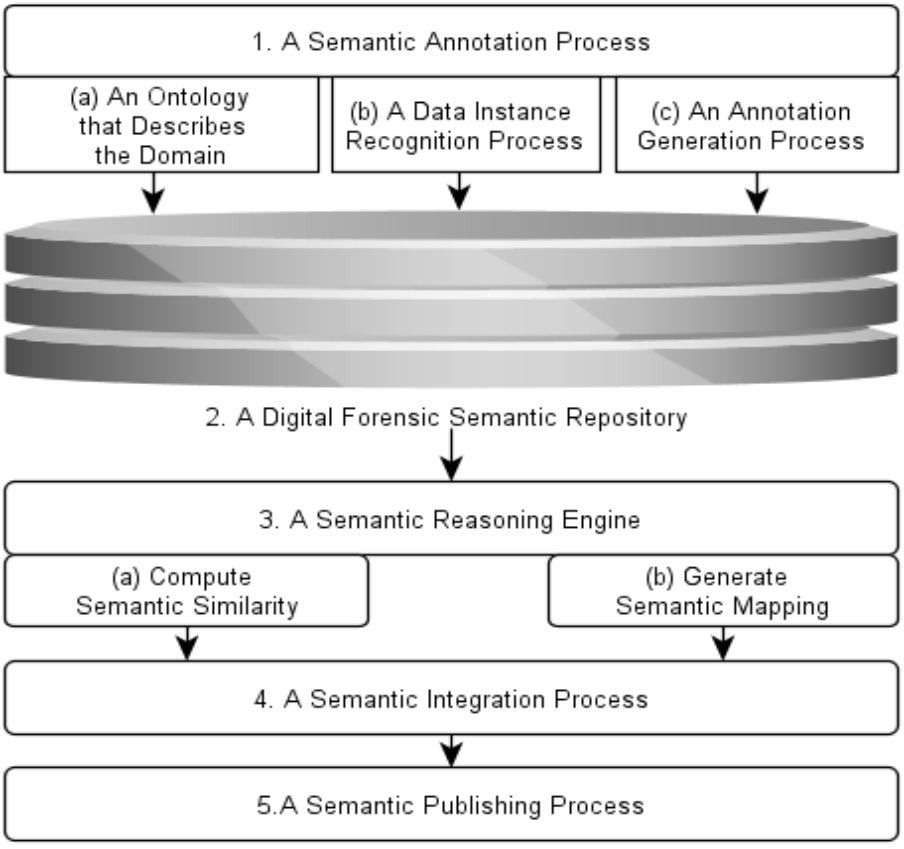


Figure 7.2 Enhanced DFSR Model

Note that Figure 7.2 is an enhanced version of the DFSR model with Phase 1 showing all three components of the semantic annotation process.

As shown in Phase 1 of Figure 7.2, semantic annotation makes it possible to assign links to existing semantic descriptions of any domain terminology in question. This makes it possible to relate one domain terminology to another. With the first phase of the DFSR model it also becomes possible to annotate different digital forensic terminologies that are later stored in the semantic repository. Figure 7.3 shows a high-level example of how terminologies in digital forensics can be related, based on their semantic annotations using the DFSR model.

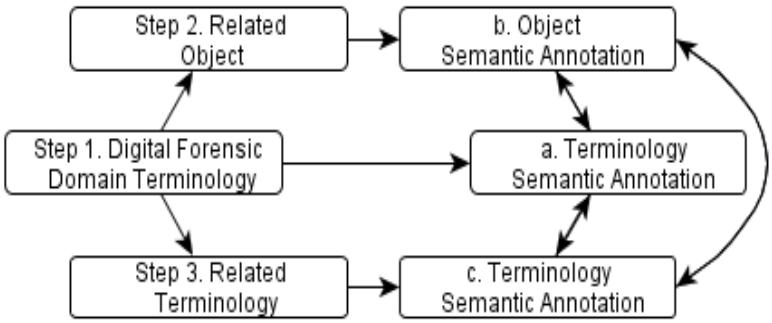


Figure 7.3 Semantic Annotations of DF Terminologies using the DFSR Model

Infer from Figure 7.3 that the process begins with a root terminology (digital forensic domain terminology) shown as Step 1. The root terminology is, however, related to an object in Step 2 as well as to another terminology in Step 3. The root terminology in Step 1, the related object in Step 2, and the related terminology in Step 3 all have semantic annotations associated with each one of them, shown as 'a', 'b' and 'c' in Figure 7.3. Note also from Figure 7.3 that the semantic annotations of the related object labelled 'b' and the related terminology labelled 'c' are connected by means of arrows, implying that they ('b' and 'c') are both related to the semantic annotation of the root terminology labelled 'a'. The whole process results in a semantic repository with related terminologies and objects that can be used as desired. The next section explains the process of creating the semantic repository.

7.4.2 A Digital Forensic Semantic Repository

At the heart of the DFSR model lies a semantic repository which is a large and structured set of texts stored in a knowledge base format. The semantic repository is a very useful information source for resolving the semantic disparities as well as computing the semantic similarities of different terminologies by using the proposed DFSR model. In

the case of resolving semantic disparities in digital forensics, such a repository needs to be developed to enable the extraction or retrieval of terminology parameters necessary for computing the semantic similarities, as well as for semantic mapping as shown in Phase 3 of Figure 7.2.

However, in place of the semantic repository, a text corpus or digital library can also be developed and used. A digital library – also known as an electronic library – is a collection of information stored in digital or electronic formats that can easily be accessed using any computer system. The content of a digital library can be stored locally in a particular computer system or in a local database. However, it can also be stored remotely in a server and made accessible via computer networks.

Unfortunately, in the case of this study, the researcher does not focus on developing a digital library, but a semantic repository. This is because, even at the time of this research study, digital forensics lacked a standardised semantic repository that can be used to resolve the semantic disparities that exist in the domain. The latter includes the testing and/or implementation of any proposed model such as the DFSR model that features in this research study. A standardised digital forensic semantic repository needs to be established for use in implementing newly developed models such as the DFSR model, as well as new tools and techniques. The next sub-section explains Phase 3 of the proposed DFSR model, which deals with semantic similarities and semantic mapping as shown in Figure 7.1.

7.4.3 A Asemantic Reasoning Engine for Computing Semantic Similarity and Generate Semantic Mapping

Semantic similarity measures numerically compute the degree of similarity and relatedness among different terminologies and, in the case of this study, the digital forensic terminologies.

An accurate measurement of semantic similarity between terminologies is a matter of concern in many different domains (Karie and Venter, 2012). Even with the importance of computing semantic similarity in different domains, the accurate measurement of semantic similarity between any two terminologies has remained a challenging task. The main difficulty lies in developing a computational method that has the ability to generate satisfactory semantic similarity results that closely resemble the way in which human beings perceive these terminologies, especially when used in their domain of expertise.

In addition, different methods for computing semantic similarity between terminologies are available as shown in Table 7.1. Each method uses different terminology parameters for computing semantic similarity between given terminologies. Some of the existing methods, for example, compute semantic similarity between terminologies based on taxonomies, while others are Web-based. Taxonomy-based methods, for example, use information theory and hierarchical taxonomy such as WordNet (Miller, 1995) to measure semantic similarity. Web-based methods, on the other hand, prefer the Web as a live and active text corpus from which to elicit a hierarchical taxonomy (Zheng et al., 2011).

Care should be exercised, though, to use a method that is widely accepted by the scientific community for computing semantic similarity. Whichever the method used, it should be able to generate satisfactory results closer to how the terminologies in question are currently perceived or interpreted by the domain experts.

For this reason, a method for computing semantic similarity between different digital forensic terminologies should be used in Phase 3 of the DFSR model as shown in Figure 7.2. Table 7.1 below shows some of the available methods for computing semantic similarity between terminologies, as proposed by different researchers.

Table 7.1Various Methods for Computing Semantic Similarity between Terms

Method Description	Author(s) and year
1. Measuring semantic similarity between words using Web search engines	Danushka et al. (2007)
2. Measuring semantic similarity between words using Web documents	Sheetal and Sushama (2010)
3. Measuring semantic similarity between words using page counts and snippets	Manasa et al. (2012)
4. A Web Search Engine-based approach to measure semantic similarity between words	Danushka et al. (2011)
5. A combined method to measure the semantic similarity between words	S. Vijay (2012)
6. Measuring semantic similarity between words using Web pages	Sujatha et al. (2012)

7. Measuring semantic similarity between digital forensics terminologies using Web Search Engines	Karie and Venter (2012)
8. Measuring semantic similarity between words using page-count and pattern clustering methods	Prathvi and Ravishankar (2013)

The different methods shown in Table 7.1 are also explained briefly below.

7.4.3.1 Measuring Semantic Similarity between Words using Web Search Engines

In this method Bollegala et al. (2007) proposed a method to measure similarity between words or entities using information that is available on the Web. This method exploits page counts and text snippets returned by a Web search engine. Their paper defines various similarity scores for two given words P and Q, using the page counts for the queries P, Q and P AND Q. They also propose an approach to compute semantic similarity using automatically extracted lexico-syntactic patterns from text snippets. These different similarity scores are integrated using support vector machines, to leverage a robust semantic similarity measure.

7.4.3.2 Measuring Semantic Similarity between Words using Web Documents

Sheetal and Sushama (2010) also proposed a method to compute semantic similarity between words or terminologies that uses web- based metrics. Their method makes use of snippets returned by Wikipedia or any encyclopaedia such as Britannica Encyclopaedia. The snippets are pre-processed for stop word removal and stemming. For suffix removal they use an algorithm by M. F. Porter (1980). Luhn's Idea is also used in Sheetal and Sushama (2010) for taking out important words from the pre-processed snippets (Luhn, 1958).

7.4.3.3 Measuring Semantic Similarity between Words using Page Counts and Snippets

Manasa et al. (2012) in their paper proposed a method to find semantic similarity between words based on text snippets and page counts. These two measures are taken from the results of a search engine like Google. Besides, lexical patterns are taken out from text snippets and page counts are used to describe word co-occurrence measures. The results of these two are combined. In addition, they proposed algorithms such as pattern clustering and pattern extraction in order to find various relationships between any given two words. They also employ Support Vector Machines, a data mining

technique to optimize the results. The empirical results reveal that their proposed techniques are finding best results that can be compared with human ratings and accuracy in web mining activities.

7.4.3.4 A Web Search Engine-based Approach to Measure Semantic Similarity between Words

Danushka et al. (2011) in their paper put forward an experiential method to determine semantic similarity with the help of page counts and text snippets obtained from search engines for two different terms. Particular, they define a variety of term co-occurrence measures with the help of page counts and consolidate those with lexical patterns obtained from text snippets. To single out the very many semantic relationships that exist between two terms, they suggest two algorithms namely: pattern extraction and pattern clustering algorithms respectively. The best integration of page counts based co-occurrence measures and lexical pattern clusters is shown with the help of support vector machines.

7.4.3.5 A Combined Method to Measure the Semantic Similarity between Words

Vijay (2012) in his paper put forward an approach that make use of the information accessible from the Web to compute semantic similarity between a pair of terms or entities as well as combine page counts for each term in the pair and lexico-syntactic patterns that come about among top ranking snippets for the AND query with the help of support vector machines.

7.4.3.6 Measuring Semantic Similarity between Words using Web Pages

Sujatha et al. (2012) presented an approach that utilizes web based metrics to calculate semantic similarity between terms. The researcher however found this method by Sujatha et al. (2012) exactly to be similar to the one proposed by Sheetal and Sushama (2010).

7.4.3.7 Measuring Semantic Similarity between Digital Forensics Terminologies using Web Search Engines

In the case of implementing and testing the feasibility of the DFSR model, a Web-based method was used to compute the semantic similarity between different digital forensic terminologies (Listed as number 7 in Table 7.1). This method is however completely different from all the above cited methods in that, the proposed method is based on the Euclidean distance, a mathematical concept used to calculate the distance

between two points. None of the cited methods employed this theory in computing the semantic similarity.

The proposed method by Karie and Venter (2012) shows how computing the absolute value of the difference of the logarithms of the hit count percentages of any given terms x and y relates to the computed Euclidean distance of x and y. Percentages are computed from the total number of hit counts reported by any Web search engine for the terms x, y and the logical x AND y together. The method then uses these concepts to deduce a formula to automatically calculate a semantic similarity measure coined as the Digital Forensic Absolute Semantic Similarity Value of the terms x and y, denoted as DFASSV(x, y). Experiments conducted using the proposed DFASSV(x, y) method focuses on the digital forensic domain and are explained in Chapter 8 where a prototype implementation is also presented. However, a comparison of the DFASSV approach with previously proposed Web-based semantic similarity measures shows that this approach is well suited for digital forensics domain terminologies.

7.4.3.8 Measuring Semantic Similarity between Words using Page-count and Pattern Clustering Methods

Finally as shown in Table 7.1, Prathvi and Ravishankar (2013) presented an approach to compute semantic similarity between terms with the help of information found on the web as well as methods that make use of page counts and snippets to measure semantic similarity between two terms. They explain different term co-occurrence measures with the help of page counts and then combine those with lexical patterns obtained from text snippets.

Table 7.1 is used in this study to show that the use of the web in computing semantic similarity has been employed several times by different researchers however; none of the methods explored the theory of Euclidean distance in the manner proposed in this research thesis. Hence the proposed method listed as number 7 in Table 7.1 adopts the use of the web and Euclidean distance concept to compute semantic similarities. Although there exist many other proposed methods for finding word similarity other than the ones stated in Table 7.1, all such methods have their own shortcomings. Therefore, new methods need to be developed in digital forensics that have the ability to generate satisfactory semantic similarity results.

As part of Phase 3 Semantic mapping is also shown in the DFSR model and it helps to display the meaning-based connections between the domain terminologies or

phrases and a set of related terminologies or concepts. The primary objective of semantic mapping in the DFSR model is to help to expand vocabulary and extend knowledge by displaying in specific categories those terminologies that are similar or related to one another, based on their computed semantic similarity (Gerald, 2009). Semantic mapping is also used to help explain how the terminology meanings are categorised (Gerald, 2009).

Semantic mapping in the proposed DFSR model furthermore helps to identify key attributes that distinguish one terminology from another. The generated semantic maps usually provide the additional benefit of helping individuals to visualise how the terminology meanings are categorised. With the help of semantic maps, it also becomes easy for individuals to identify, understand and recall the meaning of terminologies they read in the text. A semantic map allows people to conceptually explore their knowledge of a new terminology by mapping it together with other similar or related terminologies or phrases that are similar in meaning to the new terminology (Erick, 2012). Semantic mapping further enables the adaptation of concept definition and the visual display of terms or phrases and a set of related terms or concepts. Most importantly, semantic maps will help people to recall the meaning of words that they read in texts.

The next sub-section explains semantic integration, which constitutes Phase 4 of the DFSR model.

7.4.4 A Semantic Integration Process

Semantic integration is the process of interrelating information from diverse sources (Li and Clifton, 1994). This is one of the most challenging processes in the DFSR model, since the information sources to be used during semantic integration may lack consistent information architecture (Liaison, 2015). However, leveraging on the semantic integration process can lead to enhanced quality of domain data through centrally governed, locally distributed, reusable active processes, including reduced time and cost to merge digital forensic tools that work on the same data. Semantic integration can also allow for integration of different digital forensic tools without giving rise to the costs related to manually harmonizing and validating uneven domain data interchange between the tools (Liaison, 2015).

The other reason for introducing semantic integration in Phase 4 of the DFSR model is to help with the process of interrelating information from diverse sources or

different digital forensic tools. Semantic integration is also useful to automate communication between different systems using metadata publishing (Liaison, 2015). In the case of digital forensics, metadata publishing potentially offers the ability to automatically link different domain ontologies. These improvements are the motivation for introducing the semantic integration process in Phase 4 of the DFSR model. The last phase of the DFSR model deals with semantic publishing, which is explained in the section to follow.

7.4.5 A Semantic Publishing Process

Semantic publishing constitutes the fifth and last phase of the DFSR model. Semantic publishing helps people and systems to understand the structure and even the meaning of the published domain terminologies and information (Shotton, 2009). It also helps to make information search and data integration more efficient (Shotton et al., 2009). Any published information is usually accompanied by metadata that describe such information, hence providing a semantic context (Shadbolt et al., 2006).

Phase 5 is designed to have the capability to publish the domain information that is accompanied by semantic mark-ups. Semantic mark-ups are usually written to define the context of the content enclosed in the mark-up. Semantic mark-ups can also be used to reinforce the semantics or meaning of domain information or terminologies, rather than to merely define its presentation or look (Shadbolt et al., 2006). The other reason for introducing this phase in the proposed DFSR model is to help define the context and the structure of the different domain terminologies by using the appropriate semantic elements. The next section presents a discussion of the proposed DFSR model.

7.5 A Discussion of the Proposed DFSR Model

The DFSR model proposed in this chapter is a new contribution in the digital forensics domain. The scope of the model is defined by the phases as shown in Figure 7.1, and the main phases as depicted in the DFRS model include the following:

- A Semantic Annotation Process
- A Digital Forensic Semantic Repository
- A Semantic Reasoning Engine
- A Semantic Integration Process
- A Semantic Publishing Process

The specific details of the individual phases as identified in the DFSR model have been explained so far in this chapter. The proposed DFSR model can be used in the digital forensics domain as one way to resolve semantic disparities and create uniformity and a common understanding of domain data and terminologies. It can also be useful in identifying relevant terminologies to be used during the interpretation, description and representation of digital forensic evidence by helping to define the context of the terminologies used, for example, in court or legal proceedings. The DFSR model is also useful in establishing a unified and formal representation of the domain terminology semantics that are required during a digital forensic investigation process.

Developers of digital forensic tools can furthermore use the proposed DFSR model to incorporate new features in existing tools with the ability to detect and resolve semantic disparities in the domain, for instance, during forensic memory analysis. This implies that developers may also find the proposed model in this chapter useful, especially when considering the development of new digital forensic techniques and tools for resolving the semantic disparities that exist in the domain.

Finally, the proposed DFSR model presented in this chapter was designed in a way to accommodate new phases that may emerge as a result of future requirements or domain evolution. To the best of the researcher's knowledge, at the time of writing this thesis there existed no other work of this kind in the digital forensic domain. Therefore, this is a new contribution in digital forensics towards resolving semantic disparities.

7.6 CHAPTER CONCLUSION

In this chapter the need to develop a digital forensic semantic reconciliation model was discussed. A model coined as the Digital Forensic Semantic Reconciliation (DFSR) model for resolving semantic disparities in digital forensics was then proposed and explained. The model consists of five phases arranged from top to bottom, where the first phase involves creating semantic annotations and the second deals with the creation of a semantic repository.

A reasoning engine with the ability to compute semantic similarity and generating semantic mapping of the terminologies in question is dealt with in the third phase. Phase 4 handles semantic integration and semantic publishing is explained in Phase 5. The proposed DFSR model discussed constitutes one way towards resolving semantic disparities in digital forensics. The next chapter will test and determine the feasibility and implementation of the proposed DFSR model.

CHAPTER 8 : TESTING THE FEASIBILITY AND IMPLEMENTATION OF THE PROPOSED DFSR MODEL

8.1 INTRODUCTION

In order to resolve semantic disparities in digital forensics, several approaches were discussed in this research study. This chapter demonstrates the feasibility of the proposed Digital Forensic Semantic Reconciliation DFSR model in a prototype implementation called the DFSR prototype. The DFSR prototype serves to provide the fundamental specifications to implement the DFSR model proposed in Chapter 7 of this research thesis. Although the DFSR prototype is not a complete implementation of the DFSR model and not fully automated, it nevertheless implements the most essential components of the DFSR model that is necessary to demonstrate that the model is feasible[3].

Section 8.2 presents the objectives of the DFSR prototype while the DFSR prototype implementation is discussed in Section 8.3. Section 8.4 concentrates on the experimental results based on the proposed DFASSV method. Note that the experimental results discussed in section 8.4 are based on the individual methods used to develop the DFSR prototype in this study. More results based on experiments carried out in this study are explained in Section 8.5 and a chapter conclusion is presented in Section 8.6. The next section explains the objectives of the DFSR prototype.

8.2 THE OBJECTIVES OF THE DFSR PROTOTYPE

The DFSR prototype developed in this study is a functional implementation of the DFSR model. The primary objective of developing the DFSR prototype is to demonstrate that, with an active or standardised digital forensic semantic repository, it is possible for individuals and digital forensic experts to achieve semantic reconciliation. In the case of semantic disparities existing in digital forensics, for example, the DFSR prototype presented in this chapter can be incorporated as part of the existing digital forensic tools to serve as a quick guide to semantic reconciliation.

[3] Note that the contents of this chapter were presented as a research conference paper at the Annual Conference on Information Security of South Africa in August 2012. The title of the paper is "Measuring Semantic Similarity between Digital Forensics Terminologies Using Web Search Engines".

Chapter 8 is also meant to demonstrate a computational method in the digital forensic domain that has the ability to generate satisfactory semantic reconciliation results close to how individuals or domain members perceive the domain terminologies in question. This is important, especially when interpreting, describing and representing digital forensic domain data or information to different stakeholders.

The feasibility and implementation of the DFSR prototype, which is intended to facilitate the semantic reconciliation process in digital forensics, is explained in the next section.

8.3 THE FEASIBILITY AND IMPLEMENTATION OF THE DFSR PROTOTYPE

As mentioned before, digital forensics lacks a standardised semantic repository. Thus, to test the feasibility of the proposed DFSR model, the World Wide Web is used as a live and active electronic semantic repository in this study, seeing that the Web is a vast entity where an astronomical amount of information is amassed. The Web is also viewed as the biggest semantic electronic database (knowledge base) globally. This 'database' is freely accessible to everyone and can be queried with the help of Web search engines that can give back aggregate hit count approximation for a vast range of search queries (Cilibrasi and Vitànyi, 2007).

New information is also added to the Web and annotated on a daily basis. For this reason, the semantic annotation process shown as Phase 1 of the proposed DFSR model in Figure 7.2 will be assumed. The Web is used in this study to represent the semantic repository shown as Phase 2 of the proposed DFSR model in Figure 7.2. To tap into this rich bank of information (the Web), Web search engines are the most frequently used tools to query for information related to a particular term. To the researcher's knowledge, there is so far no better or easier way to search for information on the World Wide Web than simply by using Web search engines like Google. However, the researcher does not dispute the existence of other techniques that can be used to search for and extract information from the Web. For the purpose of this study, however, the Google search engine was used as a tool that has the ability to extract the different terminology parameters necessary for the testing of the proposed DFSR model.

The scope of implementation of the DFSR prototype is shown in Figure 8.1. The implementation process has six steps, of which the first step is to identify the domain terminologies in question. This is followed by extracting the terminology parameters from an existing semantic repository (Web) or a digital library in Step 2. The extracted

terminology parameters are used to compute semantic similarities of the selected domain terminologies. The semantic similarity measures help to produce semantic maps in Step 3. Step 4 is responsible for generating semantic maps while step 5 then handles the semantic integration process. Finally, Step 6 takes care of semantic publishing.

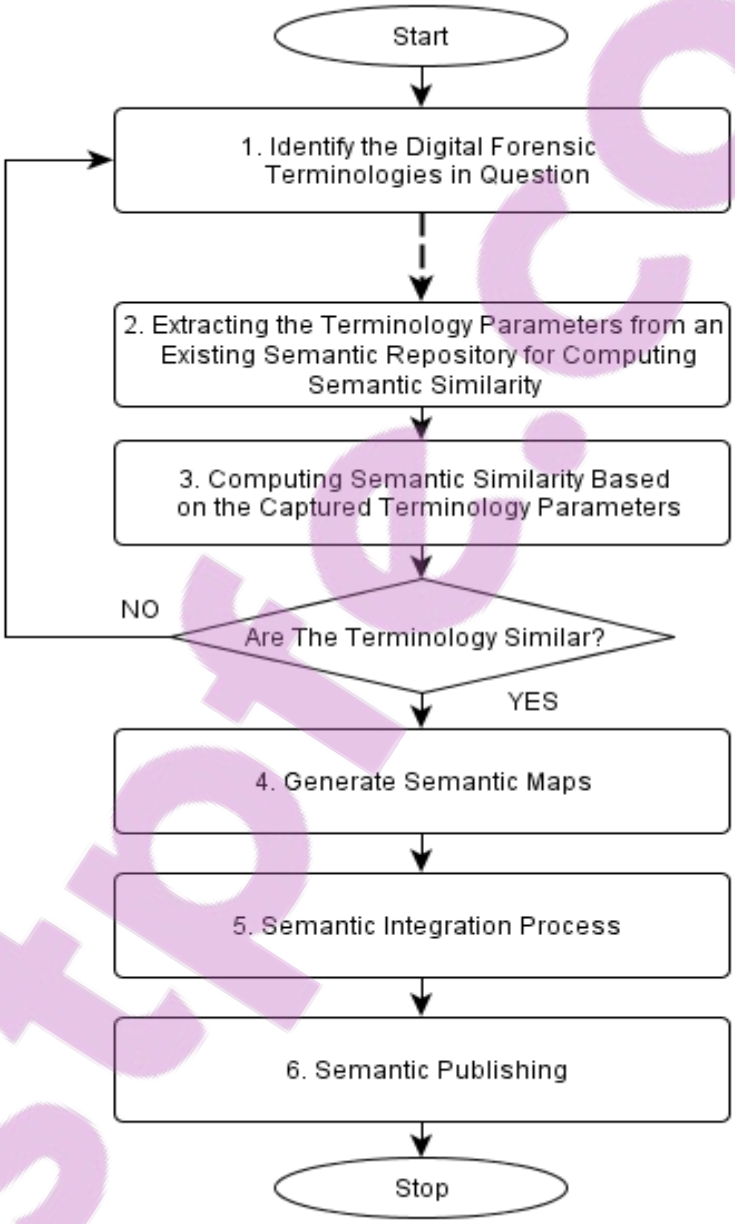


Figure 8.1 The Scope of the DFSR Prototype Implementation

From Figure 8.1 it is evident that the identification of the terminologies in question shown as Step 1 is not automated in this version of the DFSR prototype and thus requires some level of human intervention or manual activity. Steps 1 to 6, as shown in Figure 8.1, are described in detail in the next sub-sections. The different interfaces developed

and used to test the DFSR prototype are also presented and explained as a way to help show the feasibility and implementation of the proposed DFSR model in this study.

8.3.1 Identify the Digital Forensic Terminologies in Question

Identifying the terminologies in question constitutes Step 1 of the DFSR prototype which is not automated; however, future versions of the DFSR prototype will consider automating all the steps. For this reason, one needs to identify which terminologies in digital forensics are in conflict with each other and thus require semantic reconciliation. It also implies that one has to identify the different components and relationships between the domain terminologies in question. To do this, the semantic similarity measure is computed to help determine the relationship between the terminologies.

However, before computing the semantic similarity of the terminologies, the necessary terminology parameters need to be extracted from an existing semantic repository. Note from Figure 8.1 that because the Web is used as the semantic repository in this study, a dotted line is used to assume that Phases 1 and 2 of the proposed DFSR model shown earlier in Figure 7.2 have already been completed. The process of extracting the terminology parameters, however, is Step 2 of the DFSR prototype implementation and it is explained in the next sub-section.

8.3.2 Extracting the Terminology Parameters from an Existing Semantic Repository for Computing Semantic Similarity

Different methods exist for computing the semantic similarity between any given terminologies and they may as well use different terminology parameters. However, to test the feasibility and implementation of the DFSR prototype in this study, Web-based methods are used for computing semantic similarity. For this reason, Phases 1 and 2 of the DFRS model are assumed to have already executed during the testing and implementation of the DFSR prototype. This also implies that no new semantic repository is developed to test the DFSR model; the Web is used instead. New terminologies with semantic annotations are usually added to the Web on a daily basis, which results in a global live and active semantic repository.

Web-based methods, as mentioned earlier, use the Web as a live and active text corpus for measuring semantic similarity between terminologies. Hit counts reported by Web search engines are therefore useful information sources for this study and, as such,

are used as the essential terminology parameters for computing semantic similarity in this study.

The hit count of any query given is usually treated as an estimated number of Web pages containing the queried term as reported by a Web search engine. The hit count, however, may not definitely be equivalent to the term frequency. This is because the queried term may appear several times on a single page. Consequently, to accurately compute semantic similarity in this study, the hit count of any given search terms *x* and *y* is computed where the search terms *x* and *y* both appear on the same Web page. This is indicated as a logical *x* AND *y* search query. The search results of this query can be viewed as the global estimated value of the co-occurrence of the terms *x* and *y* together on the Web (Thiyagarajan et al., 2011). Logical *x* AND *y* is also used in this study to capture the context where both *x* and *y* are used together on the same Web page.

However, the researcher does not consider the hit count for the logical *x* AND *y* search query as the only terminology parameters needed for assessing semantic similarity, but also includes the hit counts for the individual terms *x* and *y* before computing semantic similarity. Therefore, the following notations will be adopted in this research study to denote the different extracted terminology parameters:

f(*x*) denotes hit count for any queried term *x*

f(*y*) denotes hit count for any queried term *y*

f(*x*, *y*) denotes hit count for logically *x* AND *y* search query where both *x* and *y* appear together on the same Web page

The next sub-section explains how the different extracted terminology parameters *(f(x), f(y)* and *f(x, y))* are used to compute semantic similarity, which also forms Step 2 of Figure 8.1.

8.3.3 Computing Semantic Similarity Based on the Captured Terminology Parameters

In order to enhance communication among domain experts and also enable faster computation of meaning between computers in a computer digestible form, many long-term projects have been initiated to try and create semantic relationships between common entities or the names of these entities. Good examples of these projects include the CYC project (Lenat, 1995) and WordNet (Miller, 1995).

The aim is to come up with a semantic Web of such huge portions that elementary intelligence and comprehension about real-world entities come out without much effort.

However, to achieve this, systems have to be designed properly and be able to manipulate knowledge, and high quality contents have to be entered in these systems by well-educated experts.

While these attempts are good and take a long-term view, the overall amount of information entered is very small when compared to what is available on the Web today (Cilibrasi and Vitànyi, 2007). We therefore take advantage in this study of the freely available information on the Web and use it to measure semantic similarities between terminologies used in the digital forensic domain. It should be possible in future to replace the World Wide Web with a standardised digital forensic semantic repository for testing or implementing the proposed DFSR model proposed in this thesis.

Figure 8.2 shows the interface of the DFSR prototype developed to compute the semantic similarity between two identified digital forensic terminologies x = 'Digital Evidence' and y = 'Electronic Evidence'. Infer from Figure 8.2 that before computing the semantic similarity measure, the domain terminologies have to be identified, i.e. x = 'Digital Evidence' and y = 'Electronic Evidence'. The identification process forms Step 1 of Figure 8.1, followed by extracting the terminology parameters *(f(x), f(y)* and *f(x, y))* which involve Step 2 of Figure 8.1.

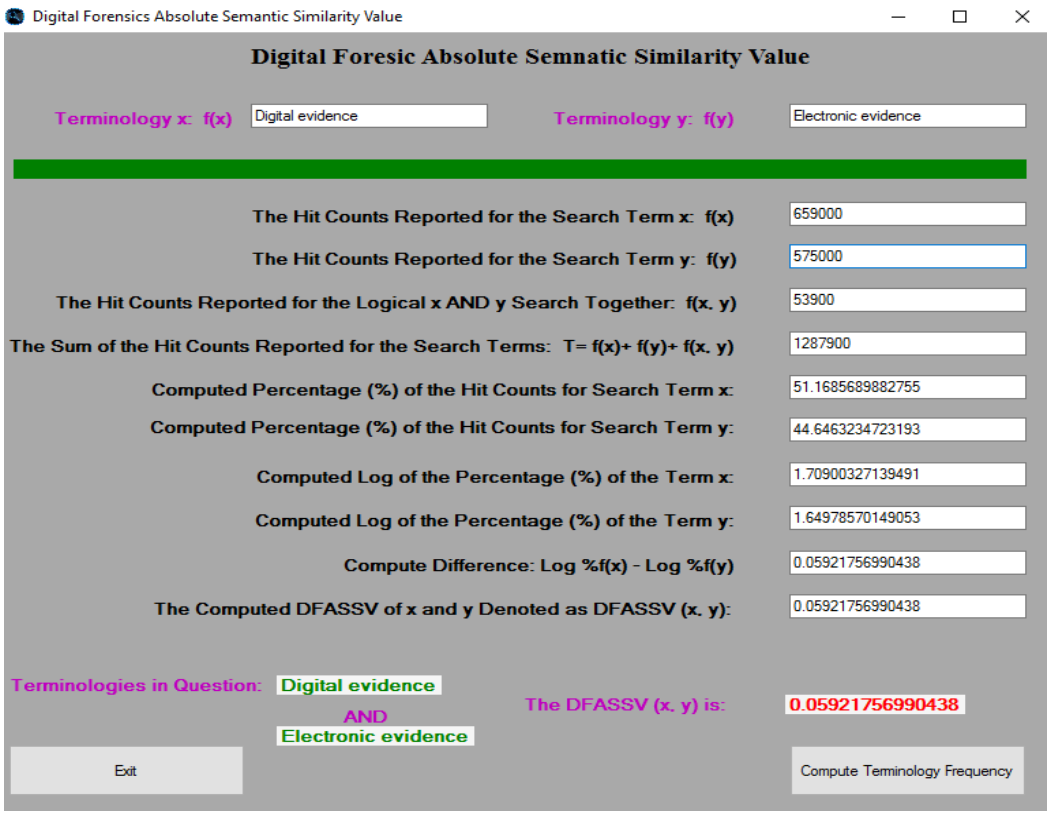


Figure 8.2 Computing Semantic Similarity using the DFSR Prototype

The presentation in Figure 8.2 is a new approach proposed in this study based on the DFASSV method to use the Web to compute semantic similarity values. The sub-sections to follow therefore explain in detail the technical background employed in this research to compute semantic similarity.

8.3.3.1 The Technical Background Employed for Computing the Terminology Semantic Similarity

An accurate measurement of semantic similarity between terminologies is a matter of concern in many different domains. For this reason, a method to assist in computing the semantic similarity between different digital forensic terminologies using the Web is proposed and explained in this section. The proposed method, coined as the Digital Forensic Absolute Semantic Similarity Value (DFASSV), uses the Web and Web search engines to compute semantic similarity between any identified pair of digital forensic terminologies.

Much of the theory employed in this regard is based on computing the Euclidean distance of any two given points in the Euclidean space, and its relationship with the computed absolute value of the difference of any given two real numbers on the number line (explained in detail later). Different distance functions result in different distance measures. However, the Euclidean distance used in this study is considered the most useful because it corresponds to the way objects are measured in the real world (Bailey, 2004).

In the sub-sections to follow, the concept of the Euclidean distance is explained, first followed by a discussion of the relationship between the Euclidean distances with the computed absolute value of the difference of any given two real numbers on the number line.

8.3.3.2 The Euclidean Distance

The Euclidean distance can be defined as the distance between any two given points in a plane that one can even measure with the help of a ruler and it is stated using the Pythagorean Theorem (Bogomolny, 2012a; Bogomolny, 2012b; Larose, 2005). If, for example, $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are two given points on the plane, then their Euclidean distance (d) can be defined as shown in Equation 1 (Sanchez et at., 2012).

$$\sqrt{(x_1\text{-}x_2)^2 + (y_1\text{-}y_2)^2} \qquad \text{Equation 1}$$

Using Equation 1 as distance, the Euclidean space can be viewed a metric space also called the distance space.

For any given two points *x* and *y,* the Euclidean distance between them is the distance of the line fragment connecting them. In a Cartesian coordinate, for example, if $x = (x_1, x_2,...,x_n)$ and $y = (y_1,y_2,\ldots, y_n)$ are two points in the Euclidean space, then the distance (*d*) from *x* to *y,* or from *y* to *x* can be defined by Equation 2,which can be seen as a generalisation of Equation 1 (Sanchez et at., 2012; Bogomolny, 2012a).

$$d(x,y) = d(y,x) = \sqrt{(x_1\text{-}y_1)^2 + (x_2\text{-}y_2)^2 + \cdots + (x_n\text{-}y_n)^2} \qquad \text{Equation 2}$$

where *n* represents any number denoting a point x_n and y_n in the Cartesian coordinate.

The location of any given point in the Euclidean n-space is usually called a Euclidean vector. Therefore, the points *x* and *y* can be referred to as Euclidean vectors. Beginning from the origin of the space, the tips of points *x* and *y* indicates the distance between these two points (also called the magnitude or the norm). The Euclidean norm or the distance of a vector *x* is a real number denoted as $||x||$ (McCracken, 2012) and measures the distance of *x* as defined by Equation 3 (Cengage, 2012):

$$||x|| = (x.x)^{1/2} = \sqrt{x.x} \qquad \text{Equation 3}$$

The distance between *x* and *y* can therefore be computed as shown in Equation 4 (Cengage, 2012):

$$d(x,y) = ||x - y|| \qquad \text{Equation 4}$$

The Euclidean norm and distance may as well be expressed in terms of components as shown in Equation 5 (McCracken, 2012; Cengage, 2012):

$$||x|| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x.x} \qquad \text{Equation 5}$$

If the length of a vector is considered as the distance from the end of a line to its head, then it becomes clear that the Euclidean length of a vector can also be treated as a special case of the Euclidean distance. Thus, the distance between *x* and *y* is the Euclidean length of the distance vector as defined in Equation 6 (McCracken, 2012):

$$\left|\left|\text{x-y}\right|\right| = \sqrt{(\text{x-y}).(\text{x-y})} \qquad \text{Equation 6}$$

Equation 6 is homogeneous to Equations 3, 4 and 5 and can be used to compute the magnitude or the norm of the numerical difference between any two real numbers *x* and *y* in the number line, denoted as $\left|\left|\text{x-y}\right|\right|$. It is also clear from Equation 6 that the one-dimensional Euclidean distance between *x* and *y* can be realised and this is briefly explained in the sub-section to follow.

8.3.3.2.1 One-dimensional Euclidean Distance

In the case of one dimension, the distance between any two given points *x* and *y* on the real number line is equivalent to the absolute value of their numerical difference. Thus, if *x* and *y* represent two real numbers, then the distance between them can be computed as shown in Equation 7 (Balu and Devi, 2011):

$$\sqrt{(\text{x-y})^2} = |\text{x-y}| \qquad \text{Equation 7}$$

In addition, in one dimension there is usually a single homogeneous, translation-invariant distance function, which is the Euclidean distance, and it defines the distance between elements of a set. Translation-invariant implies that starting from the origin, at least in one particular direction, the object is usually infinite. In higher dimensions, up to n-dimensions, there are other possible distance functions but these are beyond the scope of the research reported on in thesis. The researcher considered only up to the two-dimensional Euclidean distance in this thesis. However, future research should consider incorporating other possible distance functions of higher dimensions, up to n-dimensions.

8.3.3.2.2 Two-dimensional Euclidean Distance

In the Euclidean plane, if $x = (x_1, x_2)$ and $y = (y_1, y_2)$, then the distance (d) between the two points x and y is given by Equation 8 (Danielsson, 1980), which is homogeneous to Equations 1 and 2.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_1 - y_2)^2} \qquad \text{Equation 8}$$

For this reason, the discussion of the Euclidean distance – both one dimensional and two dimensional – presents us with the foundation of establishing its relationship with the absolute value of the difference of any given two real numbers in the number line (discussed in the sub-section to follow).

8.3.3.3 The Relationship between Euclidean Distance and Absolute Value

Having an understanding of the Euclidean distance, its relationship with the absolute value of any given real number x denoted as $|x|$ is now established in this section. The absolute value $|x|$ is the numerical value of x without consideration to its sign. For example, the absolute value of $+x$ is x, and the absolute value of $-x$ is also x. This simply means that the absolute value of any real number x may be viewed as its distance from zero (i.e. how far x is from zero on the number line) (Purplemath, 2012; Hotmath, 2012; IntAlgebra, 2012).

In practice, the absolute value of all real numbers is always positive, as shown in Equation 9. The concept of absolute value is closely associated with the notion of distance in various mathematical and physical contexts. In this thesis, though, the relationship between the Euclidean distance and the absolute value is established first. The established relationship is then used to generate a method coined as the Digital Forensic Absolute Semantic Similarity Value (DFASSV) for computing semantic similarity between any two terminologies x and y in the digital forensic domain. For any real number x, its absolute value denoted by $|x|$ can be defined as shown in Equation 9 (Ucalgary, 2012):

$$|x| = \begin{cases} x, & \text{if } x \geq 0 \\ -x, & \text{if } x < 0 \end{cases} \qquad \text{Equation 9}$$

Based on the definition shown in Equation 9, the absolute value of x is on all occasions either positive or zero, but never negative. In addition, the absolute value of the

difference of any two real numbers *x* and *y* defines the distance between *x* and *y* denoted as $|x - y|$, which is equivalent to the Euclidean distance of *x* and *y*. Since in mathematics the square root of a number *x* without considering its sign represents a positive square root, and the absolute value of *x* is invariably either positive or zero, but never negative, it follows that:

$$|x| = \sqrt{x^2} \qquad \text{Equation 10}$$

Equation 10 is homogeneous to Equation 7 and is at times used as a description of the absolute value of any real number (Anon, 2012c). The next sub-section elaborates on how to derive semantic similarity measures by using the Euclidean distance concepts and its relationship with the absolute value of the difference between any given two real numbers on the number line.

8.3.3.4 Deriving the Semantic Similarity Measures

Based on the discussions of the Euclidean distance and its relationship with the absolute value in the above sub-sections, it is clear that the absolute value of any real number can closely be associated with the concept of distance. The absolute value of any real number is the distance from that number to the origin, along the real number line (Balu and Devi, 2011). For any given two real numbers *x* and *y*, the absolute value of the difference between *x* and *y* is the distance between them. The standard Euclidean distance between any given two points, for example *x* and *y*, defined in Equation 2 affirms this, *where* $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. In the Euclidean n-space, the distance is defined as shown in Equation 11 (Balu and Devi, 2011):

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad \text{Equation 11}$$

Note that Equation 11 is homogenous to Equation 2. This can be viewed as a generalisation of $|x - y|$. Since *x* and *y* are two real numbers, from Equation 10 we can define Equation 12, where:

$$|x - y| = \sqrt{(x - y)^2} \qquad \text{Equation 12}$$

Equation 12 is homogeneous to Equation 7 and Equation 10. Equation 7 is, however, used when computing the one-dimensional Euclidean distance, while Equation 10 is

used as a definition of the absolute value. Thus, Equations 7, 10 and 12 can be used to prove that the 'absolute value' distance for any given real numbers is equivalent to the Euclidean distance defined in Equation 7, when you consider them as either one or two-dimensional Euclidean spaces defined in Equations 7 and 8 respectively. The next sub-section now elaborates on how to compute the semantic similarity values using the DFASSV method proposed in this study.

8.3.3.5 Computing Semantic Similarity Values by Using the Proposed DFASSV Method

Hit counts reported by Web search engines are useful information sources for this study and, as such, are used as input for computing semantic similarity measures by means of the proposed DFASSV method. This section therefore explains how the captured hit counts are used in this study to compute semantic similarity. As was mentioned earlier, the following notations are adopted in this research study:

$f(x)$ denotes the hit count for the queried term x

$f(y)$ denotes the hit count for the queried term y

$f(x, y)$ denotes the hit count for the logical x AND y search query where both x and y appear together on the same Web page

Re-calling the concept of the Euclidean distance and the absolute value of the difference between any given two real numbers on the number line, it is possible to establish in this thesis the relationship between these concepts with the proposed DFASSV method.

The proposed DFASSV method computes the semantic similarity value between two terms x and y in digital forensics, based on finding the one-dimensional Euclidean distance defined in Equation 7 equal to the absolute value of the difference between any two real numbers (as shown in Equations 7 and 12).

To begin with, the hit counts $f(x)$, $f(y)$, and $f(x, y)$ for any two digital forensic terminologies x and y are obtained using the Google search engine. To calculate the semantic similarity of x and y, we do not need to know the number of Web pages indexed by the Web search engine quoted as **8058044651** by Cilibrasi and Vitànyi, (2007). This is so because, according to Bar-Yossef and Gurevich (2006), the process of approximating the number of pages indexed by a search engine can be a very difficult undertaking.

Again, according to an official Google Blog (Google Blog, 2012), this number has increased significantly since 1998 when it was only 26 million. By 2000 the Google index had gone as far as the one billion mark. Over the last decade, this number has been changing and, recently, even the Google search engineers stopped calculating it due to the sheer vastness of the Web these days (Google Blog, 2012). The Google systems that process links on the Web recorded that 1 trillion distinctive URLs exist on the Web at once. Therefore, it is the researcher's opinion that, depending on this value (Web pages indexed value quoted as **8058044651**) might produce unreliable semantic similarity scores over time. The researcher will therefore not discuss the process of approximating the number of pages indexed by search engines in any further detail.

However, the number of pages indexed by search engines is replaced with a simple computed value (**T**), defined as the sum of the hit counts reported by the Web search engine for the search terminologies *x*, *y* and logical *x* AND *y* together.

Thus, $\mathbf{T} = f(x) + f(y) + f(x, y)$ Equation 13

where $f(x)$, $f(y)$ and $f(x, y)$ are as defined earlier. These parameters are then used as input into the proposed DFASSV method.

There exist four input parameters defined as $f(x)$, $f(y)$, $f(x, y)$ and **T**. Using Equation 12, which is similar to one-dimensional Euclidean distance; only two real numbers are needed as input. In order to establish a 1:1 mapping of the values of *x* and *y* in Equation 12, the DFASSV method replaces *x* and *y* with the percentage values of $f(x)$ and $f(y)$ computed as follows:

$\left(\frac{f(x)}{\mathbf{T}} * 100\right)$= percentage of the hit counts for the search term *x*

$\left(\frac{f(y)}{\mathbf{T}} * 100\right)$= percentage of the hit counts for the search term *y*

Substituting these values in Equation 12 gives Equation 14:

$$|x - y| = \sqrt{\left(\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \left(\frac{f(y)}{\mathbf{T}} * 100\right)\right)^2}$$ Equation 14

The value obtained from Equation 14 is in the fixed range of 0 per cent to 100 per cent. Treating the points *x* and *y* as Euclidean vectors and starting from the point of origin (0%) of the space, their tips (100%) shows the distance between the two points.

However, as mentioned earlier, in a one-dimensional Euclidean distance there is usually a single homogeneous, translation-invariant distance function (i.e. starting from the origin, at least in one particular direction the object is infinite). For the purpose of this study and for the sake of scalability of the semantic similarity measures, the terminologies that are paired using the proposed DFASSV method are therefore rated on a scale of 0 to ∞, where 0 denotes identical semantic similarity between the two terminologies and ∞ denotes no semantic similarity. For a similarity distance of 0 to ∞ instead of 0% to 100%, Equation 14 is further modified as follows:

The values $\left(\frac{\mathrm{f(x)}}{\mathrm{T}}*100\right)$ and $\left(\frac{\mathrm{f(y)}}{\mathrm{T}}*100\right)$, denoted as a percentage of the hit counts for the search term *x* and *y* respectively, are substituted by their computed logarithms as

$\log\left(\frac{\mathrm{f(x)}}{\mathrm{T}}*100\right)$ and $\log\left(\frac{\mathrm{f(y)}}{\mathrm{T}}*100\right)$ respectively.

A logarithm is a useful arithmetic concept used in all areas of science to help simplify the understanding of many scientific ideas. For example, logarithms may be defined and introduced in different ways as a means to simplify calculations. In this study, we adopt this simple approach to simplify the computation of the Euclidean distance, based on finding the absolute value of the difference between the logarithms of the hit count percentages of the terms *x* and *y*. There are no limits imposed on logarithms, thus their inputs and outputs can be in any range. Substituting these values in Equation 14 therefore leads to Equation 15:

$$|x-y| = \sqrt{\left(\log\left(\frac{f(x)}{\mathbf{T}}*100\right) - \log\left(\frac{f(y)}{\mathbf{T}}*100\right)\right)^2} \qquad \text{Equation 15}$$

Equations 14 and 15 are both analogous to Equations 7 and 12.

Equation 15 gives a value in the range of 0 to ∞ and can be re-written as Equation 16, which is used to automatically calculate the Digital Forensic Absolute Semantic Similarity Value of the terms *x* and *y* in digital forensics denoted as DFASSV(*x*, *y*).

Using the left-hand side of Equation 15, equivalent to the right-hand side we can define DFASSV(*x*, *y*) as shown in Equation 16:

$$\mathrm{DFASSV}(x,y) = \left|\log\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \log\left(\frac{f(y)}{\mathbf{T}} * 100\right)\right| \quad \text{Equation 16}$$

where:

$f(x)$ = the hit counts for the search term x

$f(y)$ = the hit counts for the search term y

$\mathbf{T}$ = the sum of hit counts for the search terms x and y as defined in Equation 13

Equation 16 therefore defines DFASSV(*x*, *y*), a new method for calculating semantic similarities between two terminologies *x* and *y* in digital forensics using the Web. In other words, Equation 16 denotes DFASSV as the computed absolute value of the difference between the logarithms of the hit count percentages of the terminologies *x* and *y*.

Figure 8.2 is used in this study to show the interface developed to implement the DFASSV(*x*, *y*) method shown in Equation 16. The experimental results obtained by using the new proposed DFASSV approach were found to be remarkable and are discussed in section 8.4 of this chapter. The next section explains Step 4 of Figure 8.1, which deals with generating semantic maps (Liaison, 2015) as part of the DFSR prototype implementation.

8.3.4 Generating Semantic Maps

Once the semantic similarities of the terminologies are computed, the next step is to generate a semantic map. As mentioned earlier, semantic mapping is used in this model to help display the meaning-based connections between the domain terminologies or phrases and a set of related terminologies or concepts. To achieve this, the model has to be integrated with an existing tool to aid in developing the semantic map. For the purpose of this study, the Knowledge Modelling Kit version 5.0.0.3 (see Figure 8.3) was used to generate the sample semantic maps. At this stage, the reader is again reminded that the integration between the DFSR prototype and the Knowledge Modelling Kit shown in Figure 8.3 is not automated and was tested manually. However, future versions will consider integrating the DFASSV and the Knowledge Modelling Kit to be able to work together harmoniously.

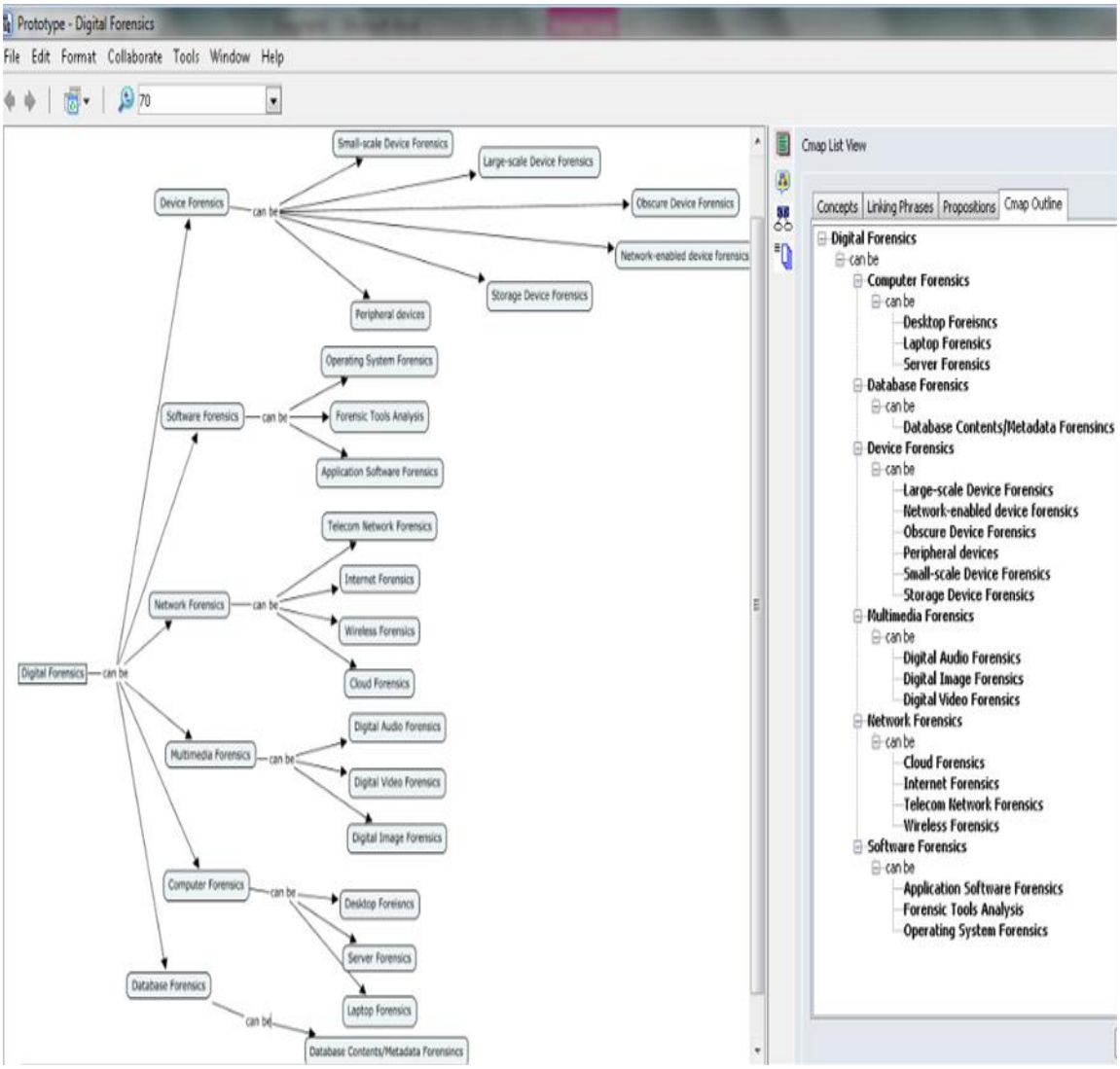


Figure 8.3 Knowledge Modelling Kit version 5.0.0.3 with a Sample Semantic Map

With the help of a map as the one shown in Figure 8.3, it is possible to show terminologies that are similar or related to one another, based on the computed semantic similarity measures. This can also help explain how the terminology meanings are categorised. Furthermore it will allow people to conceptually explore their knowledge of a new terminology by mapping it against other similar/related terminologies or phrases similar in meaning to the new terminology. Semantic integration is explained in the next section.

8.3.5 Semantic Integration Process

Semantic integration is primarily meant for interrelating information from diverse sources or different digital forensic tools. In the proposed DFSR prototype, however,

since this was the most challenging thing to implement, the researcher opted to use only one tool and consequently could not fully realise the integration of different tools in this phase. Future versions, however, will consider using various tools in order to test this phase in full. This is one of the limitations of the current prototype that need to be explored further in future versions. However, as said earlier, semantic integration is very important in automating communication between different systems using metadata publishing (Liaison, 2015). Besides, metadata publishing can potentially offer the ability to automatically link different domain ontologies as well as improve the quality of domain data through centrally governed, locally distributed, reusable operations that require reduced time and cost to merge digital forensic tools that work on the domain data. The next section explains the concept of sematic publishing which forms step 6 of the DFSR prototype.

8.3.6 Semantic Publishing Process

Once the semantic maps and semantic integration process is complete, the last step is to publish the information in a repository where it can be accessible to different software agents as well as individuals. To achieve this objective, the Ontology Web Language (OWL) is used to help in knowledge publishing. OWL, a computational logic-based language, is one among a number of knowledge representation languages or ontology languages for authoring ontologies or knowledge bases.

The primary features of OWL are formal semantics and RDF/XML-based serialisations for the Semantic Web. OWL enables knowledge expression that can be exploited using computer programs, e.g. to confirm the uniformity of that knowledge or to make implicit knowledge explicit. In addition, OWL documents, also known as ontologies, can be published on the World Wide Web and may refer to or be referred from other OWL ontologies as well. For this reason, OWL is fashioned to be used by tools that need to process the content of information instead of just presenting information to individuals. The Knowledge Modelling Kit shown in Figure 8.3 has the ability to change the semantic maps and the integrated data into the OWL before it is published.

Finally, to build an intelligent system to help resolve semantic disparities in digital forensics, semantic reconciliation is perhaps the most fundamental building block required. This is because semantic reconciliation can be used to show that two

terminologies are similar or related, despite having been described or represented differently in a context. Semantic reconciliation in digital forensics can also be used to create uniformity, for example, during evidence interpretation, presentation and reconstruction of domain data. It can also make the reporting of potential digital evidence much easier and more accurate by providing the most appropriate terminologies to use in any court of law or civil proceedings.

8.4 EXPERIMENTAL RESULTS BASED ON THE PROPOSED DFASSV METHOD

While the theories discussed in this research study might sound complicated, the resulting methods are simple enough. Knowing that there exists semantic disparity between two digital forensic terminologies *x* and *y*, for example, the computed absolute semantic similarity value denoted as the DFASSV (*x*, *y*) is used as a quick guide to show if the two terms are truly semantically related or not.

As mentioned earlier, for the purpose of this study and for the sake of the scalability of the semantic similarity measures, the terminologies that are paired using the proposed DFASSV method are rated on a scale of 0 to ∞, where 0 denotes identical semantic similarity between the two terminologies and ∞ denotes no semantic similarity. Therefore, given any two digital forensic terms *x* and *y*, we first find the number of hit counts for search term *x* denoted as *f*(*x*), the number of hit counts for the search term *y* denoted as *f*(*y*), the number of hit counts for the logical *x* AND *y* where both appear together on the same Web page denoted as *f*(*x*, *y*), and finally, the sum of hit counts denoted as (**T**). **T** is computed using Equation 13 as discussed earlier in section 8.3.3.5.

As a concrete example, consider the term *x* for 'digital evidence' and the term *y* for 'electronic evidence'. Using the Google search engine with hit counts as reported for the search terms *x* and *y* as on 14 April 2012, it follows that:

'Digital evidence' *f*(*x*) = **659,000**

'Electronic evidence' *f*(*y*) = **575,000**

'Digital evidence' AND 'Electronic evidence' *f*(*x*, *y*) = **53,900**

Therefore, **T**= *f*(*x*) +*f*(*y*) +*f*(*x*, *y*) = **1,287,900**

Substituting these values in Equation 16 as shown in Figure 8.4 below gives a semantic similarity measure of the terms 'digital evidence' and 'electronic evidence' of **0.0592**. Since this value is relatively close to zero, it proves that the two terms are very closely related in their meaning when used in digital forensics. It can also mean that, in the case

of a digital forensics investigation, the term ‘digital evidence’ can be used instead of ‘electronic evidence’ without misleading the receivers of such information.

$$\mathbf{DFASSV}(x,y) = \left|\log\left(\frac{f(x)}{\mathbf{T}} * \mathbf{100}\right) - \log\left(\frac{f(y)}{\mathbf{T}} * \mathbf{100}\right)\right| \qquad \text{Equation 16}$$

Where:

$f(x)$ = the hit counts for the search term x,

$f(y)$ = the hit counts for the search term y and

T = the sum of hit counts for the search terms x and y as defined in equation 13.

Figure 8.4 Computing the Semantic Similarity Measures using the DFASSV Method

To further analyse the performance of the proposed DFASSV method, the researcher conducted additional two sets of experiments. To start with the researcher compared the similarity scores generated by the proposed DFASSV method against the Miller and Charles benchmark dataset (Miller and Charles, 1998; Bar-Yossef and Gurevich, 2006). Secondly, the proposed DFASSV approach was tested using digital forensic domain terminologies to measure its performance against the human-perceived interpretation of the given terms. These two experiments are discussed in the sub-sections to follow.

8.4.1 The Miller and Charles Benchmark Dataset

To assess the implementation of the proposed DFASSV method, the researcher evaluated it first against the Miller and Charles dataset (Miller and Charles, 1998). The latter is a sub-set of Rubenstein and Goodenough’s original dataset of 65 word pairs (Rubenstein and Goodenough, 1965). The reason why the researcher used the Miller and Charles dataset in this study is because the Miller and Charles ratings are considered one of the most reliable benchmarks for evaluating semantic similarity measures (Bollegala et al., 2007; Vijay, 2012).

The term pairs using the proposed DFASSV method are rated on a scale of 0 to ∞ (infinite), where the value 0 means identical semantic similarity and ∞ means no similarity. This is the opposite of the Miller and Charles dataset where term pairs are

rated on a scale of 0 (dissimilarity) to 4 (identical semantic similarity) as shown in Table 8.1.

Infer from the results in Table 8.1 whose contents were sourced from Bo You et al., (2013) that the smaller the value computed using the proposed DFASSV method (shaded), the more similar the terms. For example, the word pair 'gem-jewel' in Table 8.1, with a similarity measure of **3.84** from Miller and Charles and **0.027** from the proposed DFASSV clearly depicts the accuracy of DFASSV. This is also depicted in the other columns of Table 8.1 and shows a better performance of DFASSV than some of the previous proposed methods. These results are also true as seen from the correlation coefficient value of **-0.2777**. (Note that according to Alastair, (2013) a negative correlation coefficient indicates that as one variable increases, the other decreases, and vice versa). The performance of the DFASSV method is further depicted by a graphical representation of the similarity measures in Table 8.1, shown in Figure 8.5.

Note also from Table 8.1 that the first column shows the word pairs used and column two indicates the ratings from Miller and Charles. Columns 3 to 6 sourced from Bo You et al., (2013) are used in this study to show a comparison and the ratings from previous proposed semantic similarity methods, while the last column (shaded) depicts the equivalence similarity measure computed using the proposed DFASSV method in this study.

Table 8.1Comparison of Semantic Similarity of Human Ratings and Baselines on Miller and Charles' Dataset with the Proposed DFASSV Method

Source: (You, Bo, Ting Ting He, and Fang Li., 2013)

Word Pair	**M&C**	**Web Jaccard**	**Web Dice**	**Web Overlap**	**Web PMI**	**Proposed DFASSV**
cord-smile	0.13	0.102	0.108	0.036	0.207	**0.756**
rooster-voyage	0.08	0.011	0.012	0.021	0.228	**0.828**
noon-string	0.08	0.126	0.133	0.060	0.101	**0.524**
glass-magician	0.11	0.117	0.124	0.408	0.598	**1.399**
monk-slave	0.55	0.181	0.191	0.067	0.610	**0.389**
coast-forest	0.42	0.862	0.870	0.310	0.417	**0.055**
monk-oracle	1.1	0.016	0.017	0.023	0	**0.457**
lad-wizard	0.42	0.072	0.077	0.070	0.426	**0.400**
forest-graveyard	0.84	0.068	0.072	0.246	0.494	**1.258**
food-rooster	0.89	0.012	0.013	0.425	0.207	**1.778**
coast-hill	0.87	0.963	0.965	0.279	0.350	**0.248**
car-journey	1.16	0.444	0.460	0.378	0.204	**0.865**
crane-implement	1.68	0.071	0.076	0.119	0.193	**0.418**
brother-lad	1.66	0.189	0.199	0.369	0.644	**0.970**
bird-crane	2.97	0.235	0.247	0.226	0.515	**0.051**
bird-cock	3.05	0.153	0.162	0.162	0.428	**0.024**
food-fruit	3.08	0.753	0.765	1	0.448	**0.223**
brother-monk	2.82	0.261	0.274	0.340	0.622	**0.966**
asylum-madhouse	3.61	0.024	0.025	0.102	0.813	**0.945**
furnace-stove	3.11	0.401	0.417	0.118	1	**0.180**
magician-wizard	3.5	0.295	0.309	0.383	0.863	**0.638**
journey-voyage	3.84	0.415	0.431	0.182	0.467	**0.238**
coast-shore	3.7	0.786	0.796	0.521	0.561	**0.411**
implement-tool	2.95	1	1	0.517	0.296	**0.838**
boy-lad	3.76	0.186	0.196	0.601	0.631	**0.271**
automobile-car	3.92	0.654	0.668	0.834	0.427	**0.975**
midday-noon	3.42	0.106	0.112	0.135	0.586	**0.855**
gem-jewel	**3.84**	**0.295**	**0.309**	**0.094**	**0.687**	**0.027**
Correlation	**1**	**0.259**	**0.267**	**0.382**	**0.548**	**-0.27 77**

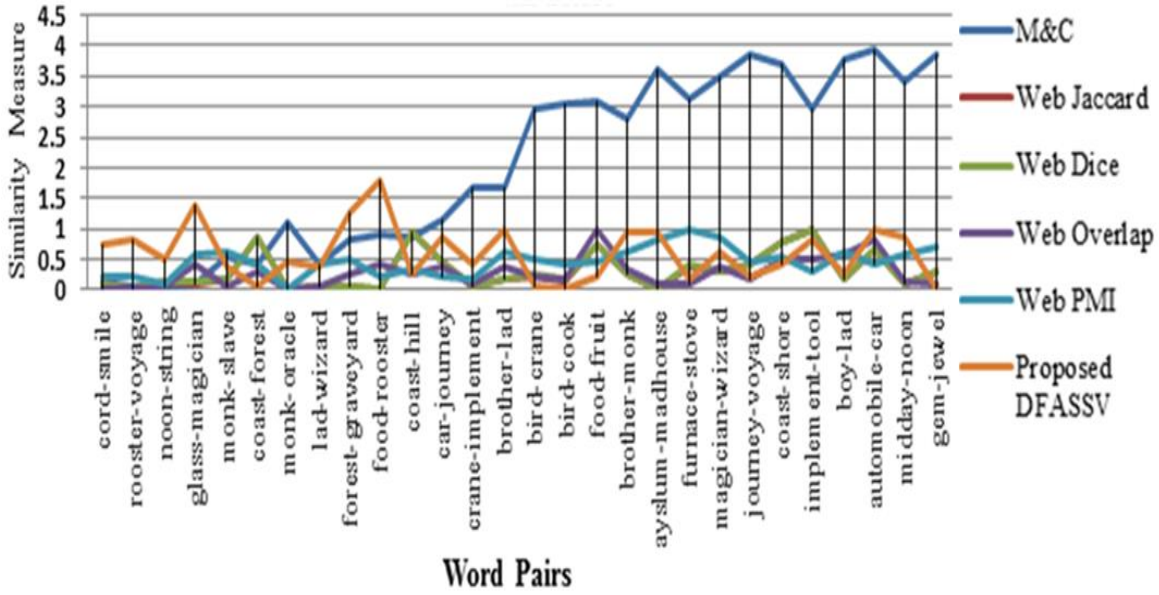


Figure 8.5 Comparison Graph of the Semantic Similarity Ratings and Baselines on Miller and Charles' Dataset with the Proposed DFASSV Method

Table 8.1 was also used in this study mainly for the purpose of comparison. It indicates different semantic similarity measures from previous methods (Bo You et al., 2013) compared to those of the proposed DFASSV method in this study. The comparison was made to provide a clear picture of the performance and accuracy of the DFASSV method.

The distance measure shown in Table 8.2, on the other hand, depicts the similarity of the digital forensic terms used in this research study and is discussed in the section to follow.

8.4.2 Experimental Results of Digital Forensic Terminologies Using the DFASSV Method

Table 8.2 presents a section of the semantic similarity results of digital forensic domain terminologies using the proposed DFASSV method. Each term shown in Table 8.2 was enclosed in double quotes " " and used as a single Google search term denoted in Table 8.2 as $f(x)$ and $f(y)$ respectively. The computed DFASSV(x, y) using Equation 16 shown in Figure 8.4 shows the semantic similarity or relatedness obtained to ascertain the performance of the DFASSV method with the human-perceived interpretation of the terminologies in question.

Table 8.2 Semantic Similarity Ratings of Digital Forensic Terms Based on the Proposed DFASSV Method

Digital Forensic Term Pairs		**Computed DFASSV(*x*, *y*)**
f(x)	***f(y)***	
Digital evidence	Electronic evidence	**0.059217**
Digital forensics	Digital evidence	**0.431534**
Digital forensics	Electronic evidence	**0.490752**
Electronic evidence	Digital and multimedia evidence	**1.833840**
Digital evidence	Digital and multimedia evidence	**1.893057**
Digital forensics	Digital and multimedia evidence	**2.324592**
Attacker	Adversary	**0.357051**
Cracker	Attacker	**0.361608**

The researcher has no knowledge of other experiments of this kind in the digital forensic domain that can be used as a baseline to judge the performance of the DFASSV method. It was therefore considered a new approach to use the Web to determine the semantic similarity of terminologies in digital forensics.

The selected terminologies used in this study are inter alia ‘digital evidence’, ‘digital forensics’, ‘electronic evidence’, ‘digital and multimedia evidence’ (Palmer, 2001; NATA, 2012). The researcher found that these terms are mostly used in discussions that involve the digital forensic investigation process and digital evidence presentations in legal proceedings; hence, the motivation for the experiment indicated in Table 8.2. In all the experiments that were conducted, DFASSV showed satisfactory results.

The proposed DFASSV was used to determine the semantic similarity measure of the terminologies as shown in Table 8.2. The first two columns of Table 8.2 show the digital forensic term pairs used for the experiments and their equivalent similarity measure is indicated in the last column. In the case of a digital forensic investigation process, for example, the proposed DFASSV method can be used to determine the relatedness of terminologies where a similarity measure closer to zero means that the two terms are closely related in meaning. This is further depicted by a graphical representation of the similarity measures in Table 8.2, shown in Figure 8.6.

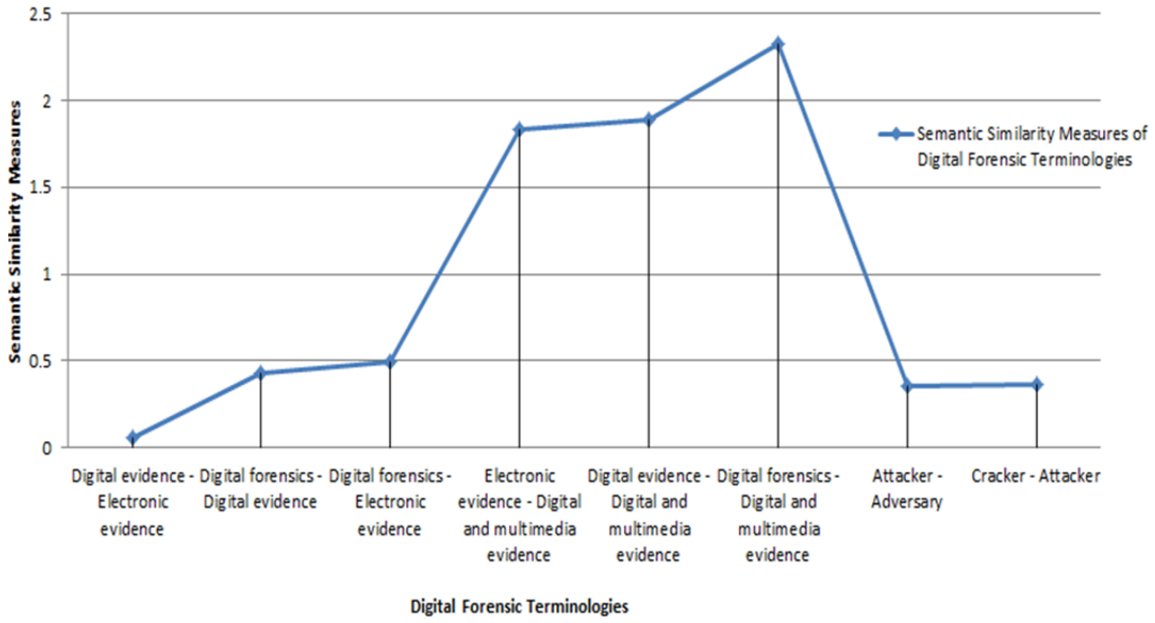


Figure 8.6 Semantic Similarity Measures Based on the DFASSV Method

Infer from Figure 8.6 that the terms ‘digital evidence’ and ‘electronic evidence’, with a similarity measure of **0.059217,** can be used interchangeably without causing confusion to the stakeholders. On the other hand, a semantic similarity measure far from zero would mean that the two terms are not closely related in meaning. Therefore, one term cannot replace the other. For example, the terms ‘digital forensics’ and ‘digital and multimedia evidence’ with a similarity value of **2.324592** means they cannot be used interchangeably. The section to follow briefly explains the use of the proposed DFASSV method in digital forensics.

8.4.3 Application of the Proposed DFASSV Method in Digital Forensics

The proposed DFASSV method as demonstrated by the experimental results in this chapter can be used in the digital forensic domain to determine for instance the semantic similarity of domain terms and also to resolve semantic disparities that exist in the domain. In addition, the DFASSV method can be used to help determine the most relevant and appropriate terminologies to use or include, for example, when building digital forensic domain ontologies. Besides, other future relevant undertakings in the digital forensic domain might benefit from applying a method such as the DFASSV proposed in this thesis. The next section presents more experimental results based on the DFSR prototype.

8.5 MORE EXPERIMENTAL RESULTS BASED ON THE DFSR PROTOTYPE

To further analyse the performance of the proposed DFSR prototype, Table 8.3 shows more results based on the DFSR prototype. Using the Google search engine with the hit counts reported for the different terminologies as on 23 November 2013, the following results as shown in Table 8.3 were recorded.

Table 8.3 Similarity Measures using the DFSR Prototype

Digital Forensic Terminologies		**Similarity measures using the DFSR prototype**
f(x)	***f(y)***	
Computer Forensics	Computer Forensics	**0**
Network Forensics	Network Analysis	**0.68434591282729**
Device Forensics	Mobile Device Forensics	**0.84438432665412**
Device Forensics	Media Analysis	**1.56290786736582**
Mobile Device Forensics	Media Analysis	**0.7185235407117**
Software Forensics	Code Analysis	**1.50323179037887**
Audio Forensics	Forensic Audio	**0.29275746969799**
Video Forensics	Video Analysis	**1.41948944913578**
Image Forensics	Image Analysis	**1.27437780089254**
Digital Audio Forensics	Forensic Audio	**0.48811663902112**
Digital Video Forensics	Video Analysis	**0.53494866504757**
Digital Image Forensics	Image Analysis	**0.94338458185111**
Multimedia Forensics	Multimedia Evidence	**0.85017396742686**

Table 8.3 is also used in this section to depict the performance of the DFSR prototype developed in this study. Infer from Table 8.3 that the closer to zero the semantic similarity measures, the more similar the terminologies, and vice versa. Table 8.3 is supported by a graphical representation of the data in Figure 8.7 as shown below, followed by a summary of this chapter.

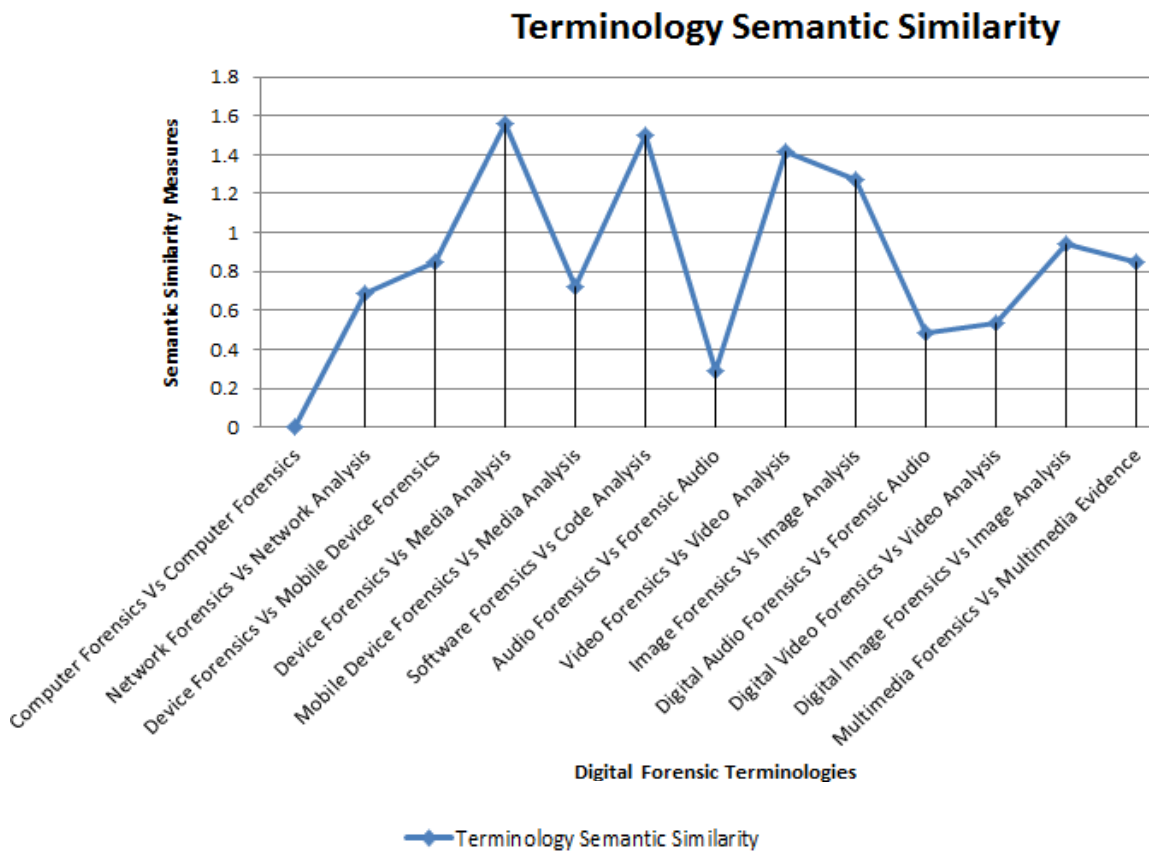


Figure 8.7 Similarity Measures using the DFSR Prototype

8.6 CHAPTER CONCLUSION

This chapter demonstrated the feasibility of the DFSR model in a prototype implementation called the DFSR prototype. In the case of semantic conflicts between digital forensic domain terminologies, the DFSR prototype can thus be employed to help with semantic reconciliation. The DFSR prototype tested in this study also helped to demonstrate the most fundamental features required for resolving semantic disparities in digital forensics. As mentioned earlier, not all the parts of the DFSR prototype are fully automated at this stage; hence, manual activities were required in some of the steps shown in Figure 8.1.

The DFSR prototype, as demonstrated in this Chapter, can be used in the digital forensic domain, for example, as a quick guide towards semantic reconciliation. In addition, the DFSR prototype can be used to help determine the most relevant and appropriate terminologies to adopt during the presentation of domain knowledge and information to different stakeholders. The experimental results to test the efficiency and feasibility of the proposed methods that were used to implement the DFSR prototype in

this chapter delivered remarkable results. This chapter also presented and explained the experimental results based on the proposed DFASSV and the DFSR prototype developed in this study. The experiments conducted using the proposed methods delivered impressive results.

The DFSR prototype, for example, can be used as a quick guide for resolving semantic disparity in digital forensics because, as a method that uses semantic similarity measures to resolve semantic disparities, it produces satisfactory results. In the researcher's opinion, other future relevant undertakings in the digital forensics domain might also benefit from applying the concepts used to implement the DFSR prototype in this study.

CHAPTER 9 : CONCLUSIONS AND FUTURE WORK

9.1 INTRODUCTION

The need for developing methodologies and specifications that can be used to resolve semantic disparities in digital forensics was identified in this thesis. In a world where digital technology keeps changing and the digital forensic domain evolves continuously, it would be appropriate to develop methodologies and specifications to take care of the semantic disparities that are bound to occur in the domain.

As a conclusion to this thesis, Chapter 9 discusses further research work that might be worthwhile in this area of study. Section 9.2 revisits the initial problem statement in brief, while Section 9.3 presents the accomplishments. Potential future research work is mentioned in Section 9.4, followed by a concluding summary of this chapter in Section 9.5.

9.2 REVISITING THE PROBLEM STATEMENT

The main problem addressed in this research involved the lack of methods and specifications specifically designed for resolving semantic disparities in the digital forensic domain. To address this problem, the researcher first set out to discuss the different challenges faced by the digital forensic domain. The study then went further to explain semantic disparities in digital forensics. Several potential causes of semantic disparities were also discussed, including identified approaches to manage semantic disparities in the digital forensic domain. Attention was also given to the significance of resolving semantic disparities to computer professionals, law enforcement agencies and other digital forensic practitioners.

Besides semantic disparities in digital forensics, several ontologies were proposed and explained in this study in an attempt to better organise digital forensic domain knowledge and explicitly describe the domain information and semantics in a common way. This implies that ontologies as presented in this thesis can be used to specify common vocabularies with which to make assertions, as well as analyse digital forensic domain information and knowledge.

Developing ontologies that clearly explain the conventional entities in which shared knowledge can be represented in digital forensics can help to create uniformity and a common understanding in the domain. Moreover, ontologies in digital forensics

can be used as a way towards resolving some of the semantic disparities that exist in the domain. Uniformity and a common understanding can also enhance or improve cooperation among computer professionals, law enforcement agencies and other digital forensic stakeholders in the case of an investigation process.

This research study proposed a systematic Digital Forensic Semantic Reconciliation (DFSR) model in an attempt to provide direction in resolving semantic disparities in the digital forensic domain. This model can be used to develop new methods and techniques for detecting and managing semantic disparities in the digital forensic field. The DFSR model can as well be incorporated as part of existing tools to help create uniformity and a common understanding of domain data or information, especially during the interpretation, description and representation of potential digital evidence.

As a way to evaluate the proposed DFSR model, a prototype known as the DFSR prototype was developed as part of this research by implementing some of the essential concepts of the DFSR model. The primary objective of developing the DFSR prototype was to demonstrate that, with an active and standardised digital forensic semantic repository, it is possible for individuals and digital forensic experts to capture different terminology parameters (from the repository) that can be used for semantic reconciliation in digital forensics. The experiments conducted in this research using the DFSR prototype delivered remarkable results and confirmed that DFSR prototype can be used as a quick guide for resolving semantic disparity in the digital forensic domain.

The next section elaborates on the accomplishments of the current research.

9.3 ACCOMPLISHMENTS

The research presented in this thesis has notable contributions in the digital forensic domain. Some of the accomplishments of this study are as explained in the sub-section to follow:

9.3.1 Ontologies for the Digital Forensic Domain

This study managed to demonstrate that ontologies play a critical role in knowledge sharing in digital forensics, for example, ontologies can make the processes of domain knowledge storage and retrieval significantly more intelligent. For this reason it seems important that many more ontologies should be developed in digital forensics. In Chapter 4, the researcher introduced the reader to the concept of ontologies as computing

models that can help in capturing domain knowledge as well as give a uniformly agreed-upon understanding of the domain information. Such domain information can also be reused and shared across different kinds of people. Chapter 4 also formed the basis for developing the ontologies discussed in Chapter 6.

The proposed ontologies in Chapter 6 can be used in the digital forensics domain to address issues of concern such as professional specialisation and certification, including development of forensics tools, curricula and education materials.

The digital forensic disciplines and sub-disciplines presented in the ontologies in Chapter 6 were found to be useful in giving direction to individuals interested in specific areas of professional specialisation. Such areas will, for example, produce specialists in computer forensics, software forensics, database forensics, multimedia forensics, device forensics and network forensics. While specialisation is important, certification cannot be ignored, especially not by individuals interested in the industry practices of digital forensics.

Developers of digital forensics tools can also use the ontologies to fine-tune such tools so as to cover as many sub-disciplines as possible in the case of digital forensic investigations. This implies that developers will find the ontologies useful, especially when considering the development of new digital forensic techniques for specific areas of interest and new high-tech digital forensic investigation tools.

Institutions of higher learning could also benefit from the ontologies in Chapter 6, especially when developing curricula and educational materials for different undergraduate and postgraduate studies. Different modules can be developed with the help of the ontologies to assist students in comprehending the concepts of digital forensics effortlessly. Prerequisites for modules can, in addition, be designed effectively with the help of the ontologies so as to avoid conflicts among and redundancy of concepts.

With the emergence of cloud computing technologies, the need for cloud forensics has become essential and the ontology for a cloud forensic environment was therefore presented in Chapter 6. Such ontology can be used as a common platform to share coherent cloud computing concepts and also promote the understanding of the cloud environments and cloud components. The ontology can furthermore serve as a basis for sharing common views on the structure and depiction of cloud computing information in a bid to enable the reuse of domain knowledge.

The ontology can help investigators to explicitly describe investigation processes and procedures that focus on specific cloud environments. Developers will also find the ontology constructive, especially when considering new cloud forensic techniques for specific cloud environments.

In the case of cloud forensics, the proposed ontology can assist in the design and development of acquisition tools that incorporate for example hybrid cloud architectural designs with shareable features such as automated acquisition, reporting, visualisation and presentation of digital evidence in a way that is admissible in court. Such tools will enhance the investigation of criminal cases involving multiple cloud computing environments.

However, since developing digital forensic domain ontologies is not an easy task, the researcher recommends that, whenever digital forensic domain ontologies are developed, digital forensic experts, computer professionals and law enforcement agencies should collaborate. In fact, collaboration and cooperation among digital forensic experts, computer professionals and law enforcement agencies is not optional if the developed ontologies are to be reused for other substantial developments in digital forensics.

9.3.2 Approaches to Manage Semantic Disparities in Digital Forensics

The concept of semantic disparities in the digital forensic domain was presented in Chapter 5. The advances in semantic disparity research, as well as the potential causes of semantic disparities in digital forensics were also explained.

However, Chapter 5 also elaborated on how to manage semantic disparities in the digital forensics domain. The different approaches found to be helpful in managing semantic disparities were identified and explained, this included a discussion on the significance of resolving semantic disparities in the digital forensic domain. Lastly, semantic reconciliation as shown in Chapter 5 is in the researcher's opinion a promising conception towards resolving semantic disparities that are apparent in the digital forensic domain.

9.3.3 A Taxonomy of the Digital Forensic Challenges

The presentation of the taxonomy of challenges for digital forensics in Chapter 3 was another exceptional accomplishment in this research study. The taxonomy is a contribution towards advancing knowledge in digital forensic challenges and founded on

the study of existing digital forensic literature. The taxonomy further classifies the large number of digital forensic challenges into a few well-defined and easily understood categories. Note that the taxonomy was designed taking into consideration the major challenges that digital forensics has faced over the past decade. However, more specific categories and sub-categories of the challenges can and should be added as the need arises in future.

Despite numerous researchers and digital forensic stakeholders having studied and examined different known challenges in digital forensics, there existed a need for a formal classification of those challenges. It is for this reason that, Chapter 3 surveyed existing research literature, identified and explained different digital forensic domain challenges that digital forensics had faced over the past decade. A taxonomy of the various challenges was subsequently proposed as a new contribution to this study field.

The proposed taxonomy can for example be useful in future developments of automated digital forensic tools by explicitly describing processes and procedures that focus on addressing the specific challenges identified in this research. Moreover, the taxonomy can help to map and categorise different digital forensic challenges, as well as create a common platform to share information in the digital forensic domain.

9.3.4 Proposed Digital Forensic Semantic Reconciliation (DFSR) Model

The Digital Forensic Semantic Reconciliation (DFSR) model for resolving semantic disparities in the digital forensic domain was also an exciting achievement. It was proposed in Chapter 7 as a way towards resolving semantic disparities in digital forensics, and supported by the implementation of a DFSR prototype in Chapter 8. The DFSR model was proposed in an attempt to provide direction for resolving semantic disparities in the digital forensic domain. While the theory of the model might sound complicated, the resulting DFSR model and the DFSR prototype presented in Chapter 8 are simple enough and the model can be employed as a quick guide towards resolving semantic disparities in the digital forensic domain.

The DFSR model can also be used to develop new methods and techniques for detecting and managing semantic disparities in the digital forensic domain. It can furthermore be incorporated as part of existing tools to help create uniformity and a common understanding of domain data or information – especially during the interpretation, description and representation of potential digital evidence. This belief

has been confirmed because the experiments conducted with the DFSR prototype discussed in Chapter 8 delivered remarkable results.

The reader should note that the initial experiments carried out using the DFSR prototype mostly focused on the digital forensic domain. However, the researcher believes that the proposed DFSR model can be applied in other domains as well, because it does not require any human-annotated knowledge.

9.3.5 Implementation of the Digital Forensic Semantic Reconciliation (DFSR) Prototype

In Chapter 8, the researcher also demonstrated the feasibility of the DFSR model in a prototype implementation called the DFSR prototype. The latter served to provide the fundamental specifications to test the DFSR model. The DFSR prototype implemented the most essential components of the DFSR model that were found necessary to demonstrate the feasibility of the DFSR model in digital forensics. The proposed model was meant to help produce satisfactory results, in other words results that are similar in interpretation, description and representation of the different domain data or information and terminologies. The results based on experiments with the proposed DFSR prototype were also discussed in Chapter 8 as well and found to be both satisfactory and remarkable.

The next section presents the opportunities for future research work, based on the research reported on in this thesis.

9.4 FUTURE RESEARCH

The current research revealed a number of topics in which further research could be worthwhile. Although the proposed DFSR model achieved the study objectives to the extent described in Chapters 7 and 8, it suffers from some limitations. Fortunately, these limitations are the basis of opportunities to extend the work discussed here in this research thesis through different potential research projects as indicated below:

- Further research is possible aimed at finding new techniques on how to fully automate the semantic reconciliation process in digital forensics. In the current research project some parts of the semantic reconciliation process were still done manually. This also means that, the experiments conducted to test the DFSR model covered only a few selected terminologies from the digital forensic domain, which served as a test bed in this study. For a comprehensive evaluation of the DFSR model, more experiments need to be conducted to determine its scalability.

- Again as part of the future work in line with this study the researcher intends to conduct a study that will try a human-based evaluation of the similarities using forensic practitioners/researchers to ascertain the results of the DFSR model
- Additionally, one might ask whether there exist other methods of implementing intelligent techniques for the semantic reconciliation process in digital forensics. The technique used in this research project is apt to be enhanced by introducing techniques such as neural network computational models that are capable of machine learning, more specifically, for grammatical induction, also known as grammatical inference and semi-supervised learning.
- The number of challenges faced by digital forensics is significantly huge. With the continued developments and research in digital forensics, much research needs to be carried out to provide direction on how to address many of the challenges faced by digital forensics today. More research also needs to be conducted to improve on the taxonomy of digital forensic challenges proposed in Chapter 3 of this thesis. This should also spark discussion on the development of new digital forensic taxonomies.
- Seeing that digital forensics lacks a standardised semantic repository, research needs to be carried out towards developing a standardised semantic repository, corpus or digital library that includes ontologies that can be used for resolving any disparities in digital forensics.
- Finally, more research is required on how to address current and future disparities in the digital forensic domain, as well as to supplement the work discussed in this thesis. The findings presented here constitute a whole new contribution towards advancing the digital forensics domain.

The next section presents a summary of this chapter and serves as a final conclusion of the research work presented in thesis.

9.5 CHAPTER CONCLUSION

Chapter 9 presented concluding remarks based on the research reported on in this thesis. These include revisiting the problem statement and the accomplishments towards resolving the identified semantic disparities in digital forensics, as well as a summary of potential future research opportunities in this study area.

Finally, when all is said and done, it is the researcher's hope that the work presented in this thesis will inspire many other researchers, practitioners and stakeholders dedicated to digital forensics to further explore this area of study. In the opinion of the researcher, this work has confirmed that time, patience, open-mindedness and the willingness to face challenges are key factors to the successful resolution of issues such as semantic disparities in digital forensics. In conclusion, the researcher also found that the time invested in issues such as the one presented in this thesis has been well worth the effort.

BIBLIOGRAPHY

Aberdeen, (2012). Petabyte of Storage Capacity. Available at: http://www.aberdeeninc.com/abcatg/petabyte-storage.htm [Accessed August 26, 2012].

Abraham, T., (2006). Event sequence mining to develop profiles for computer forensic investigation purposes. In ACSW Frontiers '06: Proceedings of the Australasian Workshops on Grid Computing and E-research, pp. 145–153. Australian Computer Society, Inc. Darlinghurst, Australia, ISBN 1-920-68236-8.

Abrams, S.M. and Weis, P.C., (2003). Knowledge of Computer Forensics is Becoming Essential for Attorneys in the Information Age, New York State Bar Journal February.

AccessData, (2012). Computer Forensics Software for Digital Investigations. Available at: http://accessdata.com/products/digital-forensics/ftk [Accessed August 29, 2012].

AccessData, (2013). Decryption & Password Cracking Software. Available at: http://www.accessdata.com/products/digital-forensics/decryption [Accessed February 21, 2013].

ACPO, (2013). Good Practice Guide for Computer-Based Electronic Evidence. Available at: http://www.7safe.com/electronic_evidence/ACPO_guidelines_computer_ evidence.pdf [Accessed February 16, 2013].

Adam W., (2015). What is Taxonomy…and how to do it well. Available at: http://www.rkosolutions.com/what-is-taxonomy/ [Accessed November 20, 2015].

Adobe Photoshop, (2012). Image editor software | Adobe Photoshop CS6. Available at: http://www.adobe.com/products/photoshop.html [Accessed August 26, 2012].

Adobe Premiere, (2012). Video editing software | Adobe Premiere Pro CS6. Available at: http://www.adobe.com/products/premiere.html [Accessed August 26, 2012].

Afcea, (2014). Cybercrime Forensics in Today's World. http://www.afcea.org/calendar/eventdet.jsp?event_id=34234&w=N. [Accessed August 08, 2015].

Alastair, L.D. (2013). Mastering Financial Mathematics in Microsoft Excel: A Practical Guide for Business Calculations. Published by Pearson UK.

Alazab, M., Venkatraman, S. and Watters, P., (2009). Digital forensic techniques for static analysis of NTFS images, Proceedings of ICIT 2009, Fourth International Conference on Information Technology, IEEE Xplore.

Alharbi, S., Weber-Jahnke, J. and Traore, I., (2011). The Proactive and Reactive Digital Forensics Investigation Process: A Systematic Literature Review. International Journal of Security and its Applications. Vol. 5, No. 4, pp. 59-71.

Allyson M.H. and Doris L.C., (2009). Weaving Ontologies to Support Digital Forensic Analysis. ISI 2009, Richardson, Texas.

Amazon, (2013). Amazon Elastic Compute Cloud (Amazon EC2), Cloud Computing Servers. Available at: http://aws.amazon.com/ec2/ [Accessed September 27, 2013].

Anon, (2012). Computer forensics. Anglia Ruskin University, Dissertation No (CSH2998A). Available at: http://www.minshawi.com/other/computer%20forensics.pdf [Accessed March 5, 2012].

Anon, (2013). Computer Forensics Privacy Issues. Available at: http://www.computerforensics1.com/privacy-computer-forensic.html [Accessed February 23, 2013].

Anon, (2013a). Computer Forensics Privacy Issues. Available at: http://www.computerforensics1.com/privacy-computer-forensic.html [Accessed February 23, 2013].

Anon, (2013b). A Communication Model. Available at: http://www.worldtrans.org/TP/TP1/TP1-17.HTML [Accessed April 25, 2013].

Arnold, E. and Soriano, E., (2013).The Recent Evolution of Expert Evidence in Selected Common Law Jurisdictions Around the World. A commissioned study for the Canadian Institute of Chartered Business Valuators.

Arthur, K.K. and Venter, H.S., (2004). An Investigation into Computer Forensic Tools. Proceedings of the ISSA Conference 2004. Midrand, South Africa.

Ashcroft, J., (2001). Electronic Crime Scene Investigation: A Guide for First Responders, Second Edition. Available at: https://www.ncjrs.gov/pdffiles1/nij/219941.pdf [Accessed August 6, 2013].

Avigdor G., Ateret A., Alberto T. and Danilo M., (2003). A framework for modeling and evaluating automatic semantic reconciliation. The VLDB Journal (2003).(DOI) 10.1007/s00778-003-0115-z

Aye, N., Hattori, F. and Kuwabara, K., (2008). Use of Ontologies for Bridging Semantic Gaps in Distant Communication. Proceedings of the International Conference on Innovations in Information Technology, (IIT 2008), pp. 371–375.

Ayers D., (2009). A second generation computer forensic analysis system. Digital Investigation: The International Journal of Digital Forensics & Incident Response, Vol. 6, pp. S34–S42.

Bailey, D.G., (2004). An Efficient Euclidean Distance Transform, IWCIA 2004, LNCS 3322, pp. 394–408.

Baker, T., Dekkers, M., Heery, R., Patel, M. and Salokhe, G., (2001). What Terms Does Your Metadata Use? Application Profiles as Machine-Understandable Narratives, Journal of Digital Information. Available at: http://jodi.ecs.soton.ac.uk/ Articles/v02/i02/Baker/

Bakshi, U.A., Bakshi, A.V. and Bakshi, K.A., (2008). Time and Frequency Measurement, Technical Publications, (Electronic Measurement Systems.) First Edition, pp. 4-1. Available at http://books.google.co.za/books?id=jvnI3Dar3b4C&pg=PT183&hl =en#v=onepage&q&f=false

Balu, R. and Devi, T., (2011). Identification of Acute Appendicitis Using Euclidean Distance on Sonographic Image. International Journal of Innovative Technology & Creative Engineering (ISSN: 2045-8711). Vol. 1, No. 7.

Barry Smith (2003). "Chapter 11: Ontology". In Luciano Floridi, ed. Blackwell Guide to the Philosophy of Computing and Information(PDF). Blackwell. pp. 155–166. ISBN 0631229183.

Barske, D., Stander, A. and Jordaan, J., (2010). A Digital Forensic Readiness Framework for South African SME's, Proceedings of ISSA Conference 2010, South Africa.

Baryamureeba, V. and Tushabe, F., (2004). The Enhanced Digital Investigation Process Model. Proceedings of the Digital Forensic Research Workshop (DFRWS), Baltimore, Maryland.

Bar-Yossef, Z. and Gurevich, M., (2006). Random sampling from a search engine's index. Proceedings of the 15th International World Wide Web Conference.

Bassett, R., Bass, L. and O'Brien, P., (2006). Computer Forensics: An Essential Ingredient for Cyber Security. Journal of Information Science and Technology, pp. 22–32.

Beebe, N.L. and Clark, J.G., (2005a). Dealing with terabyte data sets in digital investigations. Advances in Digital Forensics, pp. 3–16. Springer.

Beebe, N.L. and Clark, J.G., (2005b). A Hierarchical, Objectives-Based Framework for the Digital Investigations Process. Digital Investigation, 2(2), pp. 146–16.

Beham, G., (2012). Incident Detection and Cloud Forensics – Security at a Glance. Available at: http://ipbr.wordpress.com/2012/08/30/incident-detection-and-cloud-forensics/ [Accessed February 16, 2013].

Bennett, W.D., (2011). The Challenges Facing Computer Forensics Investigators in Obtaining Information from Mobile Devices for Use in Criminal Investigations. Available at: http://articles.forensicfocus.com/2011/08/22/the-challenges-facing-computer-forensics-investigators-in-obtaining-information-from-mobile-devices-for-use-in-criminal-investigations/ [Accessed February 16, 2013].

Bill Poser (2007) The Supreme Court Fails Semantics. Available at: http://itre.cis.upenn.edu/myl/languagelog/archives/004696.html [Accessed August 26, 2014]

Bishr, Y.A., (1998). Overcoming the Semantics and Other Barriers to GIS Interoperability. International Journal of Geographic Information Science, Vol. 12, No. 4, pp. 299–314.

Bishr, Y.A., Pundt, H., Kuhn, W. and Radwan, M., (1999). Probing the Concept of Information Communities – A First Step toward Semantic Interoperability. Interoperating Geographic Information Systems, Kluwer Academic.

Bo You, Tingting He, and Fang Li., (2013). A web Based Method For Measuring Semantic Relatedness between Words. Applied mechanics and Material 2013. Vols. 347-350, Pp 783-787

Boddington, R., Hobbs, V. and Mann, G., (2008). Validating Digital Evidence for Legal Argument. Proceedings of the 6th Australian Digital Forensics Conference, Edith Cowan University, Perth, Western Australia.

Bogomolny, A., (2012a). Pythagorean Theorem and its many proofs. Interactive Mathematics Miscellany and Puzzles. Available at: http://www.cut-the-knot.org/pythagoras/index.shtml [Accessed May 7, 2012].

Bogomolny, A., (2012b). The Distance Formula. Interactive Mathematics Miscellany and Puzzles. Available at: http://www.cut-the-knot.org/pythagoras/ DistanceFormula. shtml [Accessed May 7, 2012].

Böhme, R., Freiling, F.C., Gloe, T. and Kirchner, M., (2009). Multimedia Forensics is not Computer Forensics. Third International Workshop in Computational Forensics, IWCF 2009, The Hague.

Bollegala, D., Matsuo, Y. and Ishizuka, M., (2007). An Integrated Approach to Measuring Semantic Similarity between Words Using Information available on the Web. Association for Computational Linguistics. Proceedings of NAACL HLT 2007, pp. 340–347.

Bollegala, D., Matsuo, Y. and Ishizuka, M., (2009). A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web.

Boyce, S. and Pahl, C., (2007). Developing Domain Ontologies for Course Content. Educational Technology & Society, Vol. 10, No. 3. pp. 275–288.

Brezinski, D. and Killalea, T., (2002). Guidelines for Evidence Collection and Archiving. RFC 3227. Available at: https://tools.ietf.org/html/rfc3227 [Accessed March 7, 2013].

Brinson, A., Robinson, A. and Rogers, M., (2006). A cyber forensics ontology: Creating a new approach to studying cyber forensics. Digital investigation 3S (2006) S37–S43. Elsevier.

Brusa, G., Caliusco, M.L. and Chiotti, O., (2006). A Process for Building a Domain Ontology: An Experience in Developing a Government Budgetary Ontology. Australasian Ontology Workshop (AOW 2006), Hobart, Australia.

Bruschi, D., Martignoni, L. and Monga, M., (2004). How to Reuse Knowledge about Forensic Investigations. Proceedings of Digital Forensic Research Workshop. Baltimore, MD.

Bulbul, H.I., Yavuzcan H.G., and Ozel M., (2013). Digital forensics: An Analytical Crime Scene Procedure Model (ACSPM). Forensic Science International Vol.233 pp.244-256

Burd, S.D., Jones, D.E. and Seazzu, A.F., (2011). Bridging Differences in Digital Forensics for Law Enforcement and National Security. Proceedings of the 44th Hawaii International Conference on System Science.

Burgess, S., (2013). Computer Forensics - Criminal vs Civil: What's the Difference? Available at: http://www.burgessforensics.com/Civ_Criminal.php [Accessed February 23, 2013].

Caloyannides, M.A., (2002). Computer Forensics and Privacy. Artech House.

Caloyannides, M.A., (2004). Privacy Protection and Computer Forensics. Second Edition, Artech House.

Cameron, S.M.S., (2011). Digital Evidence, FBI Law Enforcement Bulletin, Available at: http://www.fbi.gov/stats-services/publications/law-enforcement-bulletin/august-2011/digital-evidence [Accessed February 16, 2013].

Carlos B. and Eduardo M., (2010). Enhancing the Discovery of Web Services:A Keyword-oriented Multiontology Reconciliation

Carrier BD, Spafford EH. Getting physical with the digital forensic process. International Journal of Digital Evidence 2003;2(2).

Carrier, B. and Spafford, E.H., (2003). Getting Physical with the Investigation Process International Journal of Digital Evidence. Vol. 2, No. 2.

Carrier, B., (2008). File system forensic analysis, Addison-Wesley Professional, USA.

Carrier, B.D and Spafford, E.H, (2004). Defining Event Reconstruction of Digital Crime Scenes. Journal of Forensic Sciences. Vol. 49, No. 6

Carrier, B.D. and Spafford, E.H., (2004). An Event-Based Digital Forensic Investigation Framework. Proceedings of the Digital Forensic Research Workshop (DFRWS), Baltimore, Maryland.

Carrier, B.D., (2006). Digital Investigation and Digital Forensic Basics. Available at: http://www.digital-evidence.org/di_basics.html [Accessed March 28, 2013].

Carrier, B.D., (2006a). Basic Digital Forensic Investigation Concepts. Available at: http://www.digital-evidence.org/di_basics.html [Accessed July 24, 2013].

Carrier, B.D., (2006b). A Hypothesis-based Approach to Digital Forensic Investigations. PhD Thesis.

Casey, E. and Stellatos, G.J., (2008). The Impact of Full Disk Encryption on Digital Forensics, ACM SIGOPS Operating Systems Review, Vol. 42, No. 3, pp. 93–98.

Casey, E., (2004a). Digital Evidence and Computer Crime. Forensic Science, Computers and the Internet, 2nd Edition, Elsevier.

Casey, E., (2004b). Network Traffic as a Source of Evidence: Tool strengths, weaknesses, and future needs. Digital Investigation, 1(1): 28–43.

Casey, E., (2009). Handbook of Digital Forensics an Investigation, Academic Press. p. 567.

Castañeda, V., Ballejos, L., Caliusco, M.L. and Galli, M.R., (2010). The Use of Ontologies in Requirements Engineering. Global Journal of Researches in Engineering, Vol. 10, No. 6 (Ver 1.0) GJRE Classification (FOR) 091599. pp. 2–8.

CDAC, (2012). Cyber Forensics - Device Forensics. Available at: http://www.cyberforensics.in/(A(cos8NMWQywEkAAAAODMwODM4YWMtNWFmZC00ZWNhLThkNDEtNTlhMWM3MGE5MzA5hkCziwldj9ts_CCtkjYQI68akds1))/Research/DeviceForensics.aspx?AspxAutoDetectCookieSupport=1 [Accessed March 22, 2012].

Cengage, (2012). Complex Vector Spaces and Inner Products. Available at: http://college.cengage.com/mathematics/larson/elementary_linear/4e/shared/downloads/c08s4.pdf [Accessed May 11, 2012].

CERT, (2009). Vulnerability Statistics (Historical). Available at: http://www.cert.org/stats/ [Accessed July 22, 2013].

Chaikin, D., (2006). Network Investigations of Cyber-attacks: The Limits of Digital Evidence, Crime, Law and Social Change, Vol. 46, pp. 239–256.

Chungoora, N.and Young, R.I.M.1 (2011). Semantic Reconciliation across Design and Manufacture Knowledge Models: a Logic-Based Approach. applied Ontology, 6(4), pp. 295-315

Ciardhuáin, S.O., (2004). An Extended Model of Cybercrime Investigations. International Journal of Digital Evidence, Vol. 3, No. 1.

Cichonski, P., Millar, T., Grance, T. and Scarfone, K., (2012). Computer Security Incident Handling Guide, Revision 2 (Draft), NIST Special Publication 800-61.

Cilibrasi, R.L. and Vitànyi, P.M.B. (2007). The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, March 2007, pp. 370–383.

Cobb, M., (2013). Digital forensic investigation procedure: Form a computer forensics policy. Available at: http://www.computerweekly.com/tip/Digital-forensic-investigation-procedure-Form-a-computer-forensics-policy [Accessed February 18, 2013].

Cohen, F., (2011). Digital Forensic Evidence Examination. 3rd Edition. Published by Fred Cohen & Associates. ISBN # 1-878109-46-4.

Cohen, M. and Schatz, B., (2010). Hash-based disk imaging using AFF4, Digital Investigation 7, S121 -S128.

Colomb, R.M., (1997). Impact of Semantic Heterogeneity on Federating Databases. The Computer Journal, Vol. 40, No. 5, pp. 235–244.

Conserve, (2010). Digital Storage Media, National Service Park, Number 22/5.

Ćosić, J. and Ćosić, Z., (2012). The Necessity of Developing a Digital Evidence Ontology. Proceedings of the Central European Conference on Information and Intelligent Systems, September 19-21, pp. 325–330.

Craiger, P., Swauger, J., Marberry, C. and Hendricks, C., (2006). Validation of Digital Forensics Tools. Digital Crime and Forensic Science in Cyberspace, edited by P. Kanellis, E. Kiountouzis, N. Kolokotronis, and D. Martakos, Idea Group.

Cross, M. and Shinder, D.L., (2008). Scene of the Cybercrime, Burlington, MA: Syngress 2nd edition. Available at: http://books.google.co.za/books?id=fJVcgl8IJs4C&printsec =frontcover#v=onepage&q&f=false [Accessed February 18, 2013].

Crouch, J.-Ed., (2010). An Introduction to Computer Forensics. Available at: http://www.nsci-va.org/WhitePapers/2010-12-16-Computer%20Forensics-Crouch-final.pdf [Accessed March 5, 2012].

CTDP, (2013). Incident Response Plan. Available at: http://www.comptechdoc.org/independent/security/policies/incident-response-plan.html [Accessed February 23, 2013].

Cummings, R., (2008). Computer Forensics - Detecting, Analyzing, and Reporting on Evidentiary Artifacts Found in Computer Physical Memory. Available at: http://www.evidencemagazine.com/index.php?option=com_content&task=view&id=116&Itemid=49 [Accessed July 23, 2013].

Czarnecki, C., (2011). Cloud Service Models: Comparing SaaS, PaaS and IaaS, Perspectives on Cloud Computing & Training from Learning Tree International. Available at: http://cloud-computing.learningtree.com/2011/11/09/cloud-service-models-comparing-saas-paas-and-iaas/ [Accessed February 13, 2013].

Danielsson, J. and Ingvar, T., (2004). The Need for a Structured Approach to Digital Forensic Readiness - Digital Forensic Readiness and E-Commerce, IADIS International Conference on e-Commerce, pp. 417–421.

Danielsson, P-E., (1980). Euclidean Distance Mapping, Computer Graphics and Image Processing, 14, 227–248.

Danushka, B., Yutaka, M. and Mitsuru, I., (2007). Measuring Semantic Similarity between Words Using Web Search Engines. WWW 2007 / Track: Semantic Web, Session: similarity and extraction.

Danushka, B., Yutaka, M. and Mitsuru, I., (2011). A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. IEEE Transactions on Knowledge and Data Engineering, Vol. 23, No. 7.

David, C.H. and Richard P.M., (2007). A Small-Scale Digital Device Forensics Ontology. Small-Scale Digital Device Forensics Journal, Vol. 1, No. 1. pp. 1–7.

Davies, J., Fensel, D. and Van Harmelen, F., (eds.), (2004). Towards the Semantic Web: Ontology-driven Knowledge Management. West Sussex, England, John Wiley & Sons.

Desai, A.M., Fitzgerald, D. and Hoanca, B., (2009). Offering a Digital Forensics Course in Anchorage, Alaska. Information Systems Education Journal, Vol. 7, No. 35. Available at: http://isedj.org/7/35/. ISSN: 1545-679X. (A preliminary version appears in The Proceedings of ISECON 2006: §5114. ISSN: 1542-7382.)

Dietz, J.L.G. and Delft, T.U., (2006). Enterprise Engineering: Enterprise Ontology. Available at: http://www.siks.nl/map_IO_Archi_2006/J.Dietz.pdf [Accessed September 25, 2013].

Ding, (2006). Semantic Annotation for the Semantic Web. Available at: http://www.deg.byu.edu/ding/research/SemanticAnnotation.html [Accessed July 19, 2014].

DOJ, (2013). Volatility of digital evidence. Available at: http://www.policeone.com/police-products/investigation/tips/1655664-Volatility-of-digital-evidence/ [Accessed February 18, 2013].

Dolan-Gavitt, B., Payne, B. and Lee, W., (2011). Leveraging Forensic Tools for Virtual Machine Introspection. Technical Report. Georgia Institute of Technology. Available at: http://amnesia.gtisc.gatech.edu/~moyix/vmi_forensics.pdf [Accessed August 31, 2013].

Duineveld, A.J., Stoter, R., Weiden, M.R., Kenepa, B. and Benjamins, V.R., (2000). WonderTools? A comparative study of ontological engineering tools. International Journal of Human-Computer Studies, 52(6):1111-1133.

Elancheran, A., (2013). Computer Forensics. Available at: http://uwcisa.uwaterloo.ca/Biblio2/Topic/ACC626%20Computer%20Forensics%20A%20Elancheran.pdf [Accessed February 18, 2013].

Erasani, S., (2010). Implementation of Anti-Forensic Mechanisms and Testing with Forensic Methods, Master of Science Dissertation, Texas A&M University-Corpus Christi.

Erick, N., (2012). Semantically Enhanced Composition Writing With Learners of English as a Second Language (ESL). International Journal of Humanities and Social Science. Vol. 2 No. 18

Eriksson, H., Shahar, Y., Tu, S.W., Puerta, A.R. and Musen, M.A., (1995). Task modeling with reusable problem-solving methods. Artificial Intelligence, 79, pp. 293–326.

Eroraha, I., (2010). Real-World Computer Forensics Challenges Facing Cyber Investigators, Computer Forensics Show 2010 Conference.

European Commission, (2010). Editors: Jeffery, K. and Neidecker-Lutz, B., The future of cloud computing: Opportunities for European cloud computing beyond 2010. Expert Group Report.

Falbo, R.A., Menezes, C.S. and Rocha, A.R., (1998). A Systematic Approach for Building Ontologies. Proceedings of the 6th Ibero-American Conference on AI: Progress in Artificial Intelligence, pp. 349–360.

Farquhar, A., Fikes, R. and Rice, J., (1997). Tools for Assembling Modular Ontologies in Ontolingua. Knowledge Systems, AI Laboratory.

Farshad, H. and Andreas, G. (2001). Resolving Semantic Heterogeneity in Schema Integration: An Ontology-based Approach, University of Zurich. International Conference on Formal Ontology in Information Systems (FOIS), Ogunquit, Maine.

Fei, B.K.L., (2007). Data Visualisation in Digital Forensics. University of Pretoria (MSc Dissertation).

Ferguson, R.I., (2013). Challenges in Digital Forensic Research. Available at: http://scone.cs.st-andrews.ac.uk/cybersecurity/slides/Ferguson-DigitalForensicsResearchChallenges.pdf [Accessed June 20, 2013].

Ferndndez, M., Gómez-Pérez, A. and Juristo, N., (1997). Methontology: From Ontological Art Towards Ontological Engineering. Symposium on Ontological Engineering of AAAI, Stanford (California).

Forman, G., Eshghi, K. and Chiocchetti, S., (2005). Finding similar files in large document repositories. Proceedings of the eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining, pp. 394–400, ACM, New York, NY, ISBN 1-59593-135-X.

Francia, G.A., (2006). Digital Forensics Laboratory Projects. Consortium for Computing Sciences in Colleges, Mid-South Conference.

Frederic, L., Bertrand, C., Ayoub, M. and Eric, D., (2009). Image and Video Fingerprinting: Forensic Applications. Proceedings of the SPIE Conference on Media Forensics and Security. Available at: http://dx.doi.org/10.1117/12.806580.

Freiling, F.C. and Schwittay, B. (2007). A Common Process Model for Incident Response and Computer Forensics. Proceedings of Conference on IT Incident Management and IT Forensics. Germany.

Frequencyrising, (2012). Bio Frequency Generator. Available at: http://www.frequencyrising.com/frequency-generator.html [Accessed April 17, 2012].

Gaevic, D., Djuric, D., and Devedic, V., (2009). Model Driven Engineering and Ontology Development. Second Edition. Springer: Dordrecht, Heidelberg.

Gallegos, F., (2005). Computer Forensics: An Overview. Information Systems Audit and Control Association (ISCA), Vol. 6. Available at: http://www.isaca.org/Journal/Past-Issues/2005/Volume-6/Documents/jpdf0506-Computer-Forensics-An.pdf [Accessed February 18, 2013].

Gardner, S.P., (2005). Ontologies and semantic data integration. DDT, Vol. 10, No. 14, pp. 1001–1007.

Garfinkel, S., (2007). Anti-Forensics: Techniques, Detection and Countermeasures, 2nd International Conference on i-Warfare and Security, pp. 77–84.

GAS, (2013). Infrastructure as a Service (IaaS). Available at: http://www.gsa.gov/portal/content/112063 [Accessed March 20, 2013].

Gerald G.D.,(2009).Explaining Reading, Second Edition: A Resource for Teaching Concepts, Skills, and Strategies. Copyright 2009 by Guilford Press. Printed by Guilford Press.

Gladyshev, P., (2004). Concepts of digital forensics (Ph.D. dissertation). Available at: http://www.gladyshev.info/publications/thesis/chapter3.pdf [Accessed July 22, 2013].

Gloe, T, Kirchner, M., Winkler, A. and Böhme, R., (2007). Can We Trust Digital Image Forensics? Proceedings of the 15th International Conference on Multimedia, pp. 78–86. doi>10.1145/1291233.1291252.

Gokhale, P., Deokattey, S. and Bhanumurthy, K., (2011). Ontology Development Methods. DESIDOC Journal of Library & Information Technology, Vol. 31, No. 2, pp. 77-83.

Gómez-Pérez, A., Fernandez-Lopez, M. and Corcho, O., (2004). Ontological Engineering with Examples from the Areas of Knowledge Management, e-Commerce and the Semantic Web. First Edition published by Springer-Verlag, London. eBook ISBN: 978-1-85233-840-4.

Google, B., (2012). Available at: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html [Accessed April 17, 2012].

Grüber, T., (1993). A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 5(2): 199–220.

Gruber, Thomas R. (June 1993). "A translation approach to portable ontology specifications" (PDF). Knowledge Acquisition 5 (2): 199–220. doi:10.1006/knac.1993.1008.

Grüninger, M. and Fox, M.S., (1995). Methodology for the Design and Evaluation of Ontologies. International Joint Conference on Artificial Inteligence IJCAI-95, Workshop on Basic Ontological Issues in Knowledge Sharing.

Guarino, N., (1998). Formal Ontology and Information Systems. Proceedings of the Formal Ontologies in Information Systems, N. Guarino (Ed.), IOS Press, pp. 3–15.

Guidance Software, (2012). EnCase Forensic – Computer Forensic Data Collection for Digital Evidence Examiners. Available at: http://www.guidancesoftware.com/encase-forensic.htm [Accessed August 29, 2012].

Hadzic, M., Wongthongtham, P., Dillon, T. and Chang, E., (2009). Ontology-Based Multi-Agent Systems. Studies in Computational Intelligence, p. 169.

Hanks, K.S., Knight, J.C. and Holloway, C.M., (2002). The Role of Natural Language in Accident Investigation and Reporting Guidelines. Proceedings of the 2002 Workshop on the Investigation and Reporting of Incidents and Accidents.

Harley K., (2003). Digital Evidence. Available at: http://infohost.nmt.edu/~sfs/Students/HarleyKozushko/Papers/DigitalEvidencePaper.pdf [Accessed August 29, 2015].

Harvey, R., (2012). Preserving Digital Materials - Google Books. Available at: http://books.google.co.za/books?id=Z_8glIHqKgQC&pg=PA128&lpg=PA128&dq=Limited+lifespan+of+digital+media&source=bl&ots=Qf3rNzycwR&sig=PtQPJhmT6dlifT-

dPDGDAzfYCmI&hl=en&sa=X&ei=Dz8iUebCMcmwhAe4hYDIBQ&ved=0CEoQ6AEwBQ#v=onepage&q=Limited%20lifespan%20of%20digital%20media&f=false [Accessed February 18, 2013].

Helge, K., (1989). An Introduction to the Historiography of Science. Cambridge University Press. p. 121. ISBN 0-521-38921-6.

Hjelmvik, E., (2012). Passive Network Security Analysis with Network Miner | ForensicFocus.com. Available at: http://www.forensicfocus.com/passive-network-security-analysis-networkminer [Accessed March 27, 2012].

Homeland Security, (2009). Recommended Practice: Developing an Industrial Control Systems Cybersecurity Incident Response Capability, Control Systems Security program. National Cyber Security Division. US Department of Homeland Security, Available at: http://ics-cert.us-cert.gov/practices/documents/final-RP_ics_cybersecurity_incident_response_100609.pdf [Accessed February 25, 2013].

Hoss, A.M. and Carver, D.L., (2009). Weaving Ontologies to Support Digital Forensic Analysis, ISI 2009, Richardson, Texas.

Hotmath, (2012). Absolute Value Functions. Available at: http://hotmath.com/hotmath_help/topics/absolute-value-functions.html [Accessed May 11, 2012].

Houts, A.C. and Baldwin, S., (2004). Constructs, operational definition, and operational analysis. Applied & Preventive Psychology, Vol. 11, pp. 45–46.

Ieong, R.S.C., (2006). FORZA-Digital Forensics investigation framework that incorporates legal issues. Digital Investigation: The International Journal of Digital Forensics & Incident Response, Vol. 3, pp. 29–36.

IntAlgebra, (2012). Absolute Value Functions. Department of Mathematics, College of the Redwoods. Available at: http://msenux.redwoods.edu/IntAlgText/chapter4/chapter4.pdf [Accessed May 11, 2012].

Iordan, V. and Cicortas, A. (2008). Ontologies used for Competence Management. Acta Polytechnica Hungarica, Vol. 5, No. 2. pp. 133–144.

ISIS, (2012). Multimedia Forensics. Available at: http://isis.poly.edu/projects/forensics. [Accessed August 03, 2012].

ISO/IEC 27043, (2015). Information technology - Security techniques - Digital evidence investigation principles and processes. Available at: http://www.iso27001security.com/html/27043.html [Accessed November 25, 2015].

Ithaca College Library, (2013). Primary and secondary sources. Available at: http://www.ithacalibrary.com/sp/subjects/primary [Accessed April 1, 2013].

Jagminder, K.C., Mandeep, S.S. and Sukhveer, S., (2013). Hash-based Four Level Image Cryptography. International Journal on Recent and Innovation Trends in Computing and Communication, Vol. 1, No. 5, pp. 429–432.

Jee, H., Lee, J. and Hong, D., (2008). High Speed Search for Large-Scale Digital Forensic Investigation. Proceedings of the 1st International Conference on Forensic Applications and Techniques in Telecommunications, Information, and Multimedia and Workshop. Article No. 31. Adelaide, Australia.

Jessica T., (2015). Electronic Evidence. Available at: http://www.legalmatch.com/law-library/article/electronic-evidence.html [Accessed August 29, 2015].

Johnson, C., (2000). Forensic software engineering. Proceedings of the 19th International Conference SAFECOMP 2000, pp. 420–430.

Johnson, C., (2002). Forensic Software Engineering: Are Software Failures Symptomatic of Systemic Problems? Safety Science, Vol. 40, pp. 835–847.

Jones, R., (2005). Internet Forensics, Using Digital Evidence to Solve Computer Crime. O'Reilly Media, October 2005.

Jones, S. T., Arpaci-Dusseau, A. C. and Arpaci-Dusseau R. H. (2006). Antfarm: tracking processes in a virtual machine environment. In Proceedings of the USENIX Annual Technical Conference, 2006.

Kajan, E., (2013). Electronic Business Interoperability: Concepts, Opportunities and Challenges - Google Books. Available at: http://books.google.co.za/books?id=fNh2Frjj7oUC&pg=PA287&lpg=PA287&dq=Even+if+two+ontologies+use+the+same+name+for+a+concept,+the+associated+properties+and+the+relationships+with+other+concepts+are+most+likely+to+be+different&source=bl&ots=NQ74gKNzP-&sig=yxThC1hAO27mIwqhAkoJUlPAOUI&hl=en&sa=X&ei=gNUwUeiuMI-LhQfdloDYBQ&ved=0CDYQ6AEwAg#v=onepage&q=Even%20if%20two%20ontologies%20use%20the%20same%20name%20for%20a%20concept%2C%20the%20

associated%20properties%20and%20the%20relationships%20with%20other%20concepts%20are%20most%20likely%20to%20be%20different&f=false [Accessed March 1, 2013].

Karie, N.M. (2014). personal interviews and Talking with People as a way to gather in-depth and comprehensive information about the different digital forensics disciplines and sub-disciplines.(Information Privately recorded on paper by the research during the time of the study – unpublished)

Karie, N.M. and Venter, H.S., (2012). Measuring Semantic Similarity between Digital Forensics Terminologies Using Web Search Engines. Proceedings of the 12th Annual Information Security for South Africa Conference (ISSA-2012). Johannesburg, South Africa. Published online by IEEE Xplore®.

Karie, N.M. and Venter, H.S., (2013c). Significance of Semantic Reconciliation in Digital Forensics. Proceedings of the International Conference on Digital Forensics, Security and Law (ADFSL-2013). Richmond, Virginia.

Karie, N.M. and Venter, H.S., (2015). Taxonomy of Challenges for Digital Forensics.Journal of Forensic Sciences. Vol.60 No.4. pp.885-93. doi: 10.1111/1556-4029.12809

Kent, K., Chevalier, S., Grance, T. and Dang, H., (2006). Guide to Integrating Forensic Techniques into Incident Response, NIST Special Publication 800-86. Gaithersburg - National Institute of Standards and Technology.

Kessler, G.C., (2005). The Role of Computer Forensics in Law Enforcement. Available at: http://www.garykessler.net/library/role_of_computer_forensics.html [Accessed July 23, 2013].

Khan, A., Uffe, K.W. and Nasrullah, M., (2010). Digital Forensics and Crime Investigation: Legal Issues in Prosecution at National Level. Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering, pp. 133–140.

Kilgarriff, A., (1997). Using Word Frequency Lists to Measure Corpus Homogeneity and Similarity between Corpora. Proceedings of the AISB Workshop, Falmer.

King, G.L., (2006). Forensics Plan Guide – Forensic Investigation Plan Cookbook. SANS Institute, Computer Forensics and Incidence Response.

Köhn, M., Eloff, J.H.P. and Olivier M.S., (2006). Framework for a Digital Forensic Investigation, in H.S. Venter, J.H.P. Eloff, L. Labuschagne and M.M. Eloff (Eds), Proceedings of the ISSA 2006 from Insight to Foresight Conference, Sandton, South Africa (Published electronically).

Kohn, M.D., Eloff, M.M. and Eloff, J.H.P., (2013), Integrated Digital Forensic Process Model, Computer & Security, 38(2013) pp. 103-115.

Kortsarts, Y. and Harver, W., (2007), Introduction to Computer Forensics for Non-Majors. Proceedings of the ISECON, 2007. pp. 1–7.

Kottman, C., (1999). Semantics and Information Communities, the OpenGIS Abstract Specification Topic 14. Version. 4. OpenGIS Consortium, OpenGIS™ Project Document Number 99-114.doc.

Kowalski, M., (2002). Cyber-Crime: Issues, Data Sources, and Feasibility of Collecting Police-Reported Statistics. Canadian Centre for Justice Statistics, Catalogue no. 85-558-XIE, ISBN 0-660-33200-8.

Kraetzer, C., Oermann, A., Dittmann, J. and Lang, A., (2007). Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification. Proceedings of the 9th workshop on Multimedia & Security, pp. 63–74. Dallas, Texas.

Kruse, W. and Heiser, J., (2002). Computer Forensics: Incident Response Essentials. Addison Wesley, Indianapolis.

Kuhanandha M. and Michael N.H., (1999). Ontology Tools for Semantic Reconciliation in Distributed Heterogeneous Information Environments. Intelligent Automation and Soft Computing.

Labský, M., Svátek, V. & Šváb, O., (2004). Types and Roles of Ontologies in Web Information Extraction. ECML/PKDD04 Workshop on Knowledge Discovery and Ontologies, Pisa.

Lalla, H. and Flowerday, S.V., (2010). Towards a Standardised Digital Forensic Process: E-mail Forensics. Proceedings of the Information Security South Africa Conference. Sandton, South Africa.

Larose, D.T., (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, NJ.

Larson, J.A., Navathe, S.B. and Elmasri, R., (1989). A Theory of Attribute Equivalence in Databases with Application to Schema Integration. IEEE Transactions On Software Engineering, Vol. SE-15, No. 4.

Lauren, M., (2011). Info-security - Cybercrime Knows No Borders. Available at: http://www.infosecurity-magazine.com/view/18074/cybercrime-knows-no-borders-/ [Accessed February 16, 2013].

Lee, H.C., Palmbach, T.M. and Miller, M.T., (2001). Henry Lee's Crime Scene Handbook. New York: Academic Press.

Lee, J., Chae, H., Kim, K. and Kim, C.-Han., (2006). An ontology architecture for integration of ontologies. Proceedings of the First Asian Semantic Web Conference, Beijing, China, edited by R. Mizoguchi, Z. Shi and F. Giunchiglia. pp. 205–211.

Lee, M.L. and Ling, T.W., (1995), Resolving Structural Conflicts in the Integration of Entity-Relationship Schemas. Object-Oriented and Entity-Relationship Modeling, Vol. 1021, pp. 424–433.

Leigland, R. and Krings, A.W., (2004). A Formalization of Digital Forensics, International Journal of Digital Evidence, Vol. 3, No. 2, pp. 1–32.

Lenat, D.B., (1995). CYC: A large-scale investment in knowledge infrastructure, Communications of the ACM, November 1995, Vol. 38, No. 11.

Leslie, W., Will, V. and Edgar, A.W., (2011). Meeting the Challenges of Cloud Computing. Available at: http://www.accenture.com/us-en/outlook/Pages/outlook-online-2011-challenges-cloud-computing.aspx [Accessed June 20, 2013].

Leung, N.K.Y., Lau, S.K. and Tsang, N., (2012). An Ontology Development Methodology to Integrate Existing Ontologies in an Ontology Development Process. Communications of the ICISA, An International Journal.

Li, W. and Clifton, C. (1994). Semantic Integration in Heterogeneous Databases Using Neural Networks.

Liaison, (2015). Enterprise Semantic Integration. http://www.liaison.com/solutions/data-integration/semantic-integration/ [Accessed August 08, 2015].

Libby, D.A., (2011). Distributed Computer Forensics: Challenges and Possible Solutions, Available at: http://selil.com/archives/2668 [Accessed February 16, 2013].

Lim, N. and Khoo, A., (2009). Forensics of Computers and Handheld Devices: Identical or Fraternal Twins? Communications of the ACM, Vol. 52, No. 6, pp. 132–135.

Lin, Y., Strasunskas, D., Hakkarainen, S., Krogstie, J. and Solvberg, A., (2006). Semantic Annotation Framework to Manage Semantic Heterogeneity of Process Models. Proceedings of the 18th International Conference on Advanced Information Systems Engineering, pp. 433–446.

Lin, Y., Strasunskas, D., Hakkarainen, S., Krogstie, J., and Solvberg, A., (2006), "Sematic Annotation Framework to Manage Semantic Heterogeneity of Process Models", *Proceedings of the 18th international conference on Advanced Information Systems Engineering*, pp. 433-446

Liu, V. and Brown, F., (2006). Bleeding-Edge Anti-Forensics, Infosec World Conference & Expo, MIS Training Institute.

Lizyflorance, C., Lalitha, K.R. and John, S.K., (2012). A Randomized Selective Encryption Using Hashing Technique for Securing Video Streams. International Journal on Computer Science and Engineering (IJCSE), Vol. 4, No. 11, pp. 1843–1847.

Lowman, S. (2010). The Effect of File and Disk Encryption on Computer Forensics. Available at: http://lowmanio.co.uk/share/The%20Effect%20of%20File%20and%20Disk%20Encryption%20on%20Computer%20Forensics.pdf [Accessed February 21, 2013].

Luhn, H.P.(1958), The automatic creation of literature abstracts. IBM Journal, pages 159–165.

MacDonell, S.G., Gray, A.R., MacLennan, G. and Sallis, P., (1999). Software Forensics for Discriminating between Program Authors using Case-Based Reasoning, Feed-Forward Neural Networks and Multiple Discriminant Analysis. Proceedings of the 6th International Conference on Neural Information Processing, (ICONIP '99). Vol. 1, pp. 66–71.

Madsen, B.N. and Thomsen, H.E, (2009). CAOS – A Tool for the Construction of Terminological Ontologies. 17th Nordic Conference of Computational Linguistics 279 (Vol. 283).

Manasa, C., Raman, V. and Ananda R.S.P., (2012). Measuring Semantic Similarity between Words Using Page Counts and Snippets. International Journal of Computer Science & Communication Networks, ISSN:2249-5789. Vol. 2(4), pp. 553–-558.

Mandia, K., Prosise, C. and Pepe, M., (2003). Incident Response & Computer Forensics. (Second Ed.), McGraw-Hill/Osborne, Emeryville.

Marciniak(ed), J.J. (2001). Process Models in Software Engineering. Encyclopedia of Software Engineering, 2nd Edition, John Wiley and Sons, Inc, New York, December 2001.

Mark, R., Clint, C. and Gregg, G., (2002). An Examination of Digital Forensic Models. International Journal of Digital Evidence, Vol. 1, No. 3.

Martins, A.F., and de Almeida, F.R., (2008). Models for Representing Task Ontologies. Proceedings of the Workshop on Ontologies and their Applications, WONTO.

McCracken, (2012). Vectors in Euclidean Spaces. Available at: http://scottmccracken.weebly.com/uploads/9/0/6/6/9066859/vectors-print_version.pdf [Accessed May 10, 2012].

Mell, P. and Grance, T., (2011). The NIST Definition of Cloud Computing. Recommendations of the National Institute of Standards and Technology.

Mercuri, R., (2009). Criminal Defense Challenges in Computer Forensics. Proceedings of the Digital Forensics and Cyber Crime Conference, ICDF2C 2009, Albany, NY.

Michelle, D., (2013). Mobile Devices Get Means for Tamper-Evident Forensic Auditing, Available at: http://www.businesscomputingworld.co.uk/mobile-devices-get-means-for-tamper-evident-forensic-auditing/ [Accessed June 21, 2013].

Miller, G.A. and Charles, W.G., (1998). Contextual correlates of semantic similarity. Language and Cognitive Processes. Volume 6, Issue 1.

Miller, G.A., (1995). WordNet, A Lexical Database for the English Language, Cognitive Science Lab, Princeton University.

Miller, R.J., (1998). Using schematically heterogeneous structures. Proceedings of the 1998 ACM SIGMOD International Conference on Management of data, pp. 189–200.

Mohay, G., (2005). Technical Challenges and Directions for Digital Forensics. Proceedings of the First International Workshop on Systematic Approaches to Digital Forensic Engineering, pp. 155–161.

Moore, T., Manes, G. and Shenoi, S., (2005). Using Signaling Information in Telecom Network Forensics. IFIP International Federation for Information Processing, Vol. 194.

Muhammad, G. and Alghathbar, K., (2011). Environment Recognition for Digital Audio Forensics Using Mpeg–7 and mel Cepstral Features. Journal of Electrical Engineering, Vol. 62, No. 4, pp. 199–205.

Munassar,N.M.A and Govardhan, A(2010). A Comparison Between Five Models Of Software Engineering.International Journal of Computer Science Issues (IJCSI), Vol. 7, No. 5. pp 94-101

Nagyp´al, G. (2007). Ontology development: Methodologies for ontology engineering. In Studer, R., Grimm,S., and Abecker, A., editors, Semantic Web Services:Concepts, Technologies and Applications, chapter 4,pages 107–134. Springer, Berlin.

Naiman, C.E. and Ouksel, A.M., (1995). A Classification of Semantic Conflicts in Heterogeneous Database Systems. Journal of Organizational Computing, 5(2), 167–193.

NATA, (2012). Proficiency Testing Policy in the Field of Forensic Science. Available at: http://www.nata.asn.au/phocadownload/publications/Technical_publications/Policy_Tech_circulars/technical-circular-15.pdf [Accessed March 12, 2012].

Navigli, R. and Velardi, P., (2004). Association for Computational Linguistics, Computational Linguistics, Vol. 30, No. 2.

NeOn, (2013). The NeOn Toolkit. Available at: http://neon-toolkit.org/wiki/Main_Page [Accessed September 26, 2013].

Nigel, F. and Tim M., (2000). Electronic Evidence in Civil Litigation. available at: http://www.internationallawoffice.com/newsletters/detail.aspx?g=ca20791e-ca6d-4fed-a661-9557b5a98672 [Accessed August 29, 2015].

NIJ (2015).Digital Evidence and Forensics. Available at: http://www.nij.gov/topics/forensics/evidence/digital/Pages/welcome.aspx#note1 [Accessed August 29, 2015].

Nirkhi, S.M., Dharaskar, R.V. and Thakre, V.M., (2012). Data Mining - A Prospective Approach for Digital Forensics, International Journal of Data Mining & Knowledge Management Process (IJDKP), Vol. 2, No. 6.

Noblett, M.G., Pollitt, M.M. and Presley, L.A., (2000). FBI - Recovering and Examining Computer Forensic Evidence. Forensic Science Communications, October 2000, Vol. 2, No. 4. Available at: http://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/oct2000/computer.htm [Accessed March 7, 2013].

Noy, N.F. and McGuinness, D.L., (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880.

Noy, N.F., (2004). Semantic Integration: A Survey of Ontology-based Approaches. SIGMOD Record, Vol. 33, No. 4, pp. 65–70.

NSIT, (2015). Law Enforcement Standards Office. Available at: http://www.nist.gov/oles/forensics/digital_evidence.cfm [Accessed August 29, 2015].

Obialero, R., (2003). Forensic Analysis of a Compromised Intranet Server. SANS Institute InfoSec Reading Room.

Öhgren, A., (2009). Towards an Ontology Development Methodology for Small and Medium-sized Enterprises. Dissertation (Licentiate of Engineering degree), Linköping University, Department of Computer and Information Science.

Olivier, M.S., (2009). On Metadata Context in Database Forensics. Digital Investigation: The International Journal of Digital Forensics & Incident Response, Vol. 5, No. 3-4, pp. 115–123.

OMWG, (2013). DERI Ontology Management Environment. Available at: http://dome.sourceforge.net/ [Accessed September 26, 2013].

Oracle Corporation, (2012). Making Infrastructure-as-a-Service in the Enterprise a Reality. An Oracle White Paper.

Oxford Dictionaries, (2013). Definition of disparity in Oxford Dictionaries (British & World English). Available at: http://oxforddictionaries.com/definition/english/disparity [Accessed April 12, 2013].

Palmer, G. (ed.), (2001). A Road Map for Digital Forensic Research: Report from the First Digital Forensic Workshop. DFRWS Technical Report DTR-T001-01. Available at: http://www.dfrws.org/2001/dfrws-rm-final.pdf [Accessed August 6, 2013].

Parsons, J. and Wand, Y., (2003). Attribute-Based Semantic Reconciliation of Multiple Data Sources. Journal on Data Semantics I, Vol. 2800, pp. 21–47.

PCMag, (2012). Internet forensics definition from PC Magazine Encyclopedia. Available at: http://www.pcmag.com/encyclopedia_term/0,2542,t=Internet+forensics&i= 59910,00.asp [Accessed March 22, 2012].

Pease, A. and Benzmüller, C., (2012). Sigma: An Integrated Development Environment for Formal Ontology. AI Commun. (Special Issue on Intelligent Engineering Techniques for Knowledge Bases).

Pease, A., (2003a). The Sigma Ontology Development Environment. Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems, August 9, Acapulco, Mexico.

Pease, A., (2013b). The Suggested Upper Merged Ontology (SUMO) - Ontology Portal. Available at: http://www.ontologyportal.org/ [Accessed September 26, 2013].

Perumal, S., (2009). Digital Forensic Model Based on Malaysian Investigation Process. IJCSNS International Journal of Computer Science and Network Security, Vol. 9, No. 8.

Peter H.M.(2001). A Short History of Structural Linguistics. Cambridge University Press. p. 118. ISBN 978-0-521-62568-5.

Piasecki, M., (2008). HydroTagger: A Tool for Semantic Mapping of Hydrologic Terms. AAAI Spring Symposium: Semantic Scientific Knowledge Integration, pp. 77–80.

Pierce, M., (2003). Detailed Forensic Procedure for Laptop Computers Forensic Analysis. SANS Institute InfoSec Reading Room.

Pierce, R., (2007). Evaluating Information: Validity, Reliability, Accuracy, Triangulation. pp. 79-99. Available at: http://www.sagepub.com/upm-data/17810_5052_Pierce_ Ch07.pdf [Accessed April 1, 2013].

Pinnacle Studio, (2012). Video editing software - Pinnacle Studio - The #1 selling digital video editing software. Available at: http://www.pinnaclesys.com/PublicSite/us/Products/Consumer+Products/Home+Video/Studio+Family/ [Accessed August 26, 2012].

Pollitt, M., (1995). Computer Forensics: An Approach to Evidence in Cyberspace. Proceedings of the National Information Systems Security Conference, Baltimore, MD, Vol. 2, pp. 487–491.

Porter, M.F.(1980), An algorithm for suffix stripping. Information & Knowledge Management. Vol. 14 Iss: 3, pp.130 - 137

Poulsen, K., (2007). FBI's Magic Lantern Revealed. Available at: http://www.wired.com/threatlevel/2007/07/fbis-magic-lant/ [Accessed September 18, 2013].

Prasanna, A., (2012). Cyber Crimes: Law and Practice. Available at: http://www.img.kerala.gov.in/docs/downloads/cyber%20crimes.pdf [Accessed October 7, 2012].

Prasenjit, M. and Gio, W., (2002). Resolving Terminological Heterogeneity in Ontologies. Proceedings of the ECAI-02 Workshop on Ontologies and Semantic Interoperability. Lyon, France, July.

Prathvi, K. and Ravishankar, K., (2013). Measuring Semantic Similarity between Words using Page-Count and Pattern Clustering Methods. International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075,Vol. 3, No. 2.

Purplemath, (2012). Absolute Value. Available at: http://www.purplemath.com/modules/absolute.htm [Accessed May 11, 2012].

Reddy, K. and Venter, H.S., (2013). The Architecture of a Digital Forensic Readiness Management System. Computers & Security, Vol. 32, pp. 73–89.

Reed, T., (2012). Time vs Technology and the Frailty of Digital Media, Available at: http://filmcourage.com/content/time-vs-technology-and-the-frailty-of-digital-media [Accessed August 15, 2013].

Reilly, D., Wren, C. and Berry, T., (2011). Cloud Computing: Pros and Cons for Computer Forensic Investigations, International Journal of Multimedia and Image Processing (IJMIP), Volume 1, Issue 1, March 2011.

Reith, M., Carr, C. and Gunsch, G. (2002). An Examination of Digital Forensic Models. International Journal of Digital Evidence, Vol. 1, No. 3. pp. 1–12.

Revelytix, (2013). Knoodl. Available at: http://knoodl.com/ui/home.html [Accessed September 26, 2013].

Richard, G.G. and Roussev, V., (2006). Digital Forensics Tools - The Next Generation, Idea Group Inc, pp. 76–91.

Richmond H.T. (2012). What is Semantics? available at: https://web.eecs.umich.edu/~rthomaso/documents/general/what-is-semantics.html [Accessed November 18, 2015].

Rimage, (2011). Digital Evidence Preservation and Distribution: Updating the Analog System for the Digital World. Available at: http://info.rimage.com/rs/qumu/images/RimageWhitePaperDigitalEvidence_FINAL_EN.pdf [Accessed February 25, 2013].

Roberts, J.L. and Suits, C., (2013). Admissibility of digital image data: concerns in the courtroom. Available at: http://libraries.maine.edu/Spatial/gisweb/spatdb/acsm95/ac95071.html [Accessed April 10, 2013].

Rogers, M.K., Goldman, J., Mislan, R., Wedge, T. and Debrota, S., (2006). Computer Forensics Field Triage Process Model. Proceedings of the Conference on Digital Forensics, Security and Law.

Ruan, K., Carthy, J., Kechadi, T. and Crosbie, M., (2011). Cloud forensics: An overview. Centre for Cybercrime Investigation, University College Dublin.

Rubenstein, H. and Goodenough, J.B., (1965). Contextual Correlates of Synonymy. Computational Linguistics. Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Massachusetts.

Ryan, D.J. and Shpantzer, G., (2005). Legal Aspects of Digital Forensics. Washington, Washington, D.C. Available at: http://euro.ecom.cmu.edu/program/law/08-732/Evidence/RyanShpantzer.pdf [Accessed April 1, 2013].

Sacramento, E.R, Vidal1, V.M.P., de Macêdo, J.A.F., Lóscio, B.F., Lopes, F.L.R. and Casanova, M.A., (2010). Towards Automatic Generation of Application Ontologies. Journal of Information and Data Management, Vol. 1, No. 3, pp. 535–550.

Sanchez, P., Milson, R. and Slone, M., (2012). Euclidean distance (version 11). Available at: http://planetmath.org/EuclideanDistance.html [Accessed May 7, 2012].

Schwarz, T.S.J., (2005). Teaching Ethics and Computer Forensics: The Markkula Center for Applied Ethics Approach, Proceedings of the 2nd Annual Conference on Information Security Curriculum Development, pp. 66–71.

Schwerha, J.J., (2008). Why computer forensic professionals shouldn't be required to have private investigator licenses, Digital Investigation: The International Journal of Digital Forensics & Incident Response, Vol. 5, No. 1-2, pp. 71–-72.

Sena, (2012). Network-enabled Devices. Available at: http://www.sena.com/solutions/network_enabling/ [Accessed September 5, 2012].

Shadbolt, N., Berners-Lee, T. and Hall, W., (2006). The Semantic Web Revisited. IEEE Intelligent Systems, Vol. 21, No. 3, pp. 96–101.

Sheetal, A.T. and Sushama, S.N., (2010). Measuring Semantic Similarity between Words Using Web Documents. International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 1, No. 4.

Sheth A.P. and Larse, J., (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. ACM Computing Surveys (CSUR) - Special issue on Heterogeneous Database Surveys. Vol. 22, No. 3, pp. 183–236.

Sheth, A.P. and Gala, S.K., (1989). Attribute Relationships: An Impediment in Automating Schema Integration. Proceedings of theWorkshop on Heterogeneous Database Systems, Chicago.

Sheward, M., (2012). Rock Solid: Will Digital Forensics Crack SSD's? Available at: http://resources.infosecinstitute.com/ssd-forensics/ [Accessed February 18, 2013].

Shotton, D., (2009). Semantic publishing: The coming revolution in scientific journal publishing. Learned Publishing, 22 (2): 85–94.

Shotton, D., Portwin, K., Klyne, G., Miles, A., (2009). Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article. Bourne, Philip E. PLoS Computational Biology, 5 (4).

Sibiya, G., Venter, H.S., Ngobeni, S. and Fogwill, T., (2012). Guidelines for Procedures of a Harmonised Digital Forensic Process in Network Forensics. Proceedings of the Information Security South Africa Conference. Sandton, South Africa.

Siles, R., (2012). Wireless Forensics: Tapping the Air - Part One | Symantec Connect Community. Available at: http://www.symantec.com/connect/articles/wireless-forensics-tapping-air-part-one [Accessed March 26, 2012].

Singh, T., (2012). Cyber Law & Information Technology. Available at: http://delhicourts.nic.in/ejournals/CYBER%20LAW.pdf [Accessed August 29, 2012].

Sleuth Kit and Autopsy, (2012). Open Source Digital Investigation Tools. Available at: http://www.sleuthkit.org/index.php [Accessed August 29, 2012].

Smith vs. Groover, 468 F.Supp. 105 (N.D.Ill.1979), the court noted that Congress traced the inadequacies of the self-regulatory scheme to a number of factors.

Smith, B., Kusnierczyk, W., Schober, D. and Ceusters, W., (2006). Towards a Reference Terminology for Ontology Research and Development in the Biomedical Domain. Proceedings of the 2nd International Workshop on Formal Biomedical Knowledge Representation: \Biomedical Ontology in Action. pp. 57–66.

Sommer, P., (2012). Digital Evidence, Digital Investigations and E-Disclosure: A Guide to Forensic Readiness for Organisations, Security Advisers and Lawyers. Available at: http://www.iaac.org.uk/_media/DigitalInvestigations2012.pdf [Accessed September 13, 2013].

Spoenle, J., (2010). Cloud Computing and Cybercrime Investigations: Territoriality vs. the Power of Disposal? Available at: http://www.coe.int/t/dghl/cooperation/economiccrime/cybercrime/documents/internationalcooperation/2079_Cloud_Computing_power_disposal_31Aug10a.pdf [Accessed February 23, 2013].

SRI, (2013). Open Knowledge Base Connectivity Working Group. Available at: http://www.ai.sri.com/~okbc/ [Accessed September 25, 2013].

Staab, S., Studer, R., Schnurr, H.-P. and Sure, Y., (2001). Knowledge Processes and Ontologies. IEEE Intelligent Systems, Vol. 16, No. 1, pp. 26–34.

Stamm, M.C. and Liu, K.J.R., (2011). Anti-Forensics for Frame Deletion/Addition in Mpeg Video. Proceedings of the ICASSP, 2011, pp. 1876-1879.

StatPac, (2015). Research methods. Available at: https://www.statpac.com/surveys/research-methods.htm. [Accessed August 15, 2015].

Stephenson, P., (2003). A Comprehensive Approach to Digital Incident Investigation. Information Security Technical Report, Vol. 8, No 2, pp. 42-52.

Subramanian, K., (2011a). Public Clouds. A White Paper sponsored by Trend Micro Inc.

Subramanian, K., (2011b). Hybrid Clouds. A White Paper sponsored by Trend Micro Inc.

Sujatha, T., Ramesh, N.G. and Suresh, B.P., (2012). Measuring Semantic Similarity between Words Using Web Pages. International Journal of Soft Computing and Engineering (IJSCE). ISSN: 2231-2307, Vol. 2, No. 3.

Swaminathan, A., Wu, M. and Liu, K.J.R., (2006). Image Tampering Identification Using Blind Deconvolution. Proceedings of the IEEE ICIP 2006.

Swaminathan, A., Wu, M., and Liu, K.J.R., (2008). Digital Image Forensics via Intrinsic Fingerprints. IEEE Transactions on Information Forensics and Security, Vol. 3, No. 1, March 2008.

SWGDE and IOCE (2000). Digital evidence standard and principles. Forensic Science Communications, Vol.2, No.2. available at: https://www.fbi.gov/about-us/lab/forensic-science-communications/fsc/april2000/swgde.htm/ [Accessed August 29, 2015].

Taute, B., Grobler, M. and Nare, S., (2007). Forensic Challenges for Handling Incidents and Crime in Cyberspace, Available at: http://researchspace.csir.co.za/dspace/ bitstream/10204/3756/1/Taute_d1_2009.pdf [Accessed February 18, 2013].

Techopedia, (2013). Community Cloud. Available at: http://www.techopedia.com/ definition/2 6559/community-cloud [Accessed February 8, 2013].

Thinkquest, (2013). Cryptanalysis: Introduction. Available at: http://library.thinkquest.org/27993/ crypto/classic/analysis1.shtml [Accessed April 8, 2013].

Thiyagarajan, D., Shanthi, N. and Navaneethakrishnan, S., (2011). A Web Search Engine-Based Approach to Measure Semantic Similarity between Words.

International Journal of Advanced Engineering Research and Studies, E-ISSN2249–8974.

Tschannen-Moran, M., (2001). Collaboration and the Need for Trust. Journal of Education Administration, Vol. 39, pp. 308–331.

Ubbo, V., Stuckenschmidt, H., Schlieder, C., Wache, H. and Timm, I., (2002). Terminology Integration for the Management of Distributed Information Resources. Künstliche Intelligenz, Vol. 16, pp. 31–34.

Ubuntu, (2013). Private cloud. Available at: http://www.ubuntu.com/cloud/private-cloud [Accessed February 8, 2013].

Ucalgary, (2012). Absolute Value. Available at: http://math.ucalgary.ca/sites/math.ucalgary.ca/files/courses/F07/MATH251/lec5/MATH251-F07-LEC5-Appendix_E.pdf [Accessed May 11, 2012].

Uschold, M. and Gruninger, M., (1996). Ontologies: Principles, Methods and Applications. Knowledge Engineering Review, Vol. 11, No. 2.

Uschold, M. and King, M., (1995). Towards a Methodology for Building Ontologies. Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing, held in conjunction with IJCAI-95.

Uschold, M., (1996). Building Ontologies: Towards a Unified Methodology. Proceedings of Expert Systems '96, the 16th Annual Conference of the British Computer Society Specialist Group on Expert Systems, Cambridge, UK.

Uschold, M., King, M., Moralee, S. and Zorgios, Y., (1998). The Enterprise Ontology. The Knowledge Engineering Review, Vol. 13, No. 1, pp. 31–89.

Vaciago, G., (2012). Cloud Computing and Data Jurisdiction: A New Challenge for Digital Forensics. Proceedings of the Third International Conference on Technical and Legal Aspects of the e-Society, CYBERLAWS 2012.

Valjarevic, A. and Venter, H.S., (2012). Harmonised digital forensic investigation process model. Proceedings of the 12th Annual Information Security for South Africa Conference, (ISSA-2012). Johannesburg, South Africa. Published online by IEEE Xplore®. pp. 1–10.

Van den Bos, J. and van der Storm, T., (2011). Bringing Domain-Specific Languages to Digital Forensics. Procedings of the International Conference on Systems Engineering.

Van Rees, R. and Amor, R., (2003). Clarity in the usage of the terms ontology, taxonomy and classification. Construction Informatics Digital Library. Available at: http://itc.scix.net/ paper w78-2003-432.content [Accessed August 29, 2012].

Vanitha, K., Yasudha, K., Soujanya, K.N., Sri Venkatesh, M., Ravindra,K. and Lakshmi, S.V., (2011). The Development Process of the Semantic Web and Web Ontology. International Journal of Advanced Computer Science and Applications, (IJACSA) Vol. 2, No. 7, pp. 122–125.

Vijay, S., (2012). A Combined Method to Measure the Semantic Similarity between Words. International Journal of Soft Computing and Engineering (IJSCE) ISSN: 2231–2307, Vol. 1, No. ETIC2011.

Walls, R.J., Learned-Miller, E. and Levine. B.N., (2011). Forensic Triage for Mobile Phones with DEC0DE. Proceedings of the USENIX Security Symposium.

Wang, H. and Liu, J.N.K., (2009). Analysis of Semantic Heterogeneity Using a new Ontological Structure Based on Description Logics. Sixth International Conference on Fuzzy Systems and Knowledge Discovery.

Wang, W. and Farid, H., (2007). Exposing Digital Forgeries in Video by Detecting Duplication. Proceedings of the 9th workshop on Multimedia & Security, pp. 35–42. doi>10.1145/1288869.1288876.

Wang, X., Ausdal, S.V. and Zhou, J., (2005). Managing the Life Cycle of Business Semantics. Available at: http://xtensible.net.s60489.gridserver.com/wp-content/uploads/managing-the-life-cycle-of-business-semantics.pdf [Accessed April 25, 2013].

Watkins, K., McWhorte, M., Long, J. and Hill, B., (2009). Teleporter: An analytically and forensically sound duplicate transfer system. MITRE Corporation, Mclean, VA. Digital Investigation 6, S43–S47.

Weippl, E., (2009). Database Forensics. Secure Business Austria. Available at: http://www.nii.ac.jp/issi/pdf/2/4Johannes_Heurix.pdf [Accessed March 26, 2012].

Whale, G., (1990). Software Metrics and Plagiarism Detection. Journal of Systems and Software, Vol. 13, pp. 131–138.

Whitehead, A., (2013). Weakness in Computer Forensics. Available at: http://free-backup.info/weaknesse-in-computer-forensics.html [Accessed February 23, 2013].

Xu, Z. and Lee, Y.C., (2002). Semantic Heterogeneity of Geo Data, Symposium on Geospatial Theory, Processing and Applications, Ottawa.

X-Ways, (2012). WinHex: Hex Editor and Disk Editor, Computer Forensics and Data Recovery Software. Available at: http://www.winhex.com/winhex/ [Accessed August 29, 2012].

Yan, C., (2011). Cybercrime Forensic System in Cloud Computing. Proceedings of the Image Analysis and Signal Processing (IASP) Conference, pp. 612–615.

Yang, J., Li, T., Liu, S., Wang, T., Wang, D. and Liang, G., (2007). Computer Forensics System Based on Artificial Immune Systems. Journal of Universal Computer Science, Vol. 13, No. 9, pp. 1354–1365.

Yates, M., (2010). Practical Investigations of Digital Forensics Tools for Mobile Devices. Proceedings of the Information Security Curriculum Development Conference, pp. 156–162.

Yong, G., (2007). Digital Forensics: Research Challenges and Open Problems. Available at: http://itsecurity.uiowa.edu/securityday/documents/guan.pdf [Accessed June 21, 2013].

Yusoff, Y., Ismail, R. and Hassan, Z., (2011). Common Phases of Computer Forensics Investigation Models. International Journal of Computer Science & Information Technology (IJCSIT), Vol. 3, No. 3.

Zheng et al., (2011). Measuring semantic similarity between words by removing noise and redundancy in web snippets. Concurrency and Computation: Practice and Experience, 23:2496–2510. Published online 22 September 2011, Wiley.

Zoltan S., (2012). Digital Forensics is not just HOW but WHY. Available at: http://articles.forensicfocus.com/2012/07/03/digital-forensics-is-not-just-how-but-why/ [Accessed February 23, 2013].

APPENDIX A: PAPERS PUBLISHED IN INTERNATIONAL CONFERENCE PROCEEDINGS

During the time of this research study, a number of the projects undertaken have been presented at international conferences while others are published in international scientific journals. The papers are briefly explained below.

The first presentation was on measuring semantic similarity between digital forensics terminologies using Web search engines discussed as part of chapter 8 and 9 of this research thesis. The presentation was done in August 2012, at the Information Security South Africa, 12th Annual Conference in Sandton, Johannesburg, South Africa. This paper has also been published online by IEEE Xplore®.

The second was an ontological framework for a cloud forensic environment presented as part of chapter 7 in this thesis. The presentation of this paper was done at the European information security multi-conference (EISMC-2013) in May, 2013, Lisbon, Portugal.

Third, is a presentation done on the significance of semantic disparity reconciliation in digital forensics and is discussed as part of chapter 5 of this research thesis. This paper was presented at the international conference on digital forensics, security and law (ADFSL-2013) in June 2013, Richmond, Virginia USA.

In addition to the conference papers mentioned in this research thesis, several others have been published in scientific journals. These are explained in Appendix B. All the above mentioned conference papers are presented on the pages to follow.

Measuring Semantic Similarity between Digital Forensics Terminologies Using Web Search Engines

Nickson M. Karie
Department of Computer Science,
University of Pretoria,
Private Bag X20, Hatfield 0028,
Pretoria, South Africa.
Email: menza06@hotmail.com

Hein S. Venter
Department of Computer Science,
University of Pretoria,
Private Bag X20, Hatfield 0028,
Pretoria, South Africa.
Email: hventer@cs.up.ac.za

***Abstract*—Semantic similarity between different terminologies is becoming a generic problem that extends across numerous domains, touching applications developed for computational linguistics, artificial intelligence, cognitive science and, in the case of this paper, digital forensics. Despite the usefulness of semantic similarity measures in different domains, accurately measuring semantic similarity between any two terms remains a challenging task. The main difficulty lies in developing a computational method with the ability to generate satisfactory results close to how human beings perceive these terminologies, especially when used in their domain of expertise.**

This paper presents a novel approach of using the Web to measure semantic similarity between two terms x and y in the digital forensics domain. The proposed approach is based on the Euclidean distance, a mathematical concept used to calculate the distance between two points. This paper also shows how computing the absolute value of the difference of the logarithms of the hit count percentages of any given terms *x* and *y* relates to the computed Euclidean distance of *x* and *y*. Percentages are computed from the total number of hit counts reported by any Web search engine for the search terms *x*, *y* and the logical *x* AND *y* together. Finally, these concepts are used to deduce a formula to automatically calculate a semantic similarity measure coined as the Digital Forensic Absolute Semantic Similarity Value of the terms *x* and *y*, denoted as DFASSV(*x*, *y*).

Experiments conducted using the proposed DFASSV method focuses on the digital forensics domain. However, a comparison of the DFASSV approach with previously proposed Web-based semantic similarity measures shows that this approach is well suited for digital forensics domain terminologies. In the authors' opinion however, the DFASSV approach can be applied in other domains as well because it does not require any human-annotated knowledge. DFASSV is a novel approach to semantic similarity measure and constitutes the main contribution of this paper.

Keywords-Semantic similarity; digital forensics; digital forensic domain terminologies; Euclidean distance; absolute value; Web; Web search engines

Introduction

An accurate measurement of semantic similarity between terms is a matter of concern in many different domains. For example, due to the problem of ever-changing technological trends in digital forensics, new terms are constantly introduced into the domain and new meanings assigned to existing terms. Depending on the traditional knowledge-based approach, capturing the meaning of these new terms can be very hard, if not next to impossible. Such knowledge could be useful in, for example, the definition of new digital forensic terms, especially when attempting to standardise terms in the field of digital forensics. The authors are currently involved in the creation of an international standard for the digital forensic investigation process where the need arises to carefully define and reason about specific digital forensic terms.

This paper, therefore, proposes a method to compute the semantic similarity measure between two terms in the digital forensic domain using Web search engines. This method is referred as the Digital Forensic Absolute Semantic Similarity Value (DFASSV) in this paper. For the purpose of this paper and scalability of the semantic similarity measure, the terms that are paired using the proposed DFASSV method are rated on a scale of 0 to ∞ where 0 denotes identical semantic similarity between the two terms and ∞ denotes no semantic similarity. Experiments conducted using the proposed method have delivered impressive results.

The Web is a vast entity where an astronomical amount of information is amassed. It is also the largest semantic electronic database in the world [1]. This "database" is available to all and can be queried using any Web search engine that can return aggregate hit count estimates for a large range of search queries [1]. New information is also added to the Web on a daily basis. To tap into this rich bank of information, Web search engines are the most frequently used tools to query for information related to a

particular term. To the authors' knowledge, there is so far no better or easier way to search for information on the World Wide Web than simply using Web search engines like Google. However, we do not dispute the existence of other techniques that can be used to search and extract information from the Web. For the purpose of this study, however, the Google search engine was used.

As for the remaining part of this paper, section II discusses related research work. In section III some technical background is explained, followed by a discussion of the proposed DFASSV method in section IV. Experimental results are considered in Section V, while conclusions are drawn in section VI and mention is made of future research work.

Related Work

There are several methods for measuring the semantic similarity between terms that have been proposed by other researchers. Some of these methods are based on taxonomy while others are Web-based. Taxonomy-based methods use information theory and hierarchical taxonomy such as the WordNet [4] to measure semantic similarity. Web-based methods, on the other hand, prefer the Web as a live and active corpus to a hierarchical taxonomy [5].

The concept of calculating similarity between two words based on the length of the shortest path connecting the two words in taxonomies is discussed in a paper by Roy Rada et al. [6]. If a word is polysemous (i.e. having more than one sense), then multiple paths may exist between the two words. In such cases only the shortest path between any two senses of the words is considered for calculating similarity. The problem of using this approach is that it assumes that, 'theoretically' all the paths in the taxonomy represent equal distances [7] (i.e. the path distance remains the same in all cases and at all times). In practice however, this assumption might not be true; hence the results of the computed semantic similarity measure may well be incorrect.

In another paper, Ming Li et al. [9], discuss the concept of using Web search engine hits for extracting social network information on the Web. Their approach measures the association between two personal names using the Simpson coefficient [9], [17] and [18] and is calculated based on the number of Web hits for each individual name and its conjunction. This approach however, focuses more on the strength of the relation, while the current paper focuses more on the automatic identification of the underlying semantic similarities.

In 2007, Cilibrasi and Vitànyi [1] introduced the concept of the Normalized Google Distance (NGD), which was based on a 2004 research paper on normalized information distance between two strings (discussed in [9]), and which calculates a distance metric between words using page counts indexed by a Web search engine. The NGD is evaluated in a word classification task (i.e. words are grouped based on their similarities according to the model referred to in [10]). This also means that the words in question usually display the same formal properties, especially their inflections and distribution. The problem with this method is that it uses a value (**M**) that can be defined as the total number of pages on the Web that Google will search when given a query. The value **M** is quoted as 8058044651Web pages [1]. According to an Official Google Blog [11], this number has increased significantly since 1998 when it was only 26 million. By 2000 the Google index had reached the one billion mark. Over the last decade, this number has been changing and, recently, even the Google search engineers stopped calculating it due to the sheer vastness of the Web these days [11]. The Google systems that process links on the Web recorded that 1 trillion (1,000,000,000,000) unique URLs exist on the Web at once. Therefore, it is the authors' opinion that, because of the ever changing nature of the Web, depending on this value to calculate **M** might produce unreliable similarity scores over time.

In their paper, Chen et al. [12] propose a Web-based double checking method to find similar words. They collect snippets for two words x and y from the Web search engine and use these to count the number of occurrences of x in the snippets of y, and y in the snippets of x. The two values are then combined nonlinearly to compute the similarity between x and y. The problem with this method is that it relies heavily on the search engine's ranking algorithm. Although two words may be similar, it is not a guarantee that one will find y in the snippets of x or vice versa [10]. This may also have some effect on the final computed similarity measure.

There are many other proposed methods for finding word similarity using the Web, but none of the cited references in this paper uses the reported Web hit count in the way that is introduced in this paper. Our approach uses the Web and the Web search engines to automatically calculate semantic similarity between two terms, based on the number of hit counts reported for each terms (rather than for each hit).

The hit count of a query is usually an estimated number of Web pages containing the queried term as reported by a Web search engine. The hit count, however, may not necessarily be equal to the term frequency, because the queried term may appear many times on a single Web page. Therefore, an additional hit count is computed where a search term *x* and another search term *y* appear both on the same Web page, indicated as a logical *x* AND *y* search query. The search results of this query therefore, can be considered as the global estimated value of the co-occurrence of the terms *x* and *y* together on the Web [2]. Logical *x* AND *y* is also used in this study to capture the context where both *x* and *y* are used together on the same Web page.

The presentation in this paper is a new approach to using the hit counts to calculate semantic similarity. This observation is also confirmed by the experimental results based on a benchmark data set of words from Miller and Charles (1998) [13]. This data set is a subset of Rubenstein and Goodenoughs' [15] original data set of 65 word pairs. Although the Miller and Charles experiment was carried out 25 years later than that of Rubenstein and Goodenough, the two sets of ratings are highly correlated with a correlation coefficient of 0.97 (on a scale of 0 to 1, where 0 indicates no correlation and 1 indicates complete correlation) [7]. Therefore, the Miller and Charles ratings can be considered as a reliable benchmark data set for evaluating semantic similarity measures.

Technical Background

Much of the theory explained in this paper is based on computing the Euclidean distance between any two points in the Euclidean space, and its relationship with the computed absolute value of the difference of any two real numbers in the number line. Different distance functions result in different distance measures. However, the Euclidean distance used in this paper is considered the most useful because it corresponds to the way objects are measured in the real world [25]. For more information on Euclidean distance and absolute value, please refer to [22] and [23] respectively.

A. The Euclidean distance

The Euclidean distance is defined as the distance between any two points in a plane that one would measure using a ruler and it is given by the Pythagorean Theorem [19], [20], [21] and [22]. If, for example, $x = (x_1, x_2)$ and $y = (y_1, y_2)$ are two given points on the plane, then their Euclidean distance (d) can be defined as [19],

$$\sqrt{(x_1\text{-}x_2)^2 + (y_1\text{-}y_2)^2} \quad (1)$$

Using this formula as distance, the Euclidean space becomes a metric space also called the distance space.

For any given two points *x* and *y* the Euclidean distance between them is the length of the line segment connecting them. In a Cartesian coordinate, for example, if $x = (x_1, x_2,...,x_n)$ and $y = (y_1,y_2,\ldots, y_n)$ are two points in the Euclidean space, then the distance (d) from *x* to *y*, or from *y* to *x* can be defined by equation 2,which can be seen as a generalization of equation 1 [19] and [20].

$$d(x,y) = d(y,x) = \sqrt{(x_1\text{-}y_1)^2 + (x_2\text{-}y_2)^2 + \cdots + (x_n\text{-}y_n)^2} \quad (2)$$

where *n* represents any number denoting a point x_n and y_n in the Cartesian coordinate.

The position of any point in a Euclidean n-space is usually called a Euclidean vector. Therefore, the points *x* and *y* can be referred to as Euclidean vectors. Starting from the origin of the space, their tips indicate the distance between the two points (also called the magnitude or the norm). The Euclidean norm or the length of a vector *x* is the real number denoted as $\left||x|\right|$ [24] and measures the distance of *x* as defined by equation 3 [26]:

$$\left||x|\right| = (x.x)^{1/2} = \sqrt{x.x} \quad (3)$$

The distance, therefore, between *x* and *y* can be computed as [26]:

$$d(x,y) = \left||x-y|\right| \quad (4)$$

The Euclidean norm and distance may as well be expressed in terms of components as shown in equation 5 [24] and [26]:

$$\left||x|\right| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{x.x} \quad (5)$$

If the length of a vector is considered as the distance from its tail to its tip, then it becomes clear that the Euclidean length of a vector is a special case of the Euclidean distance. Therefore, the distance between *x* and *y* is the Euclidean length of the distance vector defined as [24]:

$$\left||x\text{-}y|\right| = \sqrt{(x\text{-}y).(x\text{-}y)} \quad (6)$$

Equation 6 is homogeneous to equations 3, 4 and 5 and can be used to compute the magnitude or the norm of the numerical difference between any two real numbers *x* and *y* in the number line, denoted as $\left||x\text{-}y|\right|$.

It is also clear from equation 6 that the one-dimensional Euclidean distance between x and y can be realized.

One-dimensional Euclidean distance

In the case of one dimension, the distance between two points x and y on the real number line is equivalent to the absolute value of their numerical difference. Thus, if x and y represents two real numbers, then the distance between them can be computed as [27]:

$$\sqrt{(x-y)^2} = |x-y| \qquad (7)$$

In addition, in one-dimension there is usually a single homogeneous, translation-invariant distance function, which is the Euclidean distance and defines the distance between elements of a set. Translation-invariant implies that starting from the origin, at least in one direction, the object is infinite. In higher dimensions, up to n-dimensions, there are other possible distance functions but these are beyond the scope of this paper. We therefore consider only up to the two-dimensional Euclidean distance in this paper.

Two-dimensional Euclidean distance

In the Euclidean plane, if $x = (x_1, x_2)$ and $y = (y_1, y_2)$, then the distance (d) between the two points x and y is given by equation 8 [28], which is homogeneous to equations 1 and 2.

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_1 - y_2)^2} \qquad (8)$$

At this point the discussion of the Euclidean distance presents us with the foundation of establishing its relationship with the absolute value.

B. *The relationship between Euclidean distance and absolute value*

Having gained an understanding of the Euclidean distance, we now establish its relationship with the absolute value of any real number x denoted as $|x|$. The absolute value $|x|$ is the numerical value of x without regard to its sign. For example, the absolute value of $+x$ is x, and the absolute value of $-x$ is also x. This simply means, the absolute value of any real number x may be thought of as its distance from zero (i.e. how far x is from zero on the number line) [29], [30] and [31].

In practice, the absolute value of all real numbers is always positive. See equation 9. The concept of absolute value is closely related to the notion of distance in various mathematical and physical contexts. In this paper, therefore, the relationship between the Euclidean distance and the absolute value is established first. This relationship is then used to generate a formula that automatically calculates a semantic similarity measure (the distance) between any two terms x and y in the digital forensics domain. For any real number x its absolute value denoted by $|x|$ can be defined as shown in equation 9 [32]:

$$|x| = \begin{cases} x, & \text{if } x \geq 0 \\ -x, & \text{if } x < 0 \end{cases} \qquad (9)$$

Based on this definition, the absolute value of x is always either positive or zero, but never negative. In addition, the absolute value of the difference of any two real numbers x and y defines the distance between x and y denoted as $|x - y|$ which is equivalent to the Euclidean distance of x and y. Since in mathematics the square-root of a number x without regard to its sign represents a positive square root, and the absolute value of x is always either positive or zero, but never negative, it follows that [32]:

$$|x| = \sqrt{x^2} \qquad (10)$$

Equation 10 is homogeneous to equation 7 and is sometimes used as a definition of the absolute value of any real number [32]. For any real numbers x and y, the absolute value will always have the following four fundamental properties [32]: See Figure 1.

$\|x\| \geq 0$	Non-negativity
$\|x\| = 0 \Leftrightarrow x = 0$	Positive-definiteness
$\|xy\| = \|x\|\|y\|$	Multiplicativeness
$\|x + y\| \leq \|a\| + \|b\|\|$	Sub-additivity

Figure 1: Fundamental properties of absolute value

C. *Deriving the similarity distance*

From the discussions above, it should now be clear that the absolute value of any real number is closely related to the idea of distance. The absolute value of any real number, therefore, is the distance from that number to the origin, along the real number line. For any given two real numbers x and y, the absolute value of the difference of x and y is the distance between them. The standard Euclidean distance between two points, for example x and y, defined in equation 2 affirms this, where $x = (x_1, x_2, \ldots, x_n)$ and $y = (y_1, y_2, \ldots, y_n)$. In the Euclidean n-space, the distance is defined as [27]:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \qquad (11)$$

Note that equation 11 is homogenous to equation 2. This can be viewed as a generalisation of $|x - y|$, since if x and y are two real numbers, then (from equation 10) we can define:

$$|x\text{-}y| = \sqrt{(x\text{-}y)^2} \qquad (12)$$

Equation 12 is homogeneous to equation 7 and equation 10. Equation 7 is used when computing the one-dimension Euclidean distance while equation 10 is used as a definition of the absolute value. Thus, equations 7, 10 and 12 prove that the 'absolute value' distance for any real numbers is equal to the Euclidean distance, defined in equation 7, when you consider them as either one and/or two-dimensional Euclidean spaces both defined in equation 7 and 8 respectively. Hence, the properties of the absolute value of the difference of any two real numbers (non-negativity, identity of indiscernibles, symmetry and the triangle inequality See Figure 2) agree with the concept of the distance function used to define the distance between the elements of a set. For any real value function f on a set X × X is called a distance function (or a metric) on X, if it satisfies the following four axioms [35] and [36]. See Figure 2.

$f(x,y) \geq 0$	Non-negativity (i)
$f(x,y) = 0 \Leftrightarrow x = y$	Identity of indiscernibles (ii)
$f(x,y) = f(y,x)$	Symmetry (iii)
$f(x,y) \leq f(x,z) + f(z,y)$	Triangle inequality (iv)

Figure 2: Distance function axioms

Note that condition (i) and (ii) together produce positive definiteness and the first condition is implied by the others. The technical background discussed in this section, especially the relationship between Euclidean distance and absolute value therefore simplifies the understanding of the proposed DFASSV method.

The Proposed DFASSV Method

Hit counts reported by Web search engines are useful information sources for this study and, as such, are used as input to this study. This is first explained in the next section, where after the calculation of the DFASSV method is explained.

D. Understanding the concept of 'hit Counts'

The hit count of a query as discussed earlier, is an estimated number of Web pages containing the queried term as reported by a Web search engine. In addition, the Web constitutes the largest semantic electronic "database" available on earth. Information can be accessed and extracted via any Web search engine that can return aggregate hit count estimates for a large range of search queries [1].The Web also provides semantic information for almost every known word or term. In some cases, semantics associated with each term or word is also described. In our approach, however, as explained earlier, we do not consider just the hit count for the logical *x* AND *y* search query as the only parameter for assessing the semantic similarity, but we also include the hit counts for the individual terms *x* and *y* before computing the semantic similarity value. We will, therefore, adopt the following notations in this paper:

f(*x*) denotes the hit count for the queried term *x*,
f(*y*) denotes the hit count for the queried term *y* and
f(*x*, *y*) denotes the hit count for the logically *x* AND *y* search query where both *x* and *y* appears together on the same Web page.

To calculate the Digital Forensic Absolute Semantic Similarity Value of *x* and *y*, denoted as DFASSV(*x*, *y*), we do not need to know the number of Web pages indexed by the Web search engine quoted as 8058044651 in [1]. This is so because according to [14] the process of estimating the number of pages indexed by a search engine can be a very difficult task. This paper, however, does not discuss the process of estimating the number of pages indexed by a search engine in any further detail. (For more information in this regard, please refer to [11]).

Our approach however, replaces the number of pages indexed by a search engine with a simple computed value (**T**) defined as the sum of the hit counts reported by the Web search engine for the search terms *x*, *y* and logical *x* AND *y* together.

Thus, $\mathbf{T} = f(x) + f(y) + f(x,y) \qquad (13)$

where *f*(*x*), *f*(*y*) and *f*(*x*, *y*) are as defined earlier.
Recalling the concept of the Euclidean distance and the absolute value at this point, we now establish their relationship with the proposed DFASSV method.

E. Digital Forensic Absolute Semantic Similarity Value (DFASSV)

In order to enhance communication among domain experts and also enable faster computation of meaning between computers in a computer digestible form, many long-term projects have been initiated to try and establish semantic relations between common objects and/or names of these objects. Good examples of

these projects include the CYC project [3] and the WordNet [4]. The idea is to create a semantic Web of such vast proportions that rudimentary intelligence and knowledge about the real world objects emerge spontaneously.

However, to achieve this, structures have to be properly designed with the ability to manipulate knowledge, and high quality contents have to be entered in these structures by knowledgeable human experts. While these efforts are good and take a long-term view, the overall information entered is very small when compared to what is available on the Web today [1]. We, therefore, take advantage in this study of the freely-available information on the Web and use it to calculate a semantic similarity measure between terms used in the digital forensics domain.

The proposed method in this paper, computes the semantic similarity value between two terms *x* and *y* in digital forensics, based on finding the one-dimensional Euclidean distance defined in equation 7, which is equal to finding the absolute value of the difference of any two real numbers. See equations 7 and12.

To begin with, the hit counts *f*(*x*), *f*(*y*), *f*(*x*, *y*) and the value of **T** for any two digital forensic terms *x* and *y* is obtained using the Google search engine. These parameters are then used as input to the proposed DFASSV method. **T** is however, computed using equation 13. There are four input parameters defined as *f*(*x*), *f*(*y*), *f*(*x*, *y*) and **T**. Using equation 12, which is similar to one-dimensional Euclidean distance (see equation 7); only two real numbers are needed as input. In order to establish a 1:1 mapping of the values of *x* and *y* in equation 12, DFASSV replaces the values of *x* and *y* with the percentage values of *f*(*x*) and *f*(*y*) computed as:

$\left(\frac{f(x)}{\mathbf{T}} * 100\right)$= percentage of the hit counts for the search term *x* and

$\left(\frac{f(y)}{\mathbf{T}} * 100\right)$= percentage of the hit counts for the search term *y*.

Substituting these values in equation 12 gives equation 14

$$|x - y| = \sqrt{\left(\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \left(\frac{f(y)}{\mathbf{T}} * 100\right)\right)^2} \quad (14)$$

The value obtained from equation 14 is in the fixed range of 0 per cent to 100 per cent. Treating the points *x* and *y* as Euclidean vectors, starting from the origin (0%) of the space, their tips (100%) indicate the distance between the two points.

As mentioned earlier, in one-dimensional Euclidean distance there is usually a single homogeneous, translation-invariant distance function (i.e. starting from the origin at least in one direction the object is infinite). For a similarity distance of 0 to ∞ instead of 0 per cent to 100 per cent, equation 14 is further modified as follows:

The values $\left(\frac{\mathrm{f(x)}}{\mathrm{T}} * 100\right)$ and $\left(\frac{\mathrm{f(y)}}{\mathrm{T}} * 100\right)$, denoted as percentage of the hit counts for the search term *x* and *y* respectively, are substituted by their computed logarithms as:

$\log\left(\frac{\mathrm{f(x)}}{\mathrm{T}} * 100\right)$ and $\log\left(\frac{\mathrm{f(y)}}{\mathrm{T}} * 100\right)$ respectively

Logarithm is a useful arithmetic concept used in all areas of science to help simplify the understanding of many scientific ideas. For example, logarithms may be defined and introduced in different ways as a means to simplify calculations. For the purposes of this study, we adopt a simple approach to simplify the computation of the Euclidean distance based on finding the absolute value of the difference of the logarithms of the hit count percentages of the terms *x* and *y*. There are no limits imposed on logarithms, thus their inputs and outputs can be in any range. Therefore, substituting these values in equation 14 gives rise to equation 15:

$$|x - y| = \sqrt{\left(\log\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \log\left(\frac{f(y)}{\mathbf{T}} * 100\right)\right)^2} \quad (15)$$

Equation 14 and 15 are both analogous to equation 7 and 12.

Equation 15 therefore, gives a value in the range of 0 to ∞ and can be re-written as equation 16, which is used to automatically compute the Absolute Semantic Similarity Value of the terms *x* and *y* in digital forensics denoted as DFASSV(*x*, *y*). Using the left hand side of equation 15, equivalent to the right hand side we can define DFASSV(*x*, *y*) as,

$$\mathrm{DFASSV}(x, y) = \left|\log\left(\frac{f(x)}{\mathbf{T}} * 100\right) - \log\left(\frac{f(y)}{\mathbf{T}} * 100\right)\right| \quad (16)$$

where

f(*x*) = the hit counts for the search term *x*, *f*(*y*) = the hit counts for the search term *y* and **T** = the sum of hit counts for the search terms *x* and *y* as defined in equation 13.

Equation 16, therefore, defines DFASSV(*x*, *y*), a new approach for calculating the semantic similarity between two terms *x* and *y* in digital forensics using Web search engines. In other

words equation 16 denotes DFASSV as the computed absolute value of the difference of the logarithms of the hit count percentages of terms *x* and *y*. The experimental results obtained using the new proposed DFASSV approach was found to be remarkable and are discussed in the section that follows.

Experimental Results

While the theory discussed in this paper is rather intricate, the resulting method is simple enough. Knowing that any given two digital forensics terms are perceived to be similar, the computed absolute semantic similarity value denoted by equation 16 can be used as a quick guide (proof) to show that the two given terms are truly semantically similar or not.

For example, given any two digital forensic terms *x* and *y*, we find the number of hit counts for search term *x* denoted as *f*(*x*), the number of hit counts for search term y denoted as *f*(*y*), the number of hit counts for logical *x* AND *y* both appearing together on one page denoted as *f*(*x, y*), and finally the sum of hit counts denoted as (**T**). **T** is computed using equation 13.

As a concrete example, let search term *x* be 'Digital evidence' and search term *y* be 'Electronic evidence'. Using the Google search engine with hit counts as reported for the search terms *x* and *y* as on 14 April 2012, it follows that:

"Digital evidence" *f*(*x*) =659000,
"Electronic evidence" *f*(*y*) =575000,
"Digital evidence"AND "Electronic evidence" *f*(*x, y*) =53900.

Therefore

T= $f(x) + f(y) + f(x,y)$ =1287900.

Substituting these values in equation 16 gives a semantic similarity measure of the terms 'Digital Evidence' and 'Electronic evidence' of **0.0592**. Since this value is relatively close to 0, it proves that the two terms are very closely related to the human-perceived meaning when used in digital forensics. It can also mean that, in case of a digital forensics investigation, the term 'Digital Evidence' can be used in the place of 'Electronic Evidence' without misleading the receivers of such information.

To further analyse the performance of the proposed method, we conducted two sets of experiments. First we compared the similarity scores produced by the proposed DFASSV method against the Miller and Charles benchmark data set [13] and [14]. Secondly, the proposed DFASSV approach was tested using digital forensics domain terms to measure its performance against the human-perceived meaning of the selected terms. These two experiments are discussed in the two sub-sections that follow respectively.

The Miller and Charles Benchmark Data Set

To assess the performance of the proposed DFASSV method, we evaluated it against the Miller and Charles data set [13]. The latter is a subset of Rubenstein and Goodenough's original data set of 65 word pairs [15]. As stated earlier, the Miller and Charles ratings are considered one of the most reliable benchmarks for evaluating semantic similarity measures.

The term pairs using the proposed DFASSV method are rated on a scale of 0 to ∞ (infinite), where 0 means identical semantic similarity and ∞ means no similarity. This is the opposite of the Miller and Charles dataset where word pairs are rated on a scale of 0 (dissimilarity) to 4 (identical semantic similarity). In summary, infer from the results that the smaller the value computed by the proposed DFASSV method, the more similar the terms (See Table I).This is also true from the correlation coefficient value of -**0.2777.** (Note that a negative correlation coefficient indicates that as one variable increases, the other decreases, and vice-versa.) This is further depicted by a graphical representation of the similarity measures in Table I, shown in Figure 3.

According to Cilibrasi and Vitànyi [1] Google events capture all background knowledge about the search terms concerned available on the Web. The Google event *x*, consists of a set of all Web pages containing one or more occurrences of the search term *x*. Thus, it embodies, in every possible sense, all direct context in which *x* occurs on the Web. This constitutes the Google semantics of the term *x* [1]. For this reason, in our experiments, the Google search engine was used.

The input to the DFASSV method is therefore the reported Google hit counts for any paired terms *x* and *y* from the digital forensics domain. The DFASSV method works by calculating the Euclidean distance between the terms *x* and *y*, equated to the computed absolute value of the difference of the logarithms of the hit count percentages of *x* and *y* as shown earlier in equation 16. Given any two terms *x* and *y* as points in the Euclidean plane, the associated computed absolute value of the difference of the logarithms of the hit count percentages of *x* and *y*, determines the similarity between the terms *x* and *y*.

Word Pair	M&C	Web Jaccard	Web Dice	Web Overlap	Web PMI	Proposed DFASSV
cord-smile	0.13	0.102	0.108	0.036	0.207	0.756
rooster-voyage	0.08	0.011	0.012	0.021	0.228	0.828
noon-string	0.08	0.126	0.133	0.060	0.101	0.524
glass-magician	0.11	0.117	0.124	0.408	0.598	1.399
monk-slave	0.55	0.181	0.191	0.067	0.610	0.389
coast-forest	0.42	0.862	0.870	0.310	0.417	0.055
monk-oracle	1.1	0.016	0.017	0.023	0	0.457
lad-wizard	0.42	0.072	0.077	0.070	0.426	0.400
forest-graveyard	0.84	0.068	0.072	0.246	0.494	1.258
food-rooster	0.89	0.012	0.013	0.425	0.207	1.778
coast-hill	0.87	0.963	0.965	0.279	0.350	0.248
car-journey	1.16	0.444	0.460	0.378	0.204	0.865
crane-implement	1.68	0.071	0.076	0.119	0.193	0.418
brother-lad	1.66	0.189	0.199	0.369	0.644	0.970
bird-crane	2.97	0.235	0.247	0.226	0.515	0.051
bird-cock	3.05	0.153	0.162	0.162	0.428	0.024
food-fruit	3.08	0.753	0.765	1	0.448	0.223
brother-monk	2.82	0.261	0.274	0.340	0.622	0.966
asylum-madhouse	3.61	0.024	0.025	0.102	0.813	0.945
furnace-stove	3.11	0.401	0.417	0.118	1	0.180
magician-wizard	3.5	0.295	0.309	0.383	0.863	0.638
journey-voyage	3.84	0.415	0.431	0.182	0.467	0.238
coast-shore	3.7	0.786	0.796	0.521	0.561	0.411
implement-tool	2.95	1	1	0.517	0.296	0.838
boy-lad	3.76	0.186	0.196	0.601	0.631	0.271
automobile-car	3.92	0.654	0.668	0.834	0.427	0.975
midday-noon	3.42	0.106	0.112	0.135	0.586	0.855
gem-jewel	3.84	0.295	0.309	0.094	0.687	0.027
Correlation	**1**	**0.259**	**0.267**	**0.382**	**0.548**	**-0.27 77**

Table I: comparison of Semantic Similarity of Human Ratings and Baselines on Miller AND Charles' dataset with DFASSV

The distance measure shown in Table II depicts the relatedness of the terms in question. Table I on the other hand, was used mainly for the purpose of comparison in order to indicate different semantic similarity measures from previously-proposed methods compared to those of the proposed DFASSV method. This was done to provide a clear picture of the performance and accuracy of DFASSV.

From Table I, the first column shows the word pairs used and column two indicates the ratings from Miller and Charles. Columns 3 to 6 also used for comparison show the ratings from previously proposed sematic similarity methods and the last column depicts the equivalence similarity measure computed using the proposed DFASSV. For example, the word pair 'gem-jewel' (See Table I) with a similarity measure of **3.84** from Miller and Charles and **0.027** from the proposed DFASSV clearly depicts the accuracy of DFASSV. This is also depicted in the other columns and indicates a better performance than that of some of the previously proposed methods. See Figure 3 for a graphical representation of the Table I results.

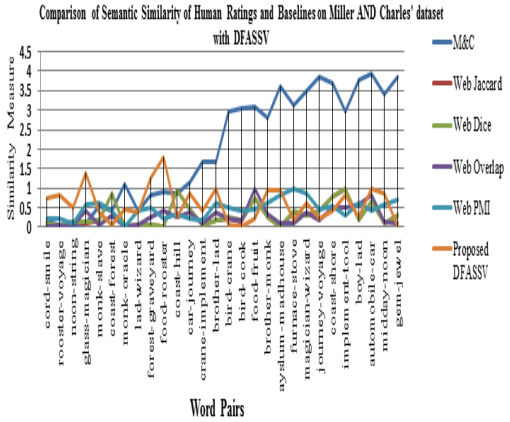


Figure 3: Comparison graph of the semantic similarity ratings and baselines on Miller and Charles' dataset with DFASSV (See Table I).

Digital Forensics Terminologies

In Table II, a part of the experimental findings is presented using the digital forensics domain terminologies. Each term enclosed in double quotes " " is used as a single Google search term denoted in Table II as *f(x)* and *f(y)* respectively. The computed DFASSV(*x*, *y*) using equation 16 therefore shows the semantic similarity measures obtained to ascertain the performance of the DFASSV method with the human-perceived meaning of the terms. The authors have no knowledge of other experiments of this kind in the digital forensics domain that can be used as a baseline to judge the performance of DFASSV. This is, therefore, a novel approach of using a Web search engine to determine the semantic similarity of terms in digital forensics.

The selected terms used are: 'digital evidence', 'digital forensics', 'electronic evidence', 'digital and multimedia evidence' [33] and [34] among other terms. The authors found that these terms are mostly used in discussions that involve the digital forensic investigation process and also in the accreditation of digital forensics laboratories, hence the motivation for the experiment indicated in Table II. In all the experiments conducted, DFASSV showed remarkable results.

To determine the semantic similarity measure of the terms as shown in Table II, the proposed DFASSV was used in all our experiments. The first two columns of Table II shows the digital forensics terms used for the experiments and their equivalent similarity measure indicated in the last column. Using the results in Table II, a random interview was conducted to a few digital forensics researchers and their understanding of these terms seemed to

agree with the results of the proposed DFASSV method.

Digital Forensics Terms		**Computed**
f(x)	***f(y)***	***DFASSV(x, y)***
Digital evidence	Electronic evidence	0.059217
Digital forensics	Digital evidence	0.431534
Digital forensics	Electronic evidence	0.490752
Electronic evidence	Digital and multimedia evidence	1.833840
Digital evidence	Digital and multimedia evidence	1.893057
Digital forensics	Digital and multimedia evidence	2.324592
Attacker	Adversary	0.357051
Cracker	Attacker	0.361608

Table II: Semantic Similarity ratings of digital forensic terms based on DFASSV

In the case of a digital forensic investigation for example, DFASSV can be used to determine the usage of terms where a similarity measure closer to 0 means that the two terms are closely related in meaning. The terms 'Digital evidence' and 'Electronic evidence', for example, with a similarity measure of **0.059217** indicates that they can be used interchangeable without causing confusion to the stakeholders. On the other hand a semantic similarity measure far from 0 would mean that the two terms are not closely related in meaning and therefore, one cannot replace the other. For example the terms 'Digital forensics' and 'Digital and multimedia evidence' with a similarity value of **2.324592** means they cannot be used interchangeable.

Application of The Proposed DFASSV Method in the Digital Forensics Domain

The proposed DFASSV method as demonstrated in this paper can be used in the digital forensics domain for example, to determine the semantic relatedness of terms and also as a way towards resolving the semantic disparities that exist in the domain. In addition, DFASSV can be used to help determine the most relevant and appropriate terminologies to use or included for example when building a specific ontology in the digital forensics domain. In addition, other future relevant undertakings in the digital forensics domain, in the authors' opinion, might as well benefit from applying such a method as DFASSV.

Conclusion

The problem that this paper addressed was that of the ever-changing technological trends in digital forensics where new terms are constantly introduced into the domain and new meanings assigned to existing terms.

In this paper a method was presented to automatically calculate a semantic similarity value between any two given digital forensics terms, using a new approach. Unlike previous methods, the Digital Forensic Absolute Semantic Similarity Value (DFASSV) approach proposed in this paper is unsupervised. No special background information is needed to understand and use this method because it utilises the existing bank of information from the Web by simply incorporating the hit counts between two digital forensics terms reported by any Web search engine. In addition, the authors also found that DFASSV is well suited for terminologies that originate from within the same domain.

Though the initial experiments were carried out on the digital forensics domain terms, the authors believe that the DFASSV method can be extended to other domains as well. This is due to the fact that the results of the experiments conducted to evaluate this method using the digital forensics domain terminologies are remarkable. The results show that this approach of measuring semantic similarity between two terms significantly outperforms some of the previous proposed measures.

As part of future research work, the authors are now planning to conduct an investigation in order to find out whether there are existing parameters other than hit counts reported by search engines that can be used with DFASSV to enhance the accuracy delivered by this method even more as a way towards resolving semantic disparities in the digital forensics domain

References

[1]. R.L. Cilibrasi and P.M.B. Vitànyi. 2007. The Google Similarity Distance. IEEE Transactions on Knowledge and Data Engineering, Vol. 19, No 3, March 2007, pp. 370–383.

[2]. D.Thiyagarajan, N. Shanthi and S. Navaneethakrishnan, A Web Search Engine-Based Approach To Measure Semantic Similarity Between Words. International Journal of Advanced Engineering Research and Studies. E-ISSN2249–8974.

[3]. D.B. Lenat, CYC: A large-scale investment in knowledge infrastructure, Communications of the ACM, November 1995/Vol. 38, No. 11.

[4]. G.A. Miller, WordNet, A Lexical Database for the English Language,

Cognitive Science Lab, PrincetonUniversity.

[5]. Zheng Xu et al. 2011. Measuring semantic similarity between words by removing noise and redundancy in web snippets. Concurrency and Computation: Practice and Experience, 23:2496–2510. Published online 22 September 2011 in Wiley Online Library (wileyonlinelibrary.com) DOI: 10.1002/cpe.1816.

[6]. R. Rada, H. Mili, E. Bichnell and M. Blettner. 1989. Development and application of a metric on semantic nets. IEEE Transactions on Systems, Man and Cybernetics, 9(1):17–30.

[7]. S. Vijay, 2012. A Combined Method to Measure the Semantic Similarity between Words, International Journal of Soft Computing and Engineering (IJSCE), ISSN: 2231–2307, Volume 1, Issue ETIC2011, January 2012.

[8]. Y. Matsuo, T. Sakaki, K. Uchiyama and M. Ishizuka. 2006. Graph- based word clustering using web search engine. In Proceedings of EMNLP 2006.

[9]. M. Li, X. Chen, X. Li, B. Ma and P.M.B. Vitànyi. 2004. The similarity metric. IEEE Transactions on Information Theory, 50(12):3250–3264.

[10]. D. Bollegala, Y. Matsuo and M. Ishizuka. 2009. A Relational Model of Semantic Similarity between Words using Automatically Extracted Lexical Pattern Clusters from the Web.

[11]. Anon., We knew the web was big... | Official Google Blog. Available at: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html [Accessed April 17, 2012].

[12]. H. Chen, M. Lin and Y. Wei. 2006. Novel association measures using web search with double checking. In Proceedings of the COLING/ACL 2006, pp. 1009–1016.

[13]. G.A. Miller and W.G. Charles. 1998. Contextual correlates of semantic similarity. Language and Cognitive Processes. Volume 6, Issue 1.

[14]. Z. Bar-Yossef and M. Gurevich. 2006. Random sampling from a search engine's index. In Proceedings of the 15th International World Wide Web Conference.

[15]. H. Rubenstein and J.B. Goodenough. 1965. Contextual Correlates of Synonymy. Computational Linguistics. Decision Sciences Laboratory, L.G. Hanscom Field, Bedford, Massachusetts.

[16]. W.G. Charles, Contextual Correlates of Meanings. Applied Psycholinguistics 21 (2000) 505–524. Cambridge Journals. Available at: http://journals.cambridge.org/action/displayFulltext?type=1&fid=66932&jid=APS&volumeId=21&issueId=04&aid=66931 [Accessed April 20, 2012]. p. 514.

[17]. Anon., Simpson's Rule. Available at: http://pages.pacificcoast.net/~cazelais/187/simpson.pdf [Accessed May 7, 2012].

[18]. Anon., Distance and Similarity Coefficients. Available at: http://paleo.cortland.edu/class/stats/documents/11_Similarity.pdf [Accessed May 7, 2012].

[19]. P. Sanchez, R. Milson and M. Slone. Euclidean distance (version 11). Available at: http://planetmath.org/EuclideanDistance.html [Accessed May 7, 2012].

[20]. Bogomolny, Pythagorean Theorem and its many proofs.Interactive Mathematics Miscellany and Puzzles. Available at: http://www.cut-the-knot.org/pythagoras/index.shtml [Accessed May 7, 2012].

[21]. Bogomolny, The Distance Formula.Interactive Mathematics Miscellany and Puzzles. Available at: http://www.cut-the-knot.org/pythagoras/DistanceFormula.shtml [Accessed May 7, 2012].

[22]. D.T. Larose. 2005. Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Hoboken, NJ.

[23]. G.B. Dantzig and M.N. Thapa. 1997. Linear Programming: Introduction. p. 147. Section 6.3. Hamilton Printing, Rensselaer, NY.

[24]. Anon, Vectors in Euclidean Spaces. Available at: http://scottmccracken.weebly.com/uploads/9/0/6/6/9066859/vectors-print_version.pdf [Accessed May 10, 2012].

[25]. D.G Bailey, An Efficient Euclidean Distance Transform**,** IWCIA 2004, LNCS 3322, pp. 394–408, 2004.

[26]. Anon, Complex Vector Spaces and Inner Products. Available at:

http://college.cengage.com/mathematics/larson/elementary_linear/4e/shared/downloads/c08s4.pdf [Accessed May 11, 2012].

[27]. R. Balu and T. Devi, Identification Of Acute Appendicitis Using Euclidean Distance On Sonographic Image. International Journal of Innovative Technology & Creative Engineering (ISSN: 2045-8711**).** VOL.1 NO.7 JULY 2011

[28]. Per-Erik Danielsson, Euclidean Distance Mapping, Computer Graphics and Image Processing 14, 227-248 (1980)

[29]. Anon, Absolute Value. Available at: http://www.purplemath.com/modules/absolute.htm [Accessed May 11, 2012].

[30]. Anon, Absolute Value Functions. Available at: http://hotmath.com/hotmath_help/topics/absolute-value-functions.html [Accessed May 11, 2012].

[31]. Anon, Absolute Value Functions. Department of Mathematics, College of the Redwoods Available at: http://msenux.redwoods.edu/IntAlgText/chapter4/chapter4.pdf [Accessed May 11, 2012].

[32]. Anon,Absolute Value. Available at: http://math.ucalgary.ca/sites/math.ucalgary.ca/files/courses/F07/MATH251/lec5/MATH251-F07-LEC5-Appendix_E.pdf [Accessed May 11, 2012].

[33]. Palmer, A Road Map for Digital Forensic Research. DFRWS TECHNICAL REPORT. DTR - T001-01 FINAL. Report from the First Digital Forensic Research Workshop (DFRWS). November 6th, 2001 - Final.

[34]. NATA, Proficiency Testing Policy in the Field of Forensic Science. Available at: http://www.nata.asn.au/phocadownload/publications/Technical_publications/Policy_Tech_circulars/technical-circular-15.pdf [Accessed March 12, 2012].

[35]. J. Fennell, Axiomatics Through The Metric Space Axioms**.** Metric Space Axiomatics. University College Cork

[36]. T. Margush, Distances Between Trees**.** Discrete Applied Mathematics 4 (1982) 281-290North-Holland Publishing Company.

An Ontological Framework for a Cloud Forensic Environment

[1,2]Nickson M. Karie[*], [1]H.S. Venter[†]
[1]Department of Computer Science, University of Pretoria,
Private Bag X20, Hatfield 0028, Pretoria, South Africa
[2]Department of Computer Science, Kabarak University,
Private Bag - 20157, Kabarak, Kenya
E-mail: menza06@hotmail.com[*], hventer@cs.up.ac.za[†]

Abstract
Cloud computing is an emerging field and is considered to be one of the most transformative technologies in the history of computing. This is so because it is radically changing the way how information technology services are created, delivered, accessed and managed. Cloud forensics, on the other hand, is utilising network forensics – a subset of digital forensic techniques – in a cloud environment. However, with the continued evolution from internet-based applications to cloud computing, the environments and components surrounding cloud forensics can easily become incomprehensible.

In this paper, therefore, we present an ontological framework meant to provide a structure and depiction of the different cloud environments and components an investigator should be acquainted with, in the case of a cloud investigation process. In addition, we show the relationships and interactions between the different environments by capturing their content and boundaries. Furthermore, the purpose of this paper is meant to provide a common ontological framework for sharing coherent cloud computing concepts and also promote the understanding of the cloud environments and cloud components. Finally, the ontological framework presents an approach towards structuring and organizing the environments and components surrounding the cloud and constitutes the main contribution of this paper.

Keywords
Cloud forensics, cloud computing, cloud environments, cloud components, ontological framework

Introduction
With the emergence of cloud computing technologies, the need for cloud forensics has become inevitable. This is due to the notion of cloud computing opening a whole new world of possibilities for criminals to exploit. This also means that criminals can now use cloud computing environments to share information and to reinforce their hacking techniques (Garfinkel, 2011). As a result, the major potential security risks, such as malicious insiders, data loss/leakage and policy violations now invade the existing cloud environments.

Cloud forensics, as defined by Ruan et al (2011), is an emerging field that deals with the application of digital forensic techniques in cloud computing environments and forms a subset of network forensics. Technically, cloud forensics follows most of the main phases of network forensic processes. The only difference is that such phases are simply extended with techniques tailored for cloud computing environments within each phase. However, the continued widespread deployment of the Internet-based applications and network-enabled devices in an effort to support mechanisms for cloud computing, can potentially render the cloud environments and components incomprehensible.

In this paper we present an ontological framework in an attempt to provide a structure and depiction of the different cloud environments (cloud deployment models) and cloud components (cloud service models) that an investigator should be well-versed with in the case of an investigation processes involving the cloud. In addition, the proposed framework also shows the relationships and interactions between the different cloud environments and the cloud components. Furthermore, this paper provides a novel contribution and offers a simplified ontological framework that can, for example, help investigators comprehend the cloud environment and components with less effort.

As for the remaining part of this paper, section 2 presents previous and related work while section 3 briefly explains the cloud environments and components. The proposed ontological framework is presented in section 4 followed by a discussion in section 5. Finally, section 6 presents the conclusion and future work.

Related Work

There exist several frameworks in cloud computing proposed by other researchers, which have made valuable contributions towards the development of the ontological framework presented in this paper. In this section, therefore, a summary of some of the most prominent efforts in previous research work is provided.

To begin with, Hoefer and Karagiannis (2010) argues that several organisations want to explore the possibilities and benefits of cloud computing. However, with the amount of cloud computing services increasing quickly, the need for taxonomy frameworks rises. In their paper they describe the available cloud computing services and propose a tree-structured taxonomy based on their characteristics, in order to easily classify cloud computing services so that it is easier to compare them. However, in this paper, we focus on an ontological framework meant to provide a common framework to share coherent cloud computing concepts as well as to promote the understanding of cloud environments and essential cloud components. Such a framework will assist investigators, for example, in planning of investigation techniques to be employed in specific cloud environments in the case of an investigation process and thus enhancing the investigation of criminal cases involving the cloud.

Yan (2011) argues that cloud computing, as a service, provides a luring environment for criminals and increases the difficulties of digital forensics. He then presents a forensic framework that focuses on the security issues of cloud services in order to beat cybercrime. Yan's framework, however, focuses on security issues of cloud services while we, in the current proposed ontological framework, focus on structuring and organising the different cloud environments and cloud components.

In their paper, Takahashi et al (2010) propose an ontological approach to cybersecurity in cloud computing. They built an ontology for cybersecurity operational information based on actual cybersecurity operations mainly focused on non-cloud computing. In order to discuss necessary cybersecurity information in cloud computing, they apply the ontology to cloud computing. Their work is centred on cybersecurity operations. However, the current framework is centred on, as mentioned earlier, cloud environments and cloud components.

Lamia et al (2009) also explains that the progress of research efforts in a novel technology is contingent on having a rigorous organisation of its knowledge domain and a comprehensive understanding of all the relevant components and their relationships of the technology. In their paper, they propose an ontology for cloud computing which demonstrates a dissection of the cloud into five main layers. However, there work does not elaborate on the cloud environments and cloud components in the way that is presented in this paper.

There also exist other related works on ontological frameworks, but neither those nor the cited references in this paper have presented an ontological framework for the cloud environments and cloud components in the way that is introduced in this paper. However, we acknowledge the fact that the previous proposed frameworks have offered useful insights toward the development of the ontological framework in this paper. In the section that follows we briefly explain the different cloud environments and components based on our review of the literature.

Cloud Environments and cloud Components

Cloud Computing is an emerging technology that uses the internet and remotely located servers to maintain data and applications. The 'cloud', therefore, can be viewed as a network of virtual machines geographically dispersed. Cloud computing technology is creating a revolution in computer architecture, software and tools development. Furthermore, it is changing the way organizations store, distribute and consume information. In this section of the paper, the authors explain the different cloud environments and cloud components that form the basis of the proposed ontological framework.

The Cloud Environments (Cloud Deployment Models)

Public Cloud Environment

A public cloud is one in which a service provider makes resources, such as applications, platforms and infrastructures available to the general public over the internet. Public clouds are owned and operated at datacentres belonging to the service providers and are shared by multiple customers (Subramanian, 2011a). This also means that, public clouds offer unlimited storage space and increased bandwidth via internet to any organisation across the globe. Such services on the public cloud may be offered free or on a pay-per-usage

model. The degree of visibility and control of public clouds depends on the delivery mode. However, there is less visibility and control in public clouds compared to private clouds because the underlying infrastructure is owned by the service providers.

Private Cloud Environment
A private cloud can be viewed as the implementation of cloud computing services on resources dedicated to an organisation (i.e. the organisation owns the hardware and software), whether they exist on-premises or off-premises. A private cloud gives an organisation the advantage of greater control over the entire stack, from the bare metal up to the services accessible to users (Ubuntu, 2013).

Community Cloud Environment
A Community cloud is one that is tailored to the shared needs of a business community. Community clouds are operated specifically for a targeted group. Usually, such groups (communities) have similar cloud requirements and their ultimate goal is to work together to achieve their business objectives. According to Techopedia (2013), community clouds are often designed for businesses and organisations working on joint projects, applications, or research, which requires a central cloud computing facility for building, managing and executing such projects, regardless of the solution rented. The infrastructure in a community cloud is shared by several organizations with common concerns such as (security, compliance, jurisdiction, etc.), whether managed internally or by a third-party or hosted internally or externally. The cost is, however, shared by all the participating organizations.

Hybrid Cloud Environment
A hybrid cloud is a combination of both public and private clouds (Subramanian, 2011b). This means that a vendor who owns a private cloud can form a partnership with a public cloud provider, or a public cloud provider can form a partnership with a vendor that provides private cloud platforms. However, according to Mell and Grance (2011) of the National Institute of Standards and Technology (NIST), a hybrid cloud is a composition of two or more public, private, or community cloud infrastructures that remain unique entities but are bound together by either standardised or proprietary technology that enables data and application portability. Using the hybrid cloud architecture, organisations and individuals are able to obtain degrees of fault tolerance combined with locally immediate usability without dependency on internet connectivity. This is due to some of the resources in a hybrid cloud being managed in-house while others are provided externally. In the next sub-section the authors elaborate on the essential cloud components which also form part of the proposed ontological framework in this paper.

The Essential Cloud Components (Cloud Service Models)
Whichever the cloud environment deployed, cloud service providers will always offer their clients (individuals and organisations) with the following three categories of cloud service models: Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). In the next sub-sections, these service models are further explained.

Infrastructure-as-a-Service (IaaS)
IaaS is a cloud computing service model that offers physical and virtual systems (cloud computing infrastructure), including an operating system, hypervisor, raw storage, and networks (Oracle Corporation, 2012). Servers represent the main computing resource in IaaS and are often virtual instances within a physical server. The service providers usually own the computing infrastructure and are responsible for housing, running and maintaining it. On the other hand, organisations pay on a per-use basis. IaaS helps organisations realize cost savings and efficiencies while modernising and expanding their information technology capabilities without spending capital resources on infrastructure (GAS, 2013).

Platform-as-a-Service (PaaS)
PaaS as explained in an expert group report by the European Commission (2010) provides computational resources (cloud computing platforms) via a platform upon which applications and services can be developed and hosted. PaaS typically makes use of dedicated APIs to control the behaviour of a server hosting engine which executes and replicates the execution according to user requests. Cloud computing platforms may include the operating system, the programming language execution environment, the database, and the web server. PaaS also allows clients to use the virtualised servers and associated services for running applications or developing and testing new applications.

Software-as-a-Service (SaaS)
SaaS sometimes referred to as Service or Application Clouds (European Commission, 2010) offers implementations of specific business functions and business processes that are provided with specific cloud capabilities. I.e. they provide cloud computing applications or services using a cloud infrastructure or platform, rather than providing cloud features themselves. Moreover, SaaS also provides internet-based access to different software, thus presenting new opportunities for software vendors to explore. In the next section, the proposed ontological framework is presented and explained.

The Proposed Ontological Framework
In this section of the paper the authors present the proposed ontological framework. Figure 1 shows the structure of the ontological framework. Note that, due to the small font size of Figure 1, Figures 2 to 4 contains enlarged extracts of the ontological framework as depicted in Figure 1.

The framework consists of five layers arranged from left to right and with the first layer depicting the main domain of focus (i.e. the cloud/cloud computing). This is followed by the cloud environments in the second layer and the essential cloud components in the third layer. Services and service providers are introduced in the fourth and fifth layer of the ontological framework as a way of representing individual, finer-grained details of the essential cloud components, also referred to as cloud service models.

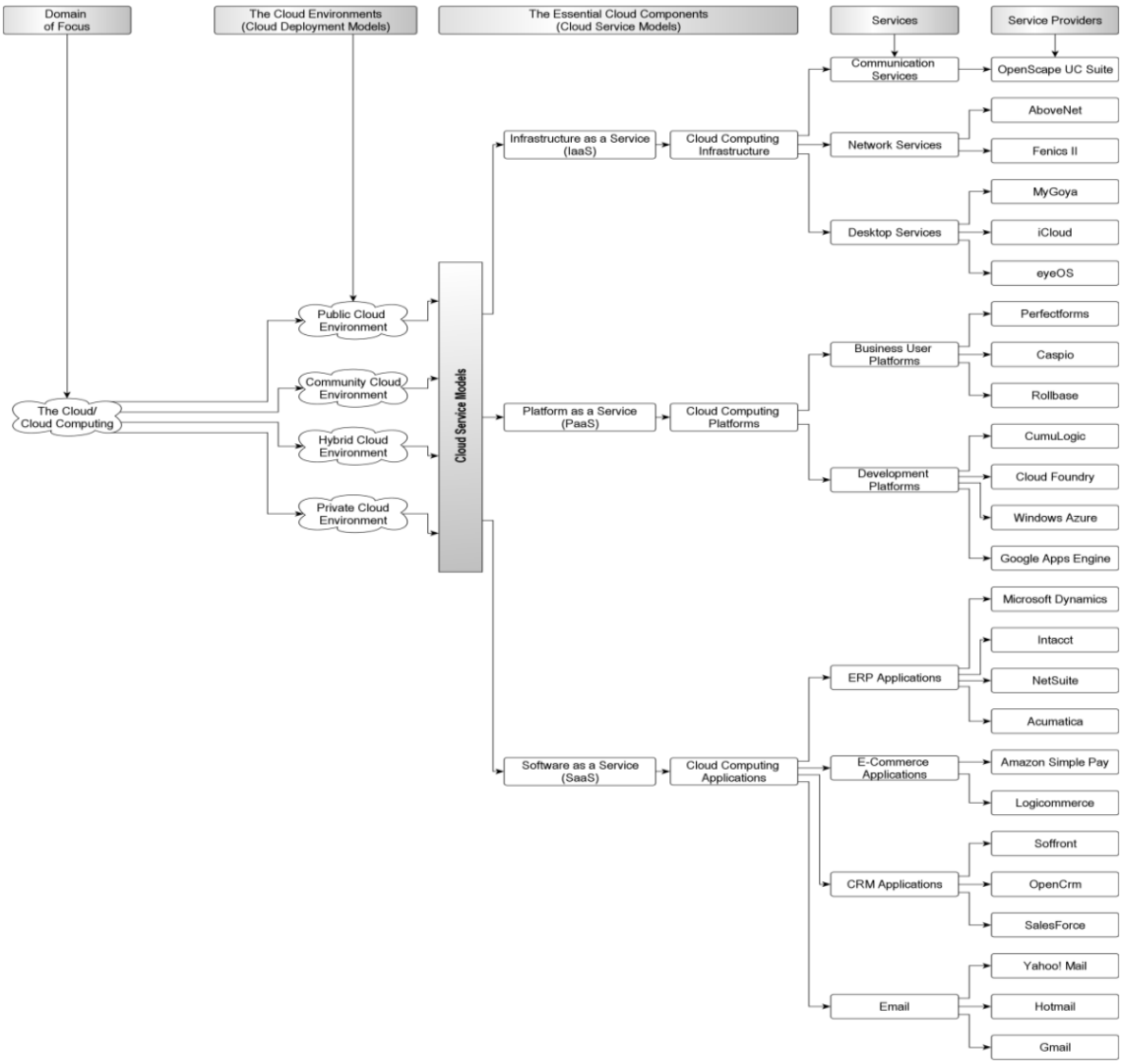


Figure 1: Conceptualisation of the cloud environments and essential components

Cloud service models enable software, platform and infrastructure to be delivered as services. The term service is used to reflect the fact that they are provided on demand and are paid for, on a usage basis (Czarnecki, 2011). In the authors' experience, organising the framework into the particular cloud environments, essential cloud components, services and service providers, was necessary to simplify the understanding of the framework as well as to present specific finer details of the framework. The services and service providers listed in the fourth and fifth layers (see Figure 1) were only selected as common examples to facilitate this study and should not be treated as an exhaustive list.

The major areas explored (with their details as shown in Figure I) include the cloud environments, the essential cloud components, services and the service providers. For the purpose of this study, the cloud environments (cloud deployment models) are divided into public cloud environment, private cloud environment, community cloud environment and hybrid cloud environment. The essential cloud components (cloud service models), on the other hand, are divided into Infrastructure-as-a-Service (IaaS), Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS). However, infer from Figure 1 that the IaaS, PaaS and SaaS are accessible through cloud computing infrastructure, cloud computing platforms and cloud computing applications respectively.

The cloud computing infrastructure (see Figure 2) is further divided into communication services, network services and desktop services forming the fourth layer of the ontological framework. The communication services show OpenScape UC Suite as one of the service providers. The network services have AboveNet™ and Fenics II as service providers. Finally, desktop services show MyGoya, iCloud and eyeOS as service providers. The service providers form the fifth layer of the framework as shown in Figure 1. However, note that, the contents of the fourth and fifth layer (services and service providers) in Figure1 were introduced in this framework to provide only selected examples for the purpose of this study. Therefore, such contents should not be treated as an exhaustive list.

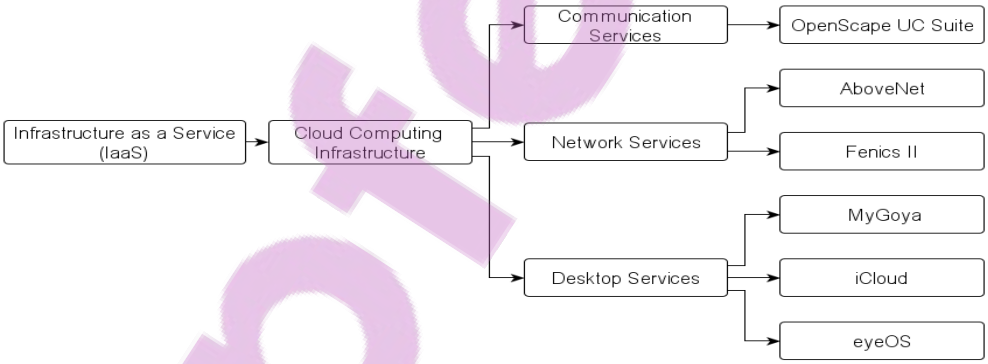


Figure 2: Infrastructure-as-a-Service

The cloud computing platforms as shown in Figure 3 are divided into two: business user platforms and development platforms. Business user platforms have PerfectForms, Caspio™ and Rollbase as service providers. The development platforms show CumuLogic, Cloud Foundry™, Windows Azure™, and Google™ Apps Engine as selected service providers. However as said earlier these are only common examples for the purpose of this study and should not be treated as an exhaustive list.

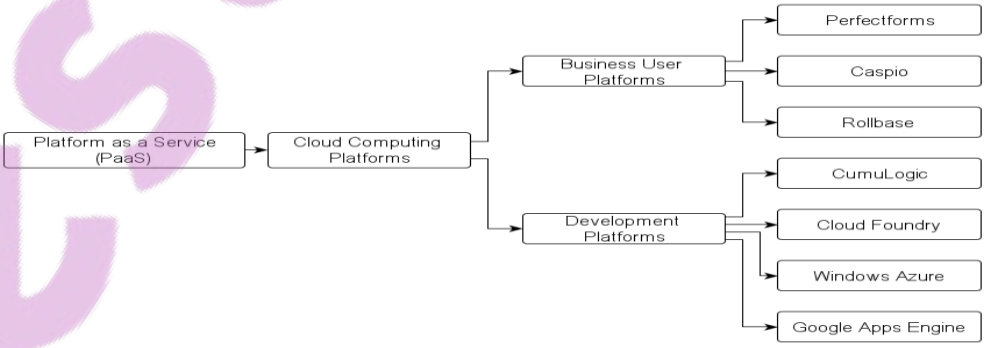


Figure 3: Platform-as-a-Service

The cloud computing applications shown in Figure 4 are divided into Enterprise Resource Planning (ERP) applications, E-commerce applications, Customer Relationship Management (CRM) applications and Email as selected examples. The ERP applications have Microsoft Dynamics™, Intacct®, NetSuite and Acumatica as service providers. E-commerce applications show Amazon Simple Pay and Logicommerce™ as examples of service providers. The CRM applications have Soffront®, OpenCrm and SalesForce® as service providers. Finally, Email has Yahoo!®, Hotmail® and Gmail™ as examples of the service providers. As said earlier, these were only selected as common examples for the purpose of this framework and, therefore, should not be treated as an exhaustive list.

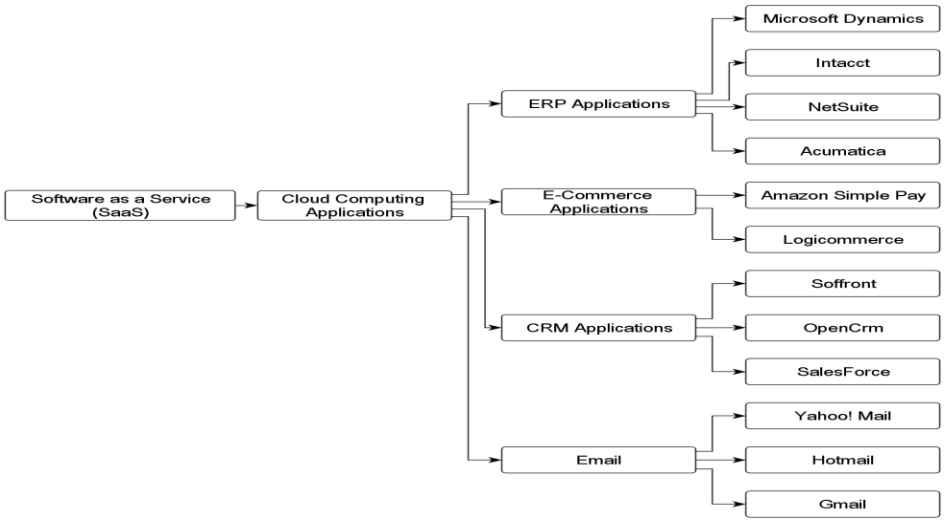


Figure 4: Software-as-a-Service

Discussion

The ontological framework presented in this paper is a new contribution and its scope is defined by the cloud environments, the essential cloud components, services and the service providers (see Figure 1). Such an ontological framework can be used, for example, as a common platform to share coherent cloud computing concepts and also promote the understanding of the cloud environments and cloud components. Moreover, the ontological framework can also serve, for example, as a basis for sharing common views of the structure and depiction of cloud computing information in a bid to enable the reuse of domain knowledge.

Furthermore, the framework in this paper can, for example, help investigators to explicitly describe investigation processes and procedures that focus on specific cloud environments in the case of cloud forensics. In addition, forensic tools developers can also use the ontological framework to fine-tune their tools so as to be able to cover as many potential security risks and policy violations experienced in the different cloud environments. This also implies that developers will find the ontological framework in this paper constructive, especially when considering new cloud forensic techniques for specific cloud environments.

In the case of cloud forensics, the proposed ontological framework can also assist in the design and development of high-tech acquisition tools incorporating, for example, hybrid cloud architectural designs with shareable features such as automated acquisition, reporting, visualisation and presentation of evidence in a manner that is acceptable in a court of law. Moreover, such high-tech tools will also enhance the investigation of criminal cases involving multiple cloud computing environments.

The proposed ontological framework can also be useful, for example, in cloud interoperability and exchanging of information between the different cloud environments. Moreover, it can be helpful in the design and development of standardised technology that also enables data and application portability in the different cloud environments. This is backed up by the fact that, the framework has explicitly described the distinctions of the various cloud environments, essential cloud components, services and service providers shown in Figure 1.

Finally, the ontological framework is, therefore, a new contribution towards advancing the field of cloud computing. To the best of the authors' knowledge, there exists no other work of this kind and, therefore, this is a novel contribution towards advancing the cloud computing and cloud forensic domain.

Conclusion and Future Work

The problem addressed in this paper was that of the incomprehensible cloud environments and components we are currently faced with. This incomprehensibility has been caused by the continued evolution from internet-based applications to cloud computing. In this paper we have proposed an ontological framework that provides a structure and a depiction of the different cloud environments and cloud components as a way to help individuals comprehend them with less effort. In addition, the cloud environments, the essential cloud components, services and service providers were also captured in the framework and explained. Therefore, the authors believe that by using this ontological framework a better understanding of the cloud environments and associated cloud components can be gained. However, more research needs to be conducted in order to identify new components and also to improve on the proposed ontological framework in this paper. Finally, the framework should spark further discussion on the development of new cloud computing ontological frameworks.

References

Czarnecki, C., (2011), "Cloud Service Models: Comparing SaaS PaaS and IaaS", *Perspectives on Cloud Computing & Training from Learning Tree International*. Available at: http://cloud-computing.learningtree.com/2011/11/09/cloud-service-models-comparing-saas-paas-and-iaas/ [Accessed February 13, 2013].

European Commission, (2010), Editors: Jeffery, K. and Neidecker-Lutz, B., "The future of cloud computing", opportunities for European cloud computing beyond 2010. *Expert Group Report*

GAS, (2013), Infrastructure as a Service (IaaS). Available at: http://www.gsa.gov/portal/content/112063 [Accessed March 20, 2013].

Garfinkel, S.L., (2011), The Criminal Cloud, *MIT Technology Review*, Available at: http://www.technologyreview.com/news/425770/the-criminal-cloud/ [Accessed February 4, 2013].

Hoefer, C.N., Karagiannis, G., (2010), "Taxonomy of cloud computing services", *Proceedings of the GLOBECOM Workshops*, pp.1345-1350

Lamia, Y., Butrico, M., and Da Silva, D., (2008), "Toward a Unified Ontology of Cloud Computing", *Proceedings of the Grid Computing Environments Workshop*, pp.1-10

Mell, P. and Grance, T., (2011), "The NIST Definition of cloud computing", *Recommendations of the National Institute of Standards and Technology*.

Oracle Corporation, (2012), "Making Infrastructure-as-a-Service in the Enterprise a Reality", *An Oracle White Paper*.

Ruan, K., Carthy, J., Kechadi, T. and Crosbie, M., (2011), "Cloud forensics", *Proceedings of the 7th IFIP WG 11.9 International Conference on Digital Forensics 2011*, Orlando, FL, USA

Subramanian, K., (2011a), "Public Clouds", *A whitepaper sponsored by Trend Micro Inc.*

Subramanian, K., (2011b), "Hybrid Clouds", *A whitepaper sponsored by Trend Micro Inc.*

Takahashi, T., Kadobayashi, Y. and Fujiwara, H., (2010) "Ontological Approach toward Cybersecurity in Cloud Computing", *Proceedings of the 3rd international conference on Security of information and networks (SIN '10), ACM, New York, NY, USA*, pp 100-109

Techopedia, (2013), "Community Cloud", Available at: http://www.techopedia.com/definition/26559/community-cloud [Accessed February 8, 2013]

Ubuntu, (2013), "Private cloud", Available at: http://www.ubuntu.com/cloud/private-cloud [Accessed February 8, 2013].

Yan, C., (2011), "Cybercrime Forensic System in Cloud Computing", *Proceedings of the Image Analysis and Signal Processing (IASP) Conference*, pp.612-615

Significance of Semantic Reconciliation in Digital Forensics

Nickson M. Karie
[1]Department of Computer Science, University of Pretoria,
Private Bag X20, Hatfield 0028, Pretoria, South Africa
[2]Department of Computer Science, Kabarak University,
Private Bag - 20157, Kabarak, Kenya
menza06@hotmail.com

H.S. Venter
[1]Department of Computer Science, University of Pretoria,
Private Bag X20, Hatfield 0028, Pretoria, South Africa
hventer@cs.up.ac.za

ABSTRACT
Digital forensics (DF) is a growing field that is gaining popularity among many computer professionals, law enforcement agencies and other stakeholders who must always cooperate in this profession. Unfortunately, this has created an environment replete with semantic disparities within the domain that needs to be resolved and/or eliminated. For the purpose of this study, semantic disparity refers to disagreements about the meaning, interpretation, descriptions and the intended use of the same or related data and terminologies. If semantic disparity is not detected and resolved, it may lead to misunderstandings. Even worse, since the people involved may not be from the same neighbourhood, they may not be aware of the existence of the semantic disparities, and probably might not easily realize it.

The aim of this paper, therefore, is to discuss semantic disparity in DF and further elaborates on how to manage it. In addition, this paper also presents the significance of semantic reconciliation in DF. Semantic reconciliation refers to reconciling the meaning (including the interpretations and descriptions) of terminologies and data used in digital forensics. Managing semantic disparities and the significance of semantic reconciliation in digital forensics constitutes the main contributions of this paper.

Keywords: Digital forensics, semantic disparity, managing semantic disparity, semantic reconciliation, significance of semantic reconciliation

1. INTRODUCTION

Digital forensics plays a very important role in both incident detection and digital investigations. However, the investigation process in most cases demands cooperation between the computer professionals, law enforcement agencies and other forensic practitioners. Unfortunately, this has created an environment replete with semantic disparity within the domain that needs to be resolve and/or eliminated. Semantic disparity as defined by Xu and Lee (2002) refers to disagreements about the meaning, interpretation, description and the intended use of the same or related data. Moreover, according to Oxford Dictionaries (2013), disparity refers to the state of being different (lack of uniformity). If semantic disparity is not detected and resolved in digital forensics, it may lead to misunderstandings. In addition, semantic disparity may become a serious problem, for example, when trying to harmonise data/information from different sources (Piasecki, 2008).

Moreover, in the case of a digital forensic investigation process, the cooperation between the computer professionals, law enforcement agencies and other forensic practitioners presupposes the reconciliation of semantic disparities that are bound to occur in the domain. Unfortunately, DF lacks comprehensive methodologies, specifications and ontologies that can assist in resolving the semantic disparities that exist between the different digital forensic practitioners.

In this paper, therefore, we discuss semantic disparities in DF and further elaborate on how to manage it. In addition, this paper also presents the significance of semantic reconciliation in digital forensics. Furthermore, the presentation in this paper is a novel contribution that offers a simplified comprehension of semantic

disparities in digital forensics. Moreover, this paper is also meant to spark further discussions on the development of methodologies and specifications for resolving semantic disparities in DF.

As for the remaining part of this paper, section 2 presents background concepts of semantic disparity while section 3 elaborates on how to manage semantic disparities in digital forensics. The significance of semantic reconciliation in digital forensics is handled in section 4. Finally, conclusions and future research work are considered in section 5.

2. BACKGROUND

In this section of the paper, the authors present background concepts on semantic disparities. Note that, semantic disparity as discussed in this paper is sometimes addressed as semantic heterogeneity in other previous research works (Xu and Lee, 2002; Sheth and Larse, 1990; Wang and Liu, 2009). However, for the purpose of this paper we adopt the use of the term semantic disparity in place of semantic heterogeneity.

To begin with, Sheth and Larsen (1990) argue that, semantic disparity is a problem that is not well understood in many domains and in the case of this paper digital forensics as well. There is not even an agreement regarding a clear definition of this problem (Xu and Lee, 2002; Sheth and Larse, 1990). However, different researchers have identified different forms of semantic disparity that are worth mentioning. A majority of these semantic disparities, however, focus more into the field of databases while others focus on distributed systems.

According to Lin et al. (2006), the problem of semantic disparity is extremely critical in situations of extensive cooperation and interoperation between distributed systems across different enterprises. In the case of digital forensics, for example, such a situation would make it difficult to manipulate distributed data/information in a centralized manner. This is because; the contextual requirements and the purpose of the information across the different systems may not be homogeneous.

Another effort by Colomb (1997) presented the case for structural semantic disparity (structural semantics define the relationships between the meanings of terminologies). Bishr (1998) on the other hand, elaborates on schematic disparity. The major problem as presented by Colomb (1997) lies in what can be called the fundamental conceptual disparity. Fundamental conceptual disparity occur when the terms used in two different ontologies, for example, have meanings that are similar, yet not quite the same (Xu and Lee, 2002). Schematic disparity, on the other hand, arises when information that is represented as data in one schema, is represented within the schema (as metadata) in another (Bishr, 1998; Miller, 1998).

Although the database perspective on semantic disparity is good and offers insights (Xu and Lee, 2002), it limits the understanding of semantic disparity and how to manage it in other domains. In the section that follows, therefore, we elaborate on how to manage semantic disparities focusing on the digital forensic domain.

3. MANAGING SEMANTIC DISPARITIES IN DIGITAL FORENSICS

Managing semantic disparities in a growing field like digital forensics can be a daunting task. This is because; the technological trends in DF are ever-changing; new terminologies are constantly introduced into the domain and new meanings assigned to existing terms (Karie and Venter, 2012). Therefore, methodologies and specifications need to be developed in digital forensics with the ability to effectively assist in managing semantic disparities that may crop up as a result of technological change or domain evolution. Such methodologies will further assist in establishing an efficient semantic reconciliation process in the domain. Furthermore, the requirement for semantic reconciliation methodologies and specifications in digital forensics is exceptionally important both for the advancement of the field as well as for the effective use of different domain terminologies and the representation of domain information.

Therefore, understanding the different potential circumstances and conflicts under which semantic disparity may arise in digital forensics can be of great significance in establishing a meaningful semantic reconciliation process.

3.1. Potential Conflicts that can Cause Semantic Disparity in Digital Forensics

Semantic disparity may occur in digital forensics, for example, when the communicating parties (computer professionals, law enforcement agencies, forensic practitioners, etc.) use different meanings, interpretations, descriptions and representations of the same or related domain terminologies and data. This causes variations in the understanding of domain information and how it is specified and structured in different components. This also implies that, perfect communication between the sender and the receiver of the information will be scanty. Having the ability to identify and avoid semantic disparities in digital forensics can assist investigators, for example, in decision making.

In the sub-sections that follow, therefore, we survey and present (based on our review of the literature) various conflicts (including examples where applicable) that can cause disparities in DF. Note that the conflicts discussed in this section only serves as common examples to facilitate this study and should not be treated as an exhaustive list.

3.1.1. Semantic Conflicts

Semantic conflicts occur when different people involved in the same domain do not perceive exactly the same set of real world objects, but instead they visualize overlapping sets (Bishr, 1998). As a result, disagreement about the meaning, interpretation and the descriptions of the same or related data and terminologies occur. Table 1 shows examples of the semantic conflicts (descriptions and interpretation of terminologies) in digital forensics.

DF Terminology	Descriptions
• First response	Include the first response to the detected incident (Valjarevic and Venter, 2012).
• Initial response	Perform an initial investigation, recording the basic details surrounding the incident, assembling the incident response team, and notifying the individuals who need to know about the incident (Mandia et al., 2003).
• Incident response	Consists of the detection and initial, pre-investigation response to a suspected computer crime related incident, such as a breach of computer security. The purpose of Incident response is also to detect, validate, assess, and determine a response strategy for the suspected security incident (Beebe and Clark, 2005).

Table 1: Semantic Conflicts in Digital Forensic Terminologies.

3.1.2. Descriptive Conflicts

Descriptive conflicts include naming conflicts due to homonyms and synonyms, as well as conflicts on attribute domain, scale, cardinalities, constraints, operations etc. (Bishr, 1998; Sheth and Gala, 1989; Larson et al. 1989). In the case of digital forensics, descriptive conflicts can occur, for example, when two terminologies representing related ideas of the domain concepts are described using different sets of properties. Table 2 present some of the descriptive conflicts identified in the digital forensic domain. Note that the terminologies in Table 1 and Table 2 are only selected examples to facilitate this study and by no means an exhaustive list.

DF Terminology	Descriptions
• Analysis	Determine significance, reconstruct fragments of data and draw conclusions based on evidence found. The distinction of analysis is that it may not require high technical skills to perform and thus more people can work on this case (Reith et al., 2002).
• Analysis	Analysis involves the use of a large number of techniques to identify digital evidence, reconstruct the evidence if needed and interpret it, in order to make hypothesis on how the incident occurred, what its exact characteristics are and who is to be held responsible (Valjarevic and Venter, 2012).
• Analysis	The use of different forensic tools and techniques to make sense of the collected evidence (Sibiya et al., 2012).

• Examination	Examination is an in-depth analysis of the digital evidence and is the application of digital forensic tools and techniques that are used to gather evidence (Lalla and Flowerday, 2010).
• Examination	An in-depth systematic search of evidence relating to the suspected crime. This focuses on identifying and locating potential evidence, possibly within unconventional locations. Construct detailed documentation for analysis (Reith et al., 2002).

Table 2: Descriptive Conflicts in Digital Forensic Terminologies.

The authors found that the terminologies in Table 1 and 2 are mostly used by digital forensic investigators and the law enforcement agencies during and after a digital forensic investigation process, hence the motivation for this study.

3.1.3. Structural Conflicts

Structural conflicts occur when two or more people use the same model, but choose different constructs to represent common real-world objects (Lee and Ling, 1995). In the context of digital forensics structural conflicts can occur, for example, when different domain members use the same digital forensic investigation process model but choose different constructs to present their results/findings. Note that, the term constructs, is used to mean ideas or theories containing various conceptual elements, and considered to be subjective but not based on any empirical evidence (Houts and Baldwin, 2004).

After attending several sessions of expert testimony (potential evidence presentation) in court and civil proceedings the authors found that different constructs are used by different digital forensic experts to convince the court that the potential digital evidence presented is worthy of inclusion into the criminal process. However, the constructs used during potential evidence presentation were based on experience rather than standardised guidelines or digital forensic logics. This is backed up by the fact that, there are currently no standardised guidelines for even presenting the most common representations of potential digital forensic evidence in court or civil proceedings (Cohen, 2011). In the sub-section that follows, we explain different approaches that can assist in managing semantic disparity in DF.

3.2. Different Approaches to Manage Semantic Disparity

There exist different approaches that can assist in resolving semantic disparities in digital forensics (Farshad and Andreas, 2001). However, as with other examples explained earlier, the list discussed in this section present only selected examples and therefore should not be treated as an exhaustive list.

3.2.1. Building Ontologies

Ontologies can help deal with the problem of semantic disparity by providing formal, explicit definitions of data and reasoning over related concepts. Moreover, ontologies in most cases capture the conceptualization of experts in a particular domain of interest (Falbo et al., 1998). Ontology mapping can also be employed to find semantic correspondences between similar elements of different ontologies, thus allowing people to agree on terms that can be used when communicating (Noy, 2004).

In digital forensics, building a proper domain ontology in terms of its explication and its accordance with the conceptualization of domain experts can help in managing the semantic disparity that occurs in the domain. However, according to Kajan (2013), considering that anyone can design ontologies according to his/her own conceptual view of the world, care must be observed during the process of designing ontologies because, ontological disparity among different parties can become an inherent characteristic.

3.2.2. Representation of Ontologies and Reasoning Based on these Ontologies

According to Farshad and Andreas (2001), the representation of ontologies and reasoning based on these ontologies makes it possible to capture and represent ontological definitions and the important features that can be used in representing ontologies for reasoning. In the case of digital forensics such an approach would help create clear definitions of the different terminologies used in the domain. Moreover, this approach can also assist in managing semantic disparity in DF because the relationships that hold among domain terminologies can be realized and structured. For more information in this regard we refer the reader to (Palmer, 2001; Caloyannides, 2004 & Crouch, 2010) respectively.

3.2.3. Semantics Integration

Semantics integration deals with the process of interrelating information from diverse sources to create a homogeneous and uniform semantic of use (Noy, 2004). In the case of digital forensics, this can make communication easier by providing precise concepts that can be used to construct domain information. Furthermore, semantic integration can facilitate or even automate communication between different systems thus offering the ability to automatically link different ontologies (Gardner, 2005).

3.2.4. Explicit use of common shared semantics

The explicit and formal definitions of semantics of terms have always guided many researchers to apply formal ontologies (Guarino, 1998) as a potential solution of semantic disparity. A formal ontology usually consists of logical axioms that convey the meaning of terms for a particular domain (Bishr et al, 1999; Kottman, 1999). Furthermore, formal ontologies are usually concerned with the understanding of the members of the domain and help to reduce ambiguity in communication (Farshad and Andreas, 2001), understanding, representation and interpretations of information.

In the next section, we present the significance of semantic reconciliation in digital forensics.

4. SIGNIFICANCE OF SEMANTIC RECONCILIATION IN DIGITAL FORENSICS

While there are a lot of research activities in digital forensics even at the time of this study very little have been towards semantic reconciliation. The authors believe that, semantic disparity in any domain can alter the context as well as the purpose of any information delivered by an individual and thus should be avoided. In digital forensics, methodologies and specifications need to be developed that can effectively assist in semantic reconciliation. Furthermore, such methodologies and specifications can also be used, for example, as fundamental building blocks in resolving the present and future semantic disparities in the domain. Semantic reconciliation, in the authors' opinion, is a promising conception towards resolving semantic disparities in digital forensics. The sub-sections that follow will explain in more details some of the significances of semantic reconciliation in digital forensics.

4.1 Perfect Communication

Semantic disparities can be a serious barrier to perfect communication in any domain. Semantic reconciliation, on the other hand, can be used to bridge the semantic gap between different communicating parties thus bringing with it perfect communication in the domain (Parsons and Wand, 2003). This also implies that, information between the different digital forensic stakeholders (computer professionals, law enforcement agencies and other digital forensic practitioners) can be interpreted in such a way that the sender's desired effect is achieved. Moreover, after a security incident has occurred, for example, if the communication, interpretation and representation of information are done correctly, it is much easier and useful in apprehending the attacker, and stands a much greater chance of being admissible in the event of a prosecution (Brezinski and Killalea, 2002). Wrong interpretation and representation of evidence information, on the other hand, might create loopholes for intruders to escape and thus making it had to convict and prosecute them. Therefore, semantic reconciliation in digital forensics is inevitable if perfect communication is to be achieved.

4.2 Common Understanding

Semantic disparities may arise in digital forensics as a result of different representation or interpretation of terminologies and data; this may include the use of different alternatives or definitions to describe the same domain information. However, with semantic reconciliation the different digital forensic experts can achieve common understanding by reconciling the meaning of terms thus having common representation or interpretation of domain terminologies (Parsons and Wand, 2003). This also implies that, the meaning of information as interpreted by the receiver will align with the meaning intended by the sender (Anon, 2013). In the case of court or civil proceedings common understanding will also help different stakeholders treat queries conveniently and at the same time maintaining consistency in their understanding of the various digital forensic terminologies and data used during such proceedings.

4.3 Correct Interpretation

When two or more independent digital forensic practitioners with varying professional backgrounds are to cooperate during an investigation process, semantic conflicts may occur. It is, therefore, very important and

critical that semantic disparities be resolved and/or eliminated to facilitate correct interpretation of domain information. Semantic reconciliation is one of the ways that can improve on correct interpretation through detecting the semantic similarities between the different terminologies and data used by the independent practitioners to describe or represent domain information (Parsons and Wand, 2003).

4.4 High-levels of collaboration

Many organisations are increasingly promoting collaborations as an important feature in organisation management (Tschannen-Moran, 2001). However, effective collaborations demands reasoning as well as effective communication. Therefore, semantic reconciliation in digital forensics can lead to high-levels of collaborations between the computer professionals, law enforcement agencies and other digital forensic practitioners. Furthermore, semantic reconciliation can also help create uniformity in the use of both terminologies and data in the digital forensic domain thus easing cooperation.

4.5 Uniform Representation of Domain Information.

In the case of potential evidence presentation in any court of law, information conveyed with very many semantic variances can be semantically unreliable. Therefore, semantic reconciliation can help create uniform representation of domain information. This is backed up by the fact that, semantic reconciliation can also make interpretation and representation of domain information much easier and more accurate (Wang et al., 2005).

4.6 Faster Harmonisation of Information from Different Sources

Efficient information management and processing have become more and more important within enterprises or when enterprises are merging together (Ubbo et al. 2002). Moreover, to achieve semantic interoperability across information system using different terminologies, the meaning of the information that is interchanged has to be harmonised across the systems (Ubbo et al. 2002). However, semantic disparity may arise whenever two contexts do not use uniform interpretation of the same information. Therefore, the use of semantic reconciliation for the explication of implicit and hidden knowledge is a promising approach to overcome the problem of semantic disparity in digital forensics and can assist in faster harmonisation of information from different sources.

4.7 Less Errors during Analysis of Potential Digital Evidence Information

Errors in analysis and interpretation of digital evidence, in the case of an investigation process, are more likely where there are semantic disparities. Even more where there are no standardised procedures or formal representation of domain information (Chaikin, 2006). Semantic reconciliation, on the other hand, will enable computer professionals, law enforcement agencies and practitioners in digital forensics to agree on terminologies or keywords to be used in representing certain key information in the case of an investigation and also establish keyword structures so that their relationship to each other are easily known. This will enhance the analysis of potential digital evidence information in the domain.

5. CONCLUSION AND FUTURE WORK

The problem addressed in this paper was that of semantic disparity in digital forensics. Different approaches to manage semantic disparities in digital forensics have also been explained. Moreover, the paper has also elaborated on the significance of semantic reconciliation in the digital forensic domain. The presentation in this paper is a new contribution in digital forensics and is meant to spark further discussion on the development of methodologies and specifications for sematic reconciliation in the domain. As part of the future work, the authors are now engaged in a research project to try and develop specification and/or ontologies that will create a unified formal representation of the digital forensic domain knowledge and information. In addition, the authors also aim at developing a digital forensic semantic reconciliatory model as a way towards resolving the semantic disparities that occur in digital forensics. However, there is still much research to be carried out so as to provide directions on how to address semantic disparities in the digital forensic domain. More research also needs to be conducted in order to add on the work discussed in this paper.

REFERENCES

Anon, (2013), A Communication Model. Available at: http://www.worldtrans.org/TP/TP1/TP1-17.HTML [Accessed April 25, 2013].

Beebe, N.L., and Clark, J.G., (2005), "A Hierarchical,Objectives-Based Framework for the Digital Investigations Process". Published in Digital Investigation 2(2). pp 146-16

Bishr Y.A., (1998). Overcoming the Semantics and Other Barriers to GIS Interoperability. *International Journal of Geographic Information Science*, Vol. 12, No. 4, pp299-314

Bishr, Y. A.; Pundt, H.; Kuhn, W., and Radwan, M. (1999). Probing the concept of information communities-a first step toward semantic interoperability. *Interoperating Geographic Information Systems*, Kluwer Academic

Brezinski, D. and Killalea, T, (2002), Guidelines for Evidence Collection and Archiving. Available at: http://tools.ietf.org/html/rfc3227 [Accessed April 25, 2013].

Caloyannides, M.A., (2004). Privacy Protection and Computer Forensics. Second Edition, Artech House, 2004.

Chaikin, D., (2006), "Network investigationis of cyber attacks: the limits of digital evidence", Crime Law Soc. Change, vol. 46, pp. 239-256.

Cohen, F., (2011), Digital Forensic Evidence Examination. 3RD Edition. Published by fred cohen & Associates. ISBN # 1-878109-46-4

Colomb R.M. (1997). Impact of Semantic Heterogeneity on Federating Databases, *The Computer Journal*, Vol. 40, No. 5 p235-244.

Crouch, J.-Ed (2010). NSCI - An Introduction to Computer Forensics, Available at: http://www.nsci-va.org/WhitePapers/2010-12-16-Computer%20Forensics-Crouch-final.pdf [Accessed March 5, 2012].

Falbo, R.A., Menezes, C.S. and Rocha, A.R., (1998) A Systematic Approach for Building Ontologies. Proceeding of the 6th Ibero-American Conference on AI: Progress in Artificial Intelligence. pp. 349-360

Farshad H. and Andreas G, (2001) Resolving Semantic Heterogeneity in Schema Integration: an Ontology Based Approach, University of Zurich. *International Conference on Formal Ontology in Information Systems (FOIS)*, Ogunquit, Maine, USA.

Gardner, S.P. (2005), Ontologies and semantic data integration. DDT, Vol.10, No 14, pp. 1001-1007

Guarino, N., (1998). "Formal Ontology in Information Systems," *Proceedings of FOIS'98*, Trento, Italy.

Houts, A.C. and Baldwin, S., (2004), Constructs, operational definition, and operational analysis. Applied & Preventive Psychology, Vol.11, pp. 45–46

Kajan, E., (2013). Electronic Business Interoperability: Concepts, Opportunities and Challenges - Google Books. Available at: http://books.google.co.za/books?id=fNh2Frjj7oUC&pg=PA287&lpg=PA287&dq=Even+if+two+ontologies+use+the+same+name+for+a+concept,+the+associated+properties+and+the+relationships+with+other+concepts+are+most+likely+to+be+different&source=bl&ots=NQ74gKNzP-&sig=yxThC1hAO27mIwqhAkoJUlPAOUI&hl=en&sa=X&ei=gNUwUeiuMI-LhQfdloDYBQ&ved=0CDYQ6AEwAg#v=onepage&q=Even%20if%20two%20ontologies%20use%20the%20same%20name%20for%20a%20concept%2C%20the%20associated%20properties%20and%20the%20relationships%20with%20other%20concepts%20are%20most%20likely%20to%20be%20different&f=false [Accessed March 1, 2013].

Karie, N.M. and Venter, H.S., (2012) Measuring Semantic Similarity Between Digital Forensics Terminologies Using Web Search Engines. In the Proceedings of the 12th Annual Information Security for South Africa Conference. Johannesburg, South Africa. Published online by IEEE Xplore®.

Kottman, C., (1999). Semantics and Information Communities, the OpenGIS Abstract Specification Topic 14. Ver. 4. *OpenGIS Consortium*, OpenGIS™ Project Document Number 99-114.doc.

Lalla, H., and Flowerday, S.V., (2010), Towards a Standardised Digital Forensic Process: E-mail Forensics. In proceeding of the Information Security South Africa Conference. Sandton, South Africa

Larson, J.A., Navathe, S.B. and Elmasri, R., (1989), "A Theory of Attribute Equivalence in Databases with Application to Schema Integration", IEEE Transactions On Software Engineering, Vol. SE-15, No. 4.

Lee, M.L. and Ling, T.W., (1995), Resolving Structural Conflicts in the Integration of Entity-Relationship Schemas. Object-Oriented and Entity-Relationship Modeling Vol. 1021, pp. 424-433

Lin, Y., Strasunskas, D., Hakkarainen, S., Krogstie, J., and Solvberg, A., (2006), "Sematic Annotation Framework to Manage Semantic Heterogeneity of Process Models", *Proceedings of the 18th international conference on Advanced Information Systems Engineering*, pp. 433-446

Mandia,K., Prosise, C., and Pepe, M., (2003), "Incident Response & Computer Forensics" (Second Ed.), McGraw-Hill/Osborne, Emeryville.

Miller, R.J., (1998), "Using schematically heterogeneous structures", *Proceedings of the 1998 ACM SIGMOD international conference on Management of data*, pp. 189-200
Noy, N.F. (2004). "Semantic Integration: A Survey of Ontology- based Approaches", *SIGMOD Record*, Vol. 33, No. 4, pp65-70
Oxford Dictionaries, (2013), Definition of disparity in Oxford Dictionaries (British & World English). Available at: http://oxforddictionaries.com/definition/english/disparity [Accessed April 12, 2013].
Palmer, G., (2001). A Road Map for Digital Forensic Research, DFRWS Technical Report. DTR - T001-01 FINAL. *Report from the First Digital Forensic Research Workshop* (DFRWS).
Parsons, J. and Wand, Y., (2003), Attribute-Based Semantic Reconciliation of Multiple Data Sources. Journal on Data Semantics I, Vol. 2800, pp 21-47
Piasecki, M., (2008). "HydroTagger: A Tool for Semantic Mapping of Hydrologic Terms". *AAAI Spring Symposium: Semantic Scientific Knowledge Integration*, page 77-80.
Reith, M., Carr, C., and Gunsch, G., (2002), An Examination of Digital Forensic Models. International Journal of Digital Evidence. Vol.1, No.3
Sheth A.P. and Larse, J., (1990). Federated database systems for managing distributed, heterogeneous, and autonomous databases. *ACM Computing Surveys (CSUR) - Special issue on heterogeneous databases Surveys*. Vol. 22 No. 3, pp.183 – 236
Sheth, A.P. and Gala, S.K. (1989), "Attribute Relationships: An Impediment in Automating Schema Integration", Proceddings of theWorkshop on Heterogeneous Database Systems, Chicago
Sibiya, G., Venter, H.S., Ngobeni, S., and Fogwill, T., (2012), Guidelines for Procedures of a Harmonised Digital Forensic Process in Network Forensics. In proceeding of the Information Security South Africa Conference. Sandton, South Africa
Tschannen-Moran, M. (2001). Collaboration and the need for Trust, Journal of Education Administaration Vol. 39, No. , pp. 308-331
Ubbo, V., Stuckenschmidt, H., Schlieder, C., Wache, H., Timm, I., (2002),Terminology Integration for the Management of distributed Information Resources. Künstliche Intelligenz, Vol. 16, pp. 31–34
Valjarevic, A. and Venter, H.S., (2012), Harmonised Digital Forensic Investigation Process Model. In proceeding of the Information Security South Africa Conference. Sandton, South Africa
Wang, H. and Liu, J.N.K., (2009), Analysis of Semantic Heterogeneity Using a new Ontological Structure Based on Description Logics. Sixth International Conference on Fuzzy Systems and Knowledge Discovery
Wang, X., Ausdal, S.V. and Zhou, J., (2005), Managing the Life Cycle of Business Semantics. Available at: http://xtensible.net.s60489.gridserver.com/wp-content/uploads/managing-the-life-cycle-of-business-semantics.pdf [Accessed April 25, 2013].
Xu, Z., and Lee, Y.C. (2002). Semantic Heterogeneity of Geo Data, *Symposium on Geospatial Theory, Processing and Applications*, Ottawa.

APPENDIX B: PAPERS PUBLISHED IN INTERNATIONAL JOURNAL

Towards a general ontology for digital forensic disciplines is a paper which is discussed as part of chapter 7 of this research thesis and has been published by the Journal of Forensic Sciences.

Taxonomy of challenges for digital forensics which is part of chapter 3 of this research thesis has also been accepted for publication by the Journal of Forensic Sciences after full review.

Based on the research work conducted in this research thesis several additional manuscripts were published both at international conferences and scientific journals and are as shown in Appendix C. However, the above mentioned Journal papers are presented on the pages to follow.

Towards a General Ontology for Digital Forensic Disciplines

1,2Nickson M. Karie* MSc, 1H.S. Venter† PhD
1Department of Computer Science, University of Pretoria,
Private Bag X20, Hatfield 0028, Pretoria, South Africa
2Department of Computer Science, Kabarak University,
Private Bag - 20157, Kabarak, Kenya
Email: menza06@hotmail.com*, hventer@cs.up.ac.za†

ABSTRACT: Ontologies are widely used in different disciplines as a technique for representing and reasoning about domain knowledge. However, despite the widespread ontology-related research activities and applications in different disciplines, the development of ontologies and ontology research activities are still wanting in digital forensic disciplines. This paper therefore presents the case for establishing an ontology for digital forensic disciplines. Such an ontology would enable better categorisation of digital forensic disciplines, as well as help with the development of methodologies that can offer direction in different areas of digital forensics, such as professional specialisation, certifications, development digital forensic tools, curricula and educational materials. In addition, the ontology presented in this paper can be used, for example, to better organise digital forensics domain knowledge and explicitly describe the discipline's semantics in a common way. Finally, this paper is meant to spark discussions and further research on an internationally agreed ontological distinction of the digital forensic disciplines. Digital forensic disciplines ontology is a novel approach towards organising the digital forensics domain knowledge and constitutes the main contribution of this paper.

KEYWORDS: forensic science, digital forensics, ontology, ontological distinction, digital forensics disciplines, digital forensics sub-disciplines

Ontology, as defined by Van Rees (1), is a set of well-defined concepts describing a specific domain of interest. According to Grüber (2), an ontology is a specification of a conceptualisation. More precisely, Smith et al (3) defines ontology as an explicit formal specification of how to represent entities that exist in a given domain of interest and the relationships that hold among them. However, for an ontology to be useful, it must represent a shared, agreed-upon conceptualisation (4), in other words it should be accepted by a group or community.

Ontologies have been used in many contexts and for many purposes (5). In recent years, however, the development of ontology has become common in many different domains (6). This is backed up by the fact that ontologies can be used to generate a common definition, knowledge and understanding (1) of a domain. Therefore, to help create a common definition that enhances the sharing and reuse of formal represented knowledge (2) in digital forensics (DF), it is important to develop ontologies that define the common entities in which the shared knowledge in this field can be represented. Ontologies in DF can also promote the reasoning about existing disciplines and sub-disciplines within the domain, as well as describe the domain.

This paper presents an ontology for the DF disciplines in an attempt to advance the domain and enhance the sharing and reuse of formal represented knowledge (2) in DF. In the authors' opinion, the ontology presented here can be viewed as a formal way of representing shared knowledge in the digital forensics domain. It can also be used to organise and reason over existing digital forensics disciplines in such a way that deductive inferences can be made (7).

The presentation in this paper is, therefore, a novel contribution in the digital forensics domain and offers a simplified platform that can help individuals comprehend the existing DF disciplines with much less effort. Moreover, the ontology has been simplified to accommodate new digital forensic disciplines and sub-disciplines that may crop up in the future as a result of technological change or domain evolution. Finally, individuals, organisations and academic institutions with an interest in areas of professional specialisation, certification, and development of digital forensic tools, curricula and/or development of educational materials should find the ontology constructive.

Background

Digital forensics is a relatively new and growing field (10) that is gaining popularity among many computer professionals, law enforcement agencies, practitioners and other stakeholders who need to cooperate in this profession. In addition, there is a strong demand for standardisation in many areas of digital forensics, for example the digital forensic investigation process (58). The number of forensic models that exist has added to the complexity of the field (60) and has led to a call for standardisation (62) so as to facilitate the investigation process (61). Recent research has also urged the need for new forensic techniques and tools that will be able to successfully investigate anti-forensics methods (59).

In a growing field like DF, developing practical methodologies for different areas is essential and as important as the research itself. Methodologies need to be developed for areas such as professional specialisation, certification, and development of digital forensic tools, curricula and/or development of educational materials. The authors believe that the ontology presented in this paper can help to provide direction in different areas of DF (such as those mentioned above).

Ontologies have been widely used in different fields as a technique for representing and reasoning about domain knowledge (1, 5). In addition, ontologies can be used to better organise domain knowledge and explicitly describe domain semantics in a common way.

As discussed by Brusa et al (12) ontology development can be divided into two phases: a specification phase and a conceptualisation phase. The goal of the specification phase is to acquire informal knowledge about the domain. In the case of this paper, the goal of the conceptualisation phase is to organise and structure this knowledge by using external representations. Basically, the main reasons for developing an ontology in any domain are to share a common understanding of the structure of information among entities in a bid to enable the reuse of domain knowledge and to make explicit those assumptions about a domain that are normally implied (13). If assumptions that underlie an implementation are made explicit in an ontology, then it is relatively easy to change the ontology when knowledge about the domain changes (13).

Hence, developing ontologies that define the common entities in which shared DF knowledge can be represented can help create uniformity and common understanding in representing DF disciplines. In the authors' opinion, uniformity and a common understanding can as well enhance and improve cooperation among computer professionals, law enforcement agencies and practitioners in the case of a digital forensic investigation. In the section that follows we examine ontology-related work in the digital forensics domain.

Related Work

Very little literature on issues related to ontology development for the digital forensics domain was available at the time of writing this paper. As a matter of fact, even what is present in literature seems to be somewhat varied. However, several previously proposed ontologies within the digital forensics domain have made valuable contributions to the development of the ontology in this paper. What follows hereafter is therefore a summary of some of the related research work on ontology development in digital forensics.

To begin with, in 2006 Brinson et al (8) presented a detailed cyber-forensics ontology in an effort to create a new way of studying cyber forensics. This ontology consists of a five-layered hierarchical structure with the final layer being specified areas that can be used for certification and specialisation. In a different paper, David and Richard (9) introduced the Small-Scale Digital Device Forensics (SSDDF) ontology. They proposed an ontology to provide law enforcement with the appropriate knowledge regarding the devices found in the Small-Scale Digital Device (SSDD) domain. Additionally, they suggest that this ontology can be used as a method to further the development of a set of standards and procedures at which to approach SSDD.

Jasmine and Zoran (63) in there paper highlights the problems encountered by investigators in the pursuit of forensic investigations of digital devices, primarily because of misunderstanding or false understanding of certain important concepts. They further propose an ontology of digital evidence as one of possible method suitable as a solution of this problem.

In 2009 Allyson and Doris (10) discussed the concept of 'Weaving Ontologies to Support Digital Forensic Analysis'. In their paper they argue that numerous challenges currently face digital forensic analysis. Although there are a variety of techniques and tools to assist with the analysis of digital evidence, they inadequately address key problems such as the vast volumes of data, lack of unified formal representation or

standardised procedures, incompatibility among heterogeneous forensic analysis tools, lack of forensic knowledge reuse, and lack of sufficient support for legal criminal/civil prosecution (10). Their paper goes further and suggests the applicability and usefulness of weaving ontologies to address some of these problems. It introduces an ontological approach that can lead to future development of automated digital forensic analysis tools.

Turk (11) presents an ontology that can be used to map a research area, design a curriculum, structure the agenda of a conference, provide keywords and classifications for bibliographic databases, or provide knowledge management in general.

There also exist other related works on ontologies, but neither those nor the cited references in this paper have presented an ontology of the digital forensics disciplines in the way that is introduced in this paper. We obviously acknowledge the fact that the previous work on ontologies has offered useful insights toward the development of the ontology in this paper. In the section that follows we provide a detailed explanation of our ontology on the digital forensic disciplines.

The Digital Forensics Disciplines Ontology

In this section of the paper, we present a detailed explanation of the ontology on the digital forensics disciplines and sub-disciplines within the domain of DF. Figure 1 shows the structure of the ontology. Note that, due to the small font size of Figure 1, Figures 2 to 7 contains enlarged extracts of the entire ontology as depicted in Figure 1.

The ontology consists of five layers arranged from left to right and with the first layer depicting the main domain of focus (i.e. digital forensics). This is followed by the DF disciplines in the second layer, and the sub-disciplines within the DF domain in the third layer. Objects and sub-objects are introduced in the fourth and fifth layers of the ontology as a way of representing individual and specific finer details of the sub-disciplines within DF. In the authors' opinion, organising the ontology into disciplines, sub-disciplines, objects and sub-objects was necessary to simplify the understanding of the ontology as well as to present specific finer details of the ontology.

In addition, the sub-disciplines, objects and sub-objects presented in the ontology focus more on areas that can be considered for professional specialisation and certification, as well as for the development of digital forensic tools, curricula and educational materials. However, infer from the ontology in Figure 1 that the objects and sub-objects listed were only selected as common examples to facilitate this study and should not be treated as an exhaustive list. More sub-disciplines, objects and sub-objects can and should be added as the need arises in future.

Note that from the ontology in Figure 1, some of the objects presented do not have sub-objects; in the authors' opinion, breaking them down to a finer-grained level would be superficial at this stage.

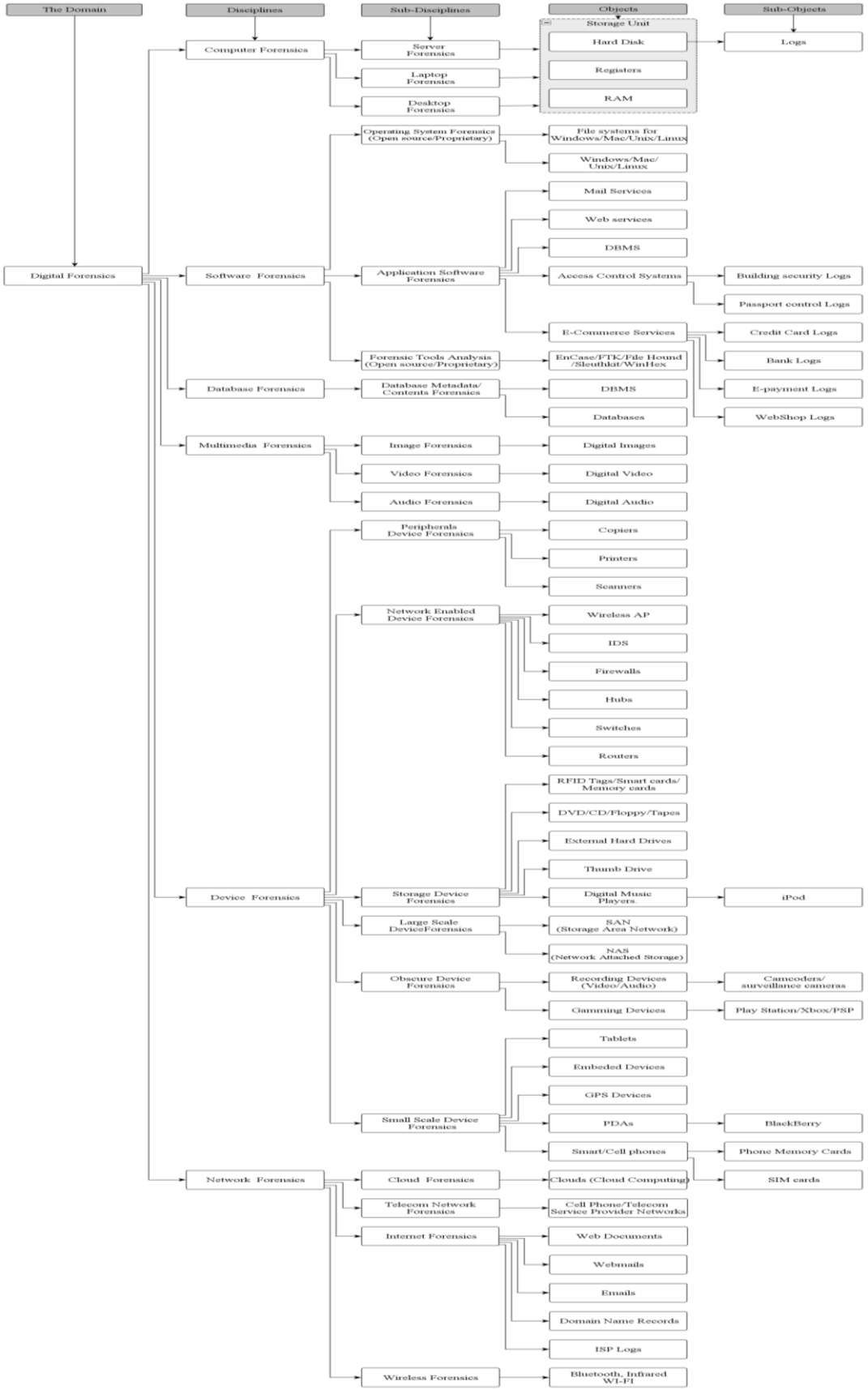


Figure 1: The DF disciplines ontology

However, in future it should be possible to mention sub-objects that can be incorporated under the applicable objects, especially when developing curricula and education materials. The major digital forensics disciplines explored in this study (with their details as shown in Figure 1) include computer forensics, software forensics, database forensics, multimedia forensics, device forensics, and network forensics.

For the purpose of this study, computer forensics is divided into server forensics, laptop forensics and desktop forensics, while software forensics focuses on application software forensics; operating system forensics (open source and proprietary) and forensic tools analysis (open source and proprietary).

Database forensics concentrates on database contents and/or database metadata, while multimedia forensics is divided into digital image forensics, digital video forensics and digital audio forensics. Device forensics is divided into peripheral device forensics, network-enabled device forensics, storage device forensics, large-scale device forensics, small-scale device forensics and obscure device forensics. Finally, the ontology concludes with network forensics, which is divided into cloud forensics, telecom network forensics, internet forensics and wireless forensics.

In the sub-sections that follow the digital forensics disciplines and sub-disciplines, as identified in the ontology in Figure 1, are explained in more detail.

Computer Forensics

According to Crouch (14), computer forensics is a branch of digital forensics that uses analysis techniques to gather potential evidence from desktops, laptops and server computers for investigating suspected illegal or unauthorised activities. More precisely, computer forensics focuses on finding potential digital evidence after a computer security incident has occurred (15). Note that we refer to 'potential' evidence throughout the paper, since digital artefacts are only considered to be 'evidence' in one of the final phases of the digital forensic investigation process, namely the reporting phase. This also implies that, for the collected potential evidence to be considered as competent evidence (50), it must possess scientific validity grounded in scientific methods and procedures. The potential evidence gathered in most cases is usually found stored on the computers' internal storage unit (see Figure 2), which includes the hard disk that also stores operating system data (e.g. log files) and application/user data (e.g. word processor files). Computer forensics also considers the value of data that may be lost by powering down a computer, and thus collection of potential evidence can be conducted while the system is still running e.g. from the Random Access Memory (RAM) or registers.

The goal of computer forensics is to perform a structured investigation while maintaining a documented chain of evidence that can withstand the legal scrutiny of a court of law, whether for a criminal or civil proceeding (14). For the purpose of this paper the areas covered under computer forensics include server forensics, laptop forensics and desktop forensics (see Figure 2 below).

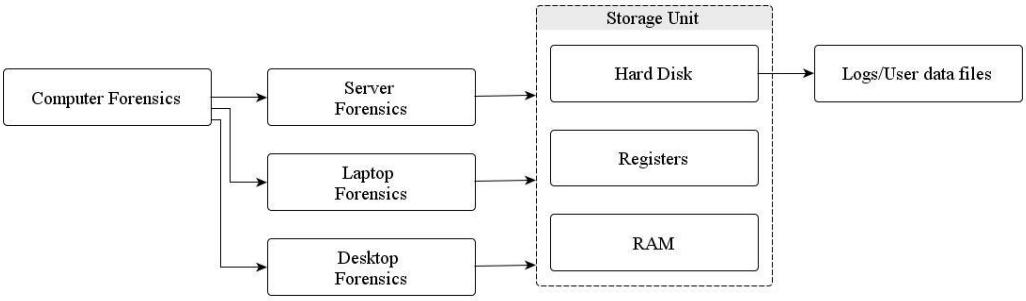


Figure 2: Computer forensics

Server Forensics

In a network environment a server is usually that powerful computer that is dedicated to managing mass system and user resources. Server forensics, therefore, focuses on finding digital evidence that is stored within the server machine (16). In essence, server forensics deals with finding potential evidence in the same way that potential evidence is found on a desktop or laptop computer, the only difference being the significantly larger storage and somewhat different access capabilities to be dealt with on a server computer.

Laptop Forensics
Laptop forensics is dedicated to finding digital evidence from laptop computers. Laptops are designed to be light and mobile. Because of their mobile nature, laptops are popular computing systems and high contenders for hosting potential evidence. The hardware in a laptop is typically custom built for that particular model. According to Pierce (17), very few components follow any given industry standard. This issue particularly complicates the process of digital forensic analysis on laptops and should be handled by a specialist who understands its configuration. However, laptop forensics still form part of computer forensics.

Desktop Forensics
Desktop forensics is meant to find digital evidence from desktop computers once a security incident has occurred. Since there are so many different ways to classify computers (8), the ones discussed above (server, laptop and desktop) serve as examples to facilitate this study. With the advancement in technology it should sooner or later be necessary to add other items to this category.

Software Forensics
Software forensics is a discipline concerned with uncovering potential evidence through examining software. However, according to MacDonell et al (18), software forensics is also a research field that attempts to investigate aspects of computer program authorship by treating pieces of program source code as linguistically and stylistically analysable entities. Software forensics can be used, for example, to detect plagiarism in an academic setting where students' assignments can be compared to see if some are "suspiciously similar" (18, 19).

According to Hanks et al (44), incidents and accidents that can be attributed to software failure often result in tragedies and other losses. The need to learn from these events turns out to be more critical as software systems become more complex and the ways they can fail become less intuitive (44). Moreover, according to Johnson (45, 46), existing software development methods do not provide clear access to retrospective information about the complex and systemic causes of incidents and accidents. In addition, what is known from forensic engineering generally, as well as the study of failure, has yet to be applied comprehensively to software (46). Software forensics (also known as software forensic engineering) can therefore be used to address such deficiencies.

A vast number of computer programs (software) are available on the software market today. However, for the purpose of this paper, the authors considered only a few. For that reason, the reader is advised to consider other software as well, especially when developing curricula and/or education materials. The list of software used in this study serves only as examples and, hence, should not be perceived as an exhaustive list. For the purpose of this paper, software forensics covers operating system forensics, application software forensics and digital forensic analysis tools (as shown in Figure 3).

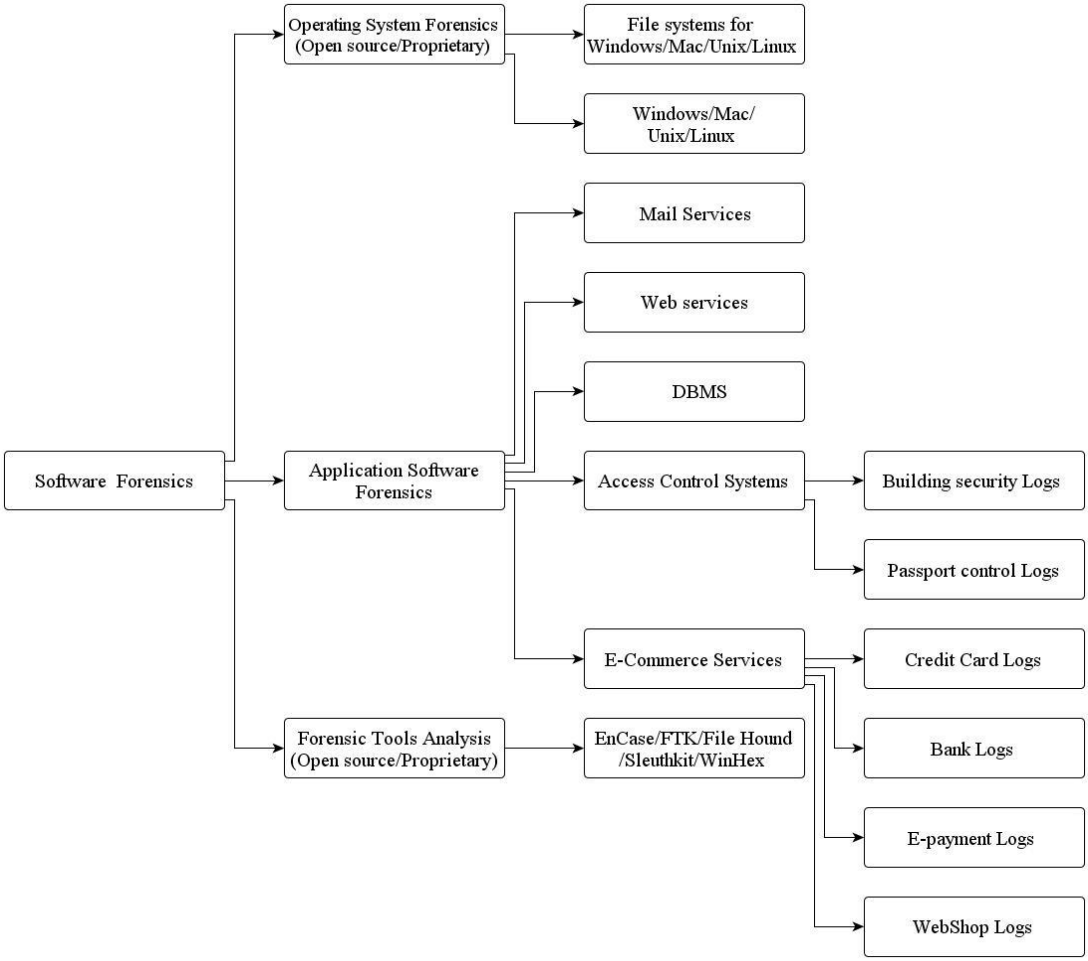


Figure 3: Software forensics

Operating System Forensics

The operating system serves as the primary software installed on any computer system and is often perceived as part and parcel of the entire computer system. Therefore, in the case of a digital investigation, the investigator should be aware of the fact that many different operating systems are available, each with its own associated file structures. By knowing in advance what particular operating system needs to be dealt with, the investigator is able to search for and locate any potential digital evidence more effectively (8).

In addition, operating systems may be categorised as open source or proprietary. Among the common and well-known operating systems are Windows, Mac, Unix and Linux, and an investigator should be acquainted with these operating systems and their different file systems in particular.

Application Software Forensics

Application software is basically designed to help end users perform specific tasks. They either come bundled together with the computer system or can be purchased separately and installed later on the system. Application software forensics focuses on analysing and retrieving potential evidence from application software such as email services, access control systems (e.g. building security logs and passport control logs), web services, database management systems, and E-commerce services (e.g. credit card logs, bank logs, e-payment logs and web shop logs) as shown in Figure 3.

Forensic Tools Analysis

There are many different open-source and proprietary digital forensic tools available for use during digital investigations. Some of the commonly known DF tools used include Encase (51), Forensic Toolkit (FTK) (52) and Sleuth kit (53). These tools are designed to perform a collection of digital forensic investigation

functions and would basically include most of the investigation techniques applied during a digital investigation process. However, there exist other digital forensic investigation tools that perform more elementary investigation functions such as WinHex, which is essentially a universal hexadecimal editor. Such a utility is particularly helpful in viewing any data in its raw form in order to perform low-level data analysis. X-Ways Imager is yet another example of such an elementary tool, which is basically a forensic disk imaging tool only (54).

Database Forensics

Database forensics, as explained by Olivier (21) and Weippl (22), focuses on databases and their related content and/or metadata. Most business' critical and sensitive information is usually recorded and stored in databases, e.g. bank accounts and medical data. Unlawful disclosure, modification and/or theft of such data can be harmful to organisations. Therefore, database forensics aims at investigating unlawful disclosure, modification and/or theft of data within a database in a bid to track down any perpetrators with such malicious intent (22, 23). An investigator's understanding of database concepts and how to use database management systems (DBMS) is clearly of crucial importance to database forensics (see Figure 4).

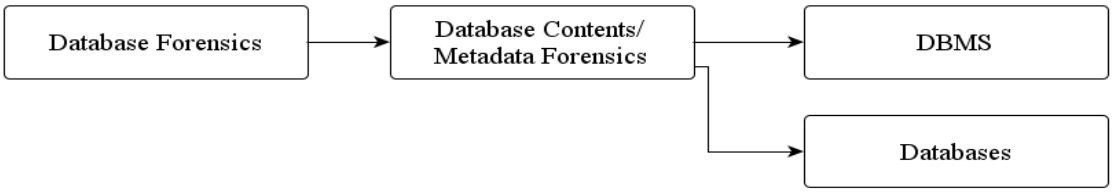


Figure 4: Database forensics

Multimedia Forensics

In today's digital age, the creation and manipulation of digital images, videos and audio have been simplified through digital processing tools that are easily and widely available (24). Such tools may include, but are not limited to, Adobe Photoshop CS6 (47), Adobe Premiere Pro CS6 (48) and Pinnacle Studio (49). Adobe Photoshop CS6 is mostly used for picture and photo editing, while Adobe Premiere Pro and Pinnacle Studio are typically used for video editing. This implies that the authenticity of images, videos and audio can no longer be taken for granted (24). According to Böhme et al (25), questions regarding media authenticity are of growing relevance and of particular interest in court, where consequential decisions might be based on evidence in the form of digital media. Multimedia forensics can be used to uncover the authenticity information of captured images, videos and audio files. Such information can also serve as potential evidence to be presented in a court of law or in civil proceedings. The main areas covered by multimedia forensics in this paper (as shown in Figure 5) include image forensics, video forensics and audio forensics. They are explained briefly in the sub-sections that follow.

Figure 5: Multimedia forensics

Digital Image Forensics

Digital image forensics is concerned with uncovering potential digital evidence found within digital images (24). This may include digital evidence such as image origin (often referred to as image file type identification), image source identification and image forgery detection (26). Digital image forensics can, thus, also be used to verify the authenticity of images (27, 28).

Digital Video Forensics

Digital video forensics, like digital image forensics, is concerned with uncovering potential digital evidence found within video files. With the advent of high-quality digital video cameras and sophisticated video-

editing software, it is becoming increasingly easier to tamper with digital video (29). Digital video forensics can be used to good effect to detect cloning or duplicating frames, or even parts of a frame when people or objects have been removed from a video (29, 30, 31).

Digital Audio Forensics

Digital audio forensics may be defined as the application of audio science and technology in a bid to investigate and establish facts in criminal or civil courts of law. Digital audio forensics is meant to uncover potential digital evidence about audio files. This may include, for example, environment recognition from digital audio files (32). Environment recognition refers to the physical environment under which digital audio samples were recorded. Audio forensics can also be used to determine what kind of microphones were used (33).

Device Forensics

Device forensics is a branch of digital forensics that deals with the gathering of digital evidence from different types of devices. Devices may range from small-scale devices such as mobile phones, Personal Digital Assistants (PDAs), printers, scanners, cameras, fax machines (34) etc., to large-scale devices such as the SAN (Storage Area Network) and NAS (Network Attached Storage) systems. The number of devices in this discipline of digital forensics is increasing daily and hence, in the authors' opinion, is the motivation why device forensics can be considered a separate and vast discipline of the digital forensics domain. For the purpose of this ontology, device forensics is divided into peripheral devices, network-enabled devices, storage devices, large-scale devices, small-scale devices, and obscure devices (see Figure 6). This list should not be considered as exhaustive as most new digital devices could well be categorised within this discipline of the digital forensic ontology.

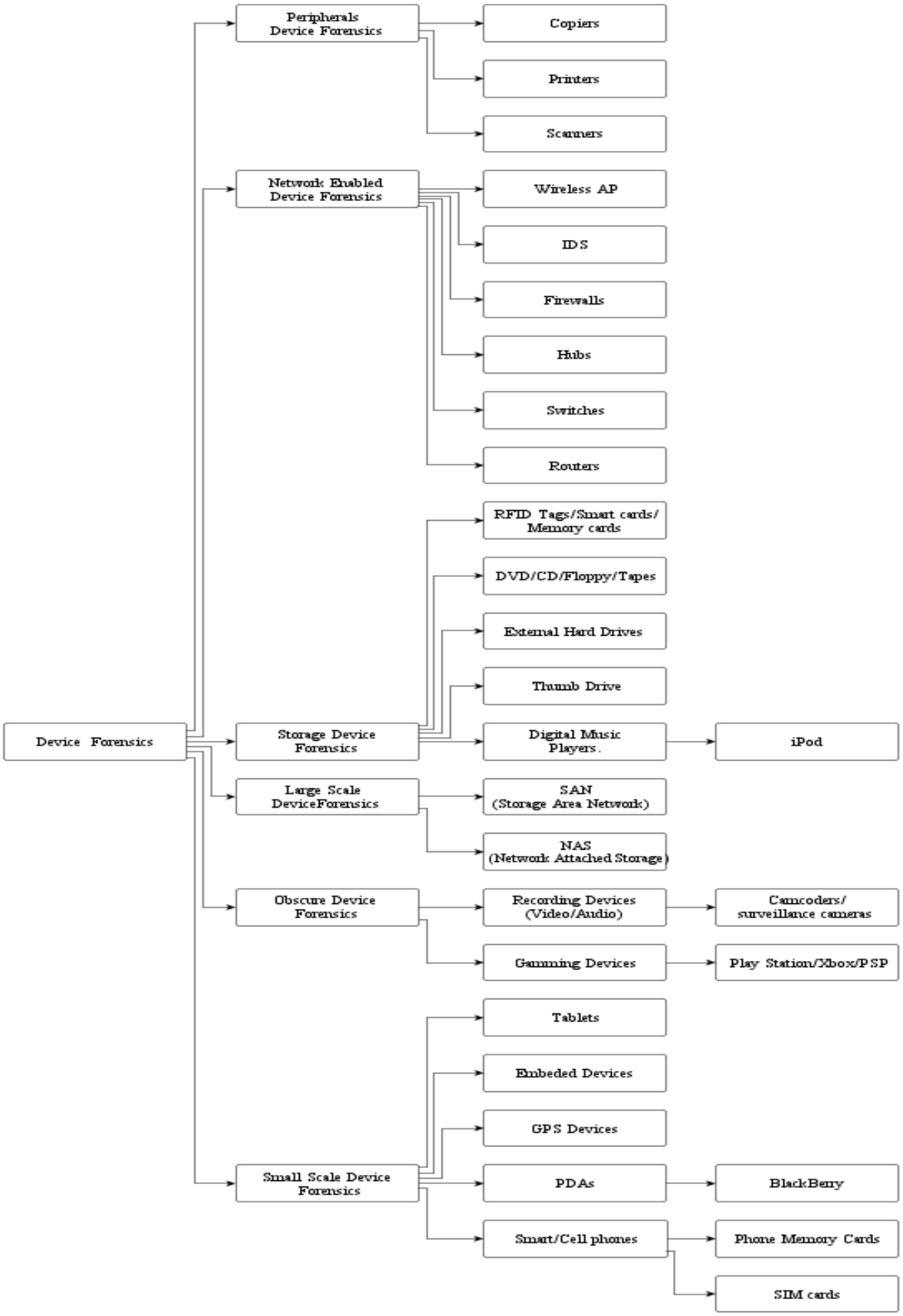


Figure 6: Device forensics

Peripheral devices
Peripheral devices are normally used to expand a system's capabilities; however, they do not actually form part of the core computer architecture. In addition, peripheral devices vary greatly and can range from external to internal peripherals. For example, external peripherals may include a mouse, keyboard, printer, monitor and scanner, among many others. Examples of internal peripheral devices (often referred to as integrated peripherals) may include devices such as a CD-ROM drive and internal modems. A thorough analysis of peripheral devices can reveal much information that is of potential value to a digital forensic investigator.

Network-enabled device forensics
With the development of network and telecommunication technologies, communication infrastructure has rapidly spread in many sectors of the industry. As a result, various network-enabled devices with Ethernet and Transmission Control Protocol/Internet Protocol (TCP/IP) communication functions can be found in different practical applications (35). Such devices may include Intrusion Detection Systems (IDSs), firewalls, hubs, switches, routers and wireless access points (to mention a few). Some of the network-enabled devices have the ability to store data and information and therefore such information can serve as potential evidence during an investigation.

Storage Device Forensics
A storage device is any hardware device that has been specifically designed to store data and information. Storage devices can be primary to a computer (e.g. the RAM) or they can be secondary (e.g. DVD, CD, Tapes, Radio-Frequency Identification (RFID) tags, smart cards, memory cards (flash drives) and external hard drives). Such devices can contain valuable potential evidence in the case of an investigation. Hence, an investigator should be aware of the different capabilities supported by different storage devices.

Large-scale Device Forensics
Nowadays, investigators and analysts increasingly have to deal with large (terabyte-sized) data sets when conducting digital investigations (36). Such large data sets are mostly found stored in large-scale devices such as the SAN (Storage Area Network) and NAS (Network Attached Storage) systems. With the evolution in large-scale storage systems technology, it is possible that petabyte storage will soon replace terabyte-sized devices (43). Petabyte-sized storage is considered the newest frontier in the ever-growing world of data storage devices (43). Therefore, an investigator needs to know how these devices operate in order to be able to effectively gather potential digital evidence. Like any other device, large-scale devices can provide potential evidence that can be presented in a court of law or in civil proceedings.

Small-scale Device Forensics
Small-scale devices, as the name suggests, are small and versatile. In addition, the proliferation of hand-held digital devices has captured the majority of the market and is primed to become the next frontier in technology (9). Therefore, a clear understanding of how these devices operate is necessary to adequately preserve, identify, and extract useful information during a digital forensic investigation (8). Examples of small-scale devices include, but are not limited to, tablets, embedded devices, Global Positioning System (GPS) devices, Personal Digital Assistants (PDAs), mobile (smart) phones, etc. Mobile phones, for example, are becoming a focus of attraction in digital forensic investigations due to the feature-rich versatility of these devices. When dealing with mobile phone device forensics, the two main artefacts of interest that may contain potential evidence are SIM (Subscriber Identity Module) cards and memory cards, of which the latter may be built in (on-board).

Obscure Device Forensics
Obscure devices are those devices that, in the opinion of the authors, cannot be classified under any of the other sub-disciplines of device forensics. Such devices have the ability to store data or information that may possess evidentiary value in a digital forensic investigation. Examples of obscure devices may include digital recording devices (video and audio) such as camcorders, surveillance cameras, gaming devices e.g., (Sony's Play Stations, Microsoft's Xboxes, Nintendo's Wii consoles, etc.), which can also be analysed for potential evidence.

Network Forensics

According to Palmer (20), network forensics "is a branch of digital forensics that basically uses scientific proven techniques to collect, use, identify, examine, correlate, analyse, and document digital evidence from multiple, actively processing and transmitting digital sources for the purpose of uncovering facts related to the planned intent, or measured success of unauthorized activities meant to disrupt, corrupt, and/or compromise system components as well as providing information to assist in response to or recovery from these activities". Unlike other branches of digital forensics, network forensics deals with volatile and dynamic information that can easily get lost after transmission in any network environment. An attacker might be able to erase all log files on a compromised host and therefore network-based evidence may be the only evidence available for forensic analysis (37). For the purpose of this study, network forensics (as shown in Figure 7) is divided into cloud forensics, telecom network forensics, internet forensics and wireless forensics.

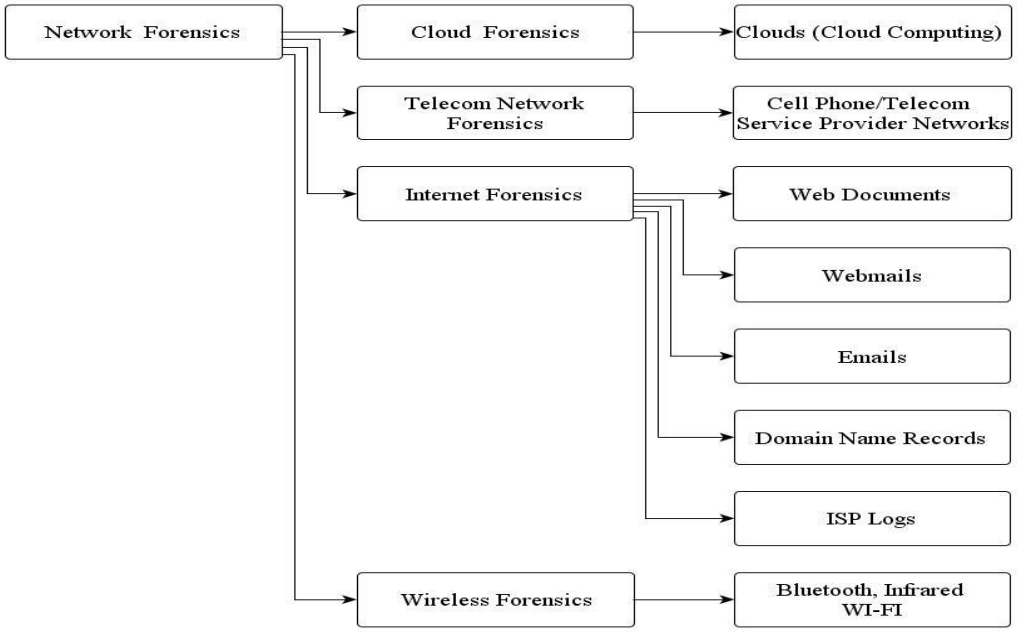


Figure 7: Network forensics

Cloud Forensics

Cloud computing is reckoned to be one of the most transformative technologies in the history of computing. This is so because it is radically changing the way in which information technology services are created, delivered, accessed and managed (41). Cloud forensics, as defined by Keyun et al (41), is an emerging field that deals with the application of digital forensics techniques in cloud computing environments and is a subset of network forensics. Therefore, technically, cloud forensics follows most of the main phases of network forensic processes with extended or novel techniques tailored for cloud computing environments in each phase. For this reason, the authors placed cloud forensics as a sub-discipline of network forensics in the ontology.

Telecom Network Forensics

Telephones are often used to facilitate criminal and terrorist acts. The signalling core of public telephone networks generates valuable data about phone calls and calling patterns that may be used in criminal investigations, especially with the widespread uptake in voice-over-IP (VoIP) systems. However, much of this data is not maintained by service providers and is, therefore, unavailable to law enforcement agencies (38). If such data can be collected and stored, it can be analysed forensically and greatly facilitate the prosecution of criminals in a court of law.

Internet Forensics

With the evolution in global commerce, many business organisations store vital business information online and/or carry out business transactions over the internet. Such organisations are under constant threat of falling victim to internet attacks. Moreover, because the internet is so large and unregulated, it has become a fertile

breeding ground for all kinds of cyber-crimes (42). If the internet is to become a safe platform for transacting business, internet forensics has to become very important as well.

Internet forensics is a research field that deals with the analysis of activities that occurred on the internet. It aims to uncover clues about people and computers involved in internet crime, most notably fraud (e.g. credit card fraud) and identity theft (39). Note that the term "internet crime" and "cyber-crime" are often used interchangeably (55). Cyber-crime is usually used to mean any criminal activity in which a computer or network is the source, tool, target or place of crime (56, 57). The Cambridge English Dictionary defines cyber-crimes as crimes committed with the use of computers or relating to computers, especially through the internet (56).

Therefore, internet forensics tries to uncover the origins, contents, patterns and transmission paths of e-mail and Web pages, as well as browser history and Web servers' scripts and header messages (39). It can also be used to extract information that lies hidden in every email message, web page and web server. Such information may contain potential digital evidence that can be analysed for forensic purposes. In this paper, the authors listed the following areas under internet forensics as common examples: Web-mail, E-mail, domain name records, Internet Service Provider (ISP) logs and web documents. However, there is much more that can be gathered from the internet as compared to what is listed in here.

Wireless Forensics

The adoption of wireless technologies by different organisations in recent years has created issues of concern such as control and security. Incident handlers and law enforcement have been forced to deal with the complexity associated with wireless technologies when managing and responding to security incidents (40). Therefore, wireless forensics, which has emerged as a result of wireless technologies, focuses on capturing and/or collecting digital evidence data that propagates over a wireless network medium. In addition, wireless forensics tries to make sense of the collected digital evidence in a forensic capacity so that it can be presented as valid digital evidence in a court of law. The evidence collected can correspond to plain data, but can include voice conversations as well (40).

Discussion

The ontology presented in this paper is a new contribution in the DF domain. The scope of the ontology is defined by the DF disciplines (refer to Figure 1). The main disciplines as defined in the ontology are computer forensics, software forensics, database forensics, multimedia forensics, device forensics and network forensics. These disciplines are further defined in terms of their scope and functions. The sub-disciplines, objects and sub-objects identified in the ontology include examples and specific finer details covered under the major disciplines. It should also be noted that most of the objects and sub-objects identified in the ontology were selected as common examples to facilitate this study. To the best of the authors' knowledge, there exists no other work of this kind in the domain of digital forensics; therefore, this is a novel contribution towards advancing the digital forensics domain.

In addition, the ontology presented in this paper can be used in the digital forensics domain, for example to address issues such as professional specialisation and certification, as well as the development of digital forensics tools, curricula and education materials.

For the case of professional specialisation, the DF disciplines and sub-disciplines presented in the ontology can be used to give direction to individuals interested in specific areas of specialisation. Such areas will, for example, produce specialists in computer forensics, software forensics, database forensics, multimedia forensics, device forensics and network forensics. While specialisation is important, certification cannot be ignored, especially not by individuals interested in the industry practices of digital forensics. Therefore, a combination of the DF sub-disciplines, objects and sub-objects identified in the ontology should be considered for certification. This will include certification as a certified wireless forensics examiner and/or investigator, certified internet forensics examiner and certified cloud forensics examiner.

Developers of digital forensics tools can use the ontology to fine-tune digital forensic tools so as to be able to cover as many sub-disciplines, objects and sub-objects as possible in the case of digital forensic investigations. This also implies that developers will find the ontology in this paper useful, especially when

considering new digital forensic techniques for specific areas of interest and new high-tech digital forensic investigation tools.

Finally, institutions of higher learning will also find the ontology in this paper constructive, especially when developing curriculums and education materials for different undergraduate and postgraduate studies. Different modules can be developed with the help of the ontology to assist students in comprehending the concepts of digital forensics less effortlessly. Prerequisites for modules can, in addition, be designed effectively with the help of the ontology so as to avoid conflicts among and redundancy of concepts. In fact, the presentation of the ontology in this paper is a whole new contribution towards advancing the digital forensics domain.

Conclusions

Digital forensics plays a very important part in both incident detection and digital investigations. Therefore, developing methodologies that can be used to offer direction in areas such as professional specialisation and certification, as well as the development of forensic tools, curricula and education materials is of utmost importance. This will help, for example, to build a foundation that can be used to solve both present and future problems arising as a result of technological change or domain evolution. Such problems may include those related to the structure of information among different DF disciplines, as well as the reuse and sharing of common domain knowledge. However, more emphasis needs to be placed on digital forensic areas that focus on preparing individuals for what they are expected to do in the case of an investigation process and on preparing them for how to accomplish their task.

This paper presented a novel contribution in the digital forensics domain by means of a guiding ontological model that indicates the placement of the different digital forensic disciplines and sub-disciplines within the domain. The ontology also allows for the addition of new digital forensic disciplines and sub-disciplines, including potential modifications in any one of the aforementioned categories.

Considering the current technological trends, more research needs to be conducted in future in order to expound on the ontology. Further research in the area of digital forensic ontologies must also be conducted to establish the various relationships that exist among the different disciplines and sub-disciplines, objects and sub-objects presented in this study, as some of the examples listed in the ontology might not be mutually exclusive to a particular discipline.

Acknowledgements

The authors wish to thank the members of the Information and Computer Security Architecture (ICSA) research group, Department of Computer Science, University of Pretoria and Kabarak University, for their support throughout the process of writing this paper.

References

1. Reinout Van Rees, Clarity in the usage of the terms ontology, taxonomy and classification. Construction Informatics Digital Library http://itc.scix.net/ paper w78-2003-432.content. (Accessed August 29, 2012).
2. Grüber, T., A translation approach to portable ontology specification. Knowledge Acquisition 5(2) (1993) 199-220.
3. Smith, B., Kusnierczyk, W., Schober, D., Ceusters, W., Towards a reference terminology for ontology research and development in the biomedical domain. In Proceedings of the 2nd Int. Workshop on Formal Biomedical Knowledge Representation: \Biomedical Ontology in Action". (2006) 57-66.
4. Verónica Castañeda, Luciana Ballejos, Ma. Laura Caliusco, Ma. Rosa Galli., The use of ontologies in requirements engineering. Global Journal of Researches in Engineering Vol. 10 Issue 6 (Ver 1.0) November 2010. *GJRE Classification (FOR) 091599*
5. N. Shadbolt, W.H., Berners-Lee, T.: The semantic web revisited. IEEE Intelligent Systems 21(3) (2006) 96-101.
6. Natalya, F. and Deborah, L., Ontology Development 101: A Guide to Creating Your First Ontology.
7. Glen D, Peter S, Revisiting Ontology-Based Requirements Engineering in the age of the Semantic Web.

8. Ashley Brinson, Abigail Robinson, Marcus Rogers, A cyber forensics ontology: Creating a new approach to studying cyber forensics. (2006 DFRWS) Digital investigation 3S (2006) S37–S43. Published by Elsevier Ltd
9. David C.H. and Richard P.M, A Small Scale Digital Device Forensics ontology. Small Scale Digital Device Forensics Journal, Vol.1, No.1, June 2007
10. Allyson M.H and Doris L.C, Weaving Ontologies to Support Digital Forensic Analysis. ISI 2009, June 8-11, 2009, Richardson, TX, USA
11. Ziga Turk, Construction informatics: Definition and ontology. Advanced Engineering Informatics 20 (2006) 187–199
12. Graciela Brusa, Ma. Laura Caliusco and Omar Chiotti, A Process for Building a Domain Ontology: an Experience in Developing a Government Budgetary Ontology. Australasian Ontology Workshop (AOW 2006), Hobart, Australia.
13. Boyce, S., & Pahl, C. (2007). Developing Domain Ontologies for Course Content. *Educational Technology & Society*, 10 (3), 275-288.
14. Jim Ed Crouch, NSCI December 16, 2010, An Introduction to Computer Forensics. Available at: http://www.nsci-va.org/WhitePapers/2010-12-16-Computer%20Forensics-Crouch-final.pdf (Accessed March 5, 2012).
15. Computer forensics. Anglia Ruskin University, Dissertation No (CSH2998A). Available at: http://www.minshawi.com/other/computer%20foransics.pdf (Accessed March 5, 2012).
16. Roberto Obialero, SANS Institute 2000 - 2005. Forensic Analysis of a Compromised Intranet Server.
17. Matt Pierce, SANS Institute 2003. Detailed Forensic Procedure for Laptop computers Forensic analysis 06-11-2003.
18. Stephen G. MacDonell, Andrew R. Gray, Grant MacLennan, and Philip Sallis, Software Forensics for Discriminating between Program Authors using Case-Based Reasoning, Feed-Forward Neural Networks and Multiple Discriminant Analysis.
19. G. Whale. Software metrics and plagiarism detection. *Journal of Systems and Software*, 13:131–138, 1990.
20. Gary Palmer, A Road Map for Digital Forensic Research. DFRWS Technical Report. DTR - T001-01 Final. Report from the First Digital Forensic Research Workshop (DFRWS). November 6th, 2001 - Final.
21. Martin S. Olivier, On metadata context in Database Forensics. ICSA Research Group.
22. Edgar Weippl, Database Forensics. Available at: http://www.nii.ac.jp/issi/pdf/2/4Johannes_Heurix.pdf (Accessed March 26, 2012).
23. Mario A.M et al, Database Forensics. Available at: http://delivery.acm.org/10.1145/1950000/1940958/p62-guimaraes.pdf?ip=137.215.6.53&acc=ACTIVE%20SERVICE&CFID=74282261&CFTOKEN=62192917&__acm__=1332837302_057fa5962ff148a2ab5a9daccbec521b (Accessed March 27, 2012).
24. Multimedia Forensics, 2012. URL http://isis.poly.edu/projects/forensics. (Accessed August 03, 2012).
25. Rainer Böhme, Felix C. Freiling, Thomas Gloe, and Matthias Kirchner, Multimedia Forensics Is Not Computer Forensics
26. Ashwin S., Min Wu and K. J. Ray Liu, Image Tampering Identification Using Blind Deconvolution
27. Thomas G. et al, Can We Trust Digital Image Forensics? MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany. Copyright 2007 ACM 978-1-59593-701-8/07/0009
28. Ashwin S. et al, Digital Image Forensics via Intrinsic Fingerprints. IEEE Transactions On Information Forensics And Security, Vol. 3, No. 1, March 2008
29. Weihong W. and Hany F., Exposing Digital Forgeries in Video by Detecting Duplication. MM&Sec'07, September 20–21, 2007, Dallas, Texas, USA. Copyright 2007 ACM 9781595938572/07/0009
30. Frédéric L. et al, Image And Video Fingerprinting: Forensic Applications
31. Matthew C. and K. J. Ray Liu, Anti-Forensics For Frame Deletion/Addition In Mpeg Video
32. Ghulam M. and Khalid A., Environment Recognition For Digital Audio Forensics Using Mpeg–7 Andmel Cepstral Features. Journal Of Electrical Engineering, Vol. 62, No. 4, 2011, 199–205
33. Christian K. et al, Digital Audio Forensics: A First Practical Evaluation on Microphone and Environment Classification. MM&Sec'07, September 20–21, 2007, Dallas, Texas, USA. Copyright 2007 ACM 978-1-59593-857-2/07/0009

34. Cyber Forensics - Device Forensics. Available at: http://www.cyberforensics.in/(A(cos8NMWQywEkAAAAODMwODM4YWMtNWFmZC00ZWNhLThkNDEtNTlhMWM3MGE5MzA5hkCziwldj9ts_CCtkjYQI68akds1))/Research/DeviceForensics.aspx?AspxAutoDetectCookieSupport=1 (Accessed March 22, 2012).
35. Network-enabled Devices, 2012 . URL http://www.sena.com/solutions/network_enabling/ (Accessed September 5, 2012).
36. Hyungkeun J., High Speed Search for Large-Scale Digital Forensic Investigation. E-Forensics 2008. Adelaide, Australia.
37. Erik Hjelmvik, Passive Network Security Analysis with Network Miner | ForensicFocus.com. Available at: http://www.forensicfocus.com/passive-network-security-analysis-networkminer (Accessed March 27, 2012).
38. T. Moore et al, Using Signaling Information in Telecom Network Forensics. IFIP International Federation for Information Processing, 2005, Volume 194/2005.
39. Internet forensics Definition from PC Magazine Encyclopedia. Available at: http://www.pcmag.com/encyclopedia_term/0,2542,t=Internet+forensics&i=59910,00.asp (Accessed March 22, 2012).
40. Raul Siles, GSE, Wireless Forensics: Tapping the Air - Part One | Symantec Connect Community. Available at: http://www.symantec.com/connect/articles/wireless-forensics-tapping-air-part-one (Accessed March 26, 2012).
41. Keyun et al, Cloud forensics: An overview. Centre for Cybercrime Investigation, University College Dublin.
42. Robert Jones, Internet Forensics, Using Digital Evidence to Solve Computer Crime Publisher: O'Reilly Media October 2005
43. Anon, 2012. Petabyte of Storage Capacity. URL http://www.aberdeeninc.com/abcatg/petabyte-storage.htm (Accessed August 26, 2012).
44. Kimberly S. Hanks, John C. Knight and C. Michael Holloway, The Role of Natural Language in Accident Investigation and Reporting Guidelines
45. Chris Johnson, Forensic software engineering. Proceedings of 19th International Conference SAFECOMP 2000, 420-430.
46. Chris Johnson, Forensic software engineering: are software failures symptomatic of systemic problems? Safety Science 40 (2002) 835–847
47. Anon, Image editor software | Adobe Photoshop CS6. Available at: http://www.adobe.com/products/photoshop.html (Accessed August 26, 2012).
48. Anon, Video editing software | Adobe Premiere Pro CS6. Available at: http://www.adobe.com/products/premiere.html (Accessed August 26, 2012).
49. Anon, Video editing software - Pinnacle Studio - The #1 selling digital video editing software. Available at: http://www.pinnaclesys.com/PublicSite/us/Products/Consumer+Products/Home+Video/Studio+Family/ (Accessed August 26, 2012).
50. Daniel J. Ryan and Gal Shpantzer, Legal Aspects of Digital Forensics
51. Guidance Software, EnCase Forensic - Computer Forensic Data Collection for Digital Evidence Examiners. Available at: http://www.guidancesoftware.com/encase-forensic.htm (Accessed August 29, 2012).
52. AccessData, Computer Forensics Software for Digital Investigations. Available at: http://accessdata.com/products/digital-forensics/ftk (Accessed August 29, 2012).
53. The Sleuth Kit (TSK) & Autopsy: Open Source Digital Investigation Tools. Available at: http://www.sleuthkit.org/index.php (Accessed August 29, 2012).
54. X-Ways Software Technology AG, WinHex: Hex Editor & Disk Editor, Computer Forensics & Data Recovery Software. Available at: http://www.winhex.com/winhex/ (Accessed August 29, 2012).
55. Melanie Kowalski, Cyber-Crime: Issues, Data Sources, and Feasibility of Collecting Police-Reported Statistics. Canadian Centre for Justice Statistics, Catalogue no. 85-558-XIE, ISBN 0-660-33200-8.
56. A. Prasanna, Cyber Crimes: Law and Practice
57. Talwant Singh, 2012. CYBER LAW & INFORMATION TECHNOLOGY. URL http://delhicourts.nic.in/ejournals/CYBER%20LAW.pdf (Accessed August 29, 2012).
58. Himal Lalla and Stephen V. Flowerday, Towards a Standardised Digital Forensic Process:E-mail Forensics

59. Soltan Alharbi1, Jens Weber-Jahnke and Issa Traore, The Proactive and Reactive Digital Forensics Investigation Process: A Systematic Literature Review. International Journal of Security and Its Applications. Vol. 5 No. 4, October, 2011
60. 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.
61. Leigland, R., & Krings, A. W. (2004). A Formalization of Digital Forensics. International Journal of Digital Evidence , 3 (2), 1-32.
62. ISO/IEC WD 27043.2, working draft. Information technology — Security techniques — Investigation principles and processes.
63. Jasmine Ćosić and Zoran Ćosić, The Necessity of Developing a Digital Evidence Ontology, Proceedings of the Central European Conference on Information and Intelligent Systems. September 19-21, 2012. Page 325-330

Taxonomy of Challenges for Digital Forensics

[1,2]Nickson M. Karie[*] MSc, [1]Hein S. Venter[†] PhD
[1]ICSA Research Group, Department of Computer Science,
University of Pretoria, Private Bag X20, Hatfield 0028,
Pretoria, South Africa
[2]Department of Computer Science,
Kabarak University, Private Bag - 20157, Kabarak, Kenya
Email: menza06@hotmail.com[*], hventer@cs.up.ac.za[†]

ABSTRACT: Since its inception, over a decade ago, the field of digital forensics has faced numerous challenges. Despite different researchers and digital forensic practitioners having studied and analysed various known digital forensic challenges, as of 2013, there still exists a need for a formal classification of these challenges. This paper, therefore, reviews existing research literature and highlights the various challenges that digital forensics has faced for the last ten years. In conducting this research study, however, it was difficult for the authors to review all the existing research literature in the digital forensic domain, hence, sampling and randomisation techniques were employed to facilitate the review of the gathered literature. Taxonomy of the various challenges is subsequently proposed in this paper based on our review of the literature. The taxonomy classifies the large number of digital forensic challenges into four well-defined and easily understood categories. The proposed taxonomy can be useful, for example, in future developments of automated digital forensic tools by explicitly describing processes and procedures that focus on addressing specific challenges identified in this paper. However, it should also be noted that the purpose of this paper is not to propose any solutions to the individual challenges that digital forensics face, but to serve as a survey of the state of the art of the research area.

KEYWORDS: Forensic sciences, digital forensics, taxonomy, digital forensic challenges, categories, formal classification of challenges

Over the last decade, the evolution in digital technology has greatly influenced the way we live our daily lives and conduct business. Consequently, as this evolution continues, numerous challenges emerge that are to be faced by the digital forensic domain. The particular problem that this paper addresses is stated as follows. Due to the fact that digital forensics (DF) is still considered a relatively new field in both research and industry, the number of challenges faced in this field is bound to increase in line with Moore's Law (1). The simplified version of this law states that processor speeds or overall processing power for computers will double every two years, resulting in numerous other challenges in DF.

This paper therefore aims at reviewing existing digital forensic literature and highlights the various challenges that digital forensics have faced over the last ten years. Taxonomy of the various challenges is subsequently proposed in this paper based on our review of the existing literature. The taxonomy classifies the large number of digital forensic challenges into four well-defined and easily understood categories.

The presentation in this paper can be useful, for example, in future developments of automated digital forensic tools as well as in explicitly describing processes and procedures that focus on addressing the individual digital forensic challenges identified. Institutions of higher learning will also find the proposed taxonomy in this paper constructive, especially when developing curriculums and educational material for different undergraduate courses, as well as research projects for postgraduate studies.

Furthermore, the presentation of the taxonomy in this paper is a novel contribution in the digital forensic domain and offers a comprehendible categorisation that may shed more light on existing digital forensic challenges. The taxonomy has been designed in a way to accommodate new categories of digital forensic challenges that may crop up as a result of technological change and domain evolution.

Background

As mentioned earlier, DF is a new and growing field in both research and industry (2). It is also considered a branch of forensic science that deals with the recovery and investigation of material found in digital devices, often in relation to digital crimes. By 2013, research in digital forensics has been conducted for over a decade. However, because of the ever-evolving nature of digital technology, the challenges faced during the recovery and investigations of materials found in digital devices are obviously increasing as well.

For this reason, rigorous and flexible process models and frameworks need to be developed to overcome the different challenges faced by DF. This includes challenges such as the vast volumes of data (3), education and certification, lack of unified formal representation of domain knowledge, legal system challenges, semantic disparities that occur in the domain among others. Developing practical methodologies that can aid in resolving different challenges in DF is inevitable and is as important as the research itself. Besides, for DF to remain effective and relevant to the law enforcement, academia, and the private sector, the domain experts must constantly endeavour to address these challenges.

Recent developments in digital forensics are geared towards standardising the digital forensic investigation process model (4). This development is backed up by the fact that the number of forensic process models that exist has added to the complexity of the digital forensic field (5), hence the need for harmonisation and/or standardisation. In the next section the authors will examine existing related work on taxonomy development in digital forensics.

Related Work

Several taxonomies and frameworks have been proposed by different researchers in the digital forensic domain. Most of these taxonomies and frameworks, though, have their major focus on the digital forensic investigation process. Nevertheless, the literature in this regard offered valuable contributions towards the development of the taxonomy of challenges for digital forensics, presented in this paper.

To begin with, in a paper by Altschaffel et al. (6), the authors argue that digital forensic investigations are usually conducted to solve crimes committed by perpetrators and/or intruders using IT systems. They then propose a taxonomy that helps to perform a forensic examination and to establish answers to a set of well-defined questions during such examination.

Efforts by Hoefer and Karagiannis (7), culminated in taxonomy of cloud computing services. Their paper describes the available cloud computing services and further proposes a tree-structured taxonomy based on their characteristics, so as to easily classify cloud computing services and make it easier to compare them. In contrast, the proposed taxonomy in this paper, offers a simplified platform that sheds more light on the classification of existing digital forensic challenges.

Strauch et al. (8) argue that cloud computing allows the reduction of capital expenditure by using resources on demand. Thus, they investigate how to build a database layer in the cloud and present pure and hybrid cloud data-hosting solutions. They then organised the solutions in a taxonomy which they use to categorise existing cloud data-hosting solutions. Lupiana et al. (9), on the other hand, proposed a taxonomy for classifying disparate research efforts in ubiquitous computing environments. Their taxonomy classifies ubiquitous computing environments into two major categories namely: interactive environments and smart environments.

Sansurooah (10) explains in his paper that the increased risk and incidences of computer misuse have raised awareness in public and private sectors of the need to develop defensive and offensive responses. He then compares the different methodologies and procedures that are in place for the gathering and acquisition of digital evidence and subsequently defines which model will be the most appropriate taxonomy for the electronic evidence in the computer forensics analysis phase. Sriram (11), however, argues that in recent years the exponential growth of technology has also brought with it some serious challenges for digital forensic research. Therefore, in his paper, he reviews the research literature since 2000 and categorise developments in the field into 4 major categories. He further highlights the observations made by previous researchers and summarise the research directions for the future.

Kara et al. (3) explains that while many fields have well-defined research agendas, evolution of the field of digital forensics has been largely driven by practitioners in the field. Their paper then goes further and

outlines new research categories (taxonomy) and areas identified at the Colloquium for Information Systems Security Education (CISSE-2008), as well as a plan for future development of a formalized research agenda for digital forensics.

Garfinkel (12) in this paper states that, the golden age of computer forensics is quickly coming to an end. He then summarizes current digital forensic research directions and argues that to move forward the community needs to adopt standardised, modular approaches for data representation and digital forensic processing. In addition, he argues that, without a clear research agenda aimed at dramatically improving the efficiency of both digital forensic tools and the research process, our hard-won capabilities will be degraded and eventually lost in the coming years.

Other related research works on taxonomies also exist, but none of those or the cited references in this paper have to date presented a taxonomy of the different challenges faced by the digital forensic domain in the way introduced in this paper.

Thus: in contrast to all the research efforts referred to above, we propose a taxonomy that classifies the various challenges faced by digital forensics into 4 well-defined and easily understood categories. Nevertheless, the authors acknowledge the fact that the previous work on the proposed frameworks and taxonomies has offered us useful insights into the development of the taxonomy of challenges for digital forensics in this paper. The scope of the proposed taxonomy is explained in the section to follow.

Scope of the Proposed Taxonomy

While there are many challenges in digital forensics and several attempts to address specific and/or individual challenges have been done by different researchers. The presentation in this paper is an exceptional effort towards a novel taxonomy of digital forensic challenges based on the review of existing digital forensic literature. The scope of the taxonomy is, however, restricted to the boundaries of the literature surveyed by the authors (not more than ten years old). The authors' also acknowledge that, the various challenges presented in this paper are not, in whatever way, an exhaustive list. This is backed up by the fact that, it is difficult to gain an exhaustive list - because an exhaustive list is hard to create and even if created it would not be easy to handle or manage because of its size.

The taxonomy, hence, has been designed taking into consideration the major challenges that digital forensic has faced over the last decade. The authors, though, did not establish a precise distinction between the old and the most recent digital forensic challenges in this paper. This is because; some of the challenges captured in the taxonomy are inherent to digital forensics, e.g. the vast volumes of data. Future research will, however, consider the possibility of developing an extensive taxonomy with distinctions between the old and the most recent challenges. The next section, thus, explains in detail the proposed taxonomy of challenges for digital forensics in this paper.

The Proposed Taxonomy of Challenges for Digital Forensics

In this section of the paper, we present a detailed explanation of the taxonomy of challenges for digital forensics. Table 1 shows the structure of the proposed taxonomy.

The taxonomy consists of four rows arranged from top to bottom with the first row depicting the technical challenges faced by digital forensics. This is followed by the legal systems and/or law enforcement challenges in the second row, the personnel-related challenges in the third row and finally the operational challenges faced by digital forensics in the fourth row.

Table 1: The Taxonomy of Challenges for Digital Forensics

Categories of DF Challenges	**Identified Sub-Categories**	
Technical Challenges	xiii.	Encryption
	xiv.	Vast Volumes of Data
	xv.	Incompatibility Among Heterogeneous Forensic Tools
	xvi.	Volatility of Digital Evidence
	xvii.	Bandwidth Restrictions
	xviii.	Limited Lifespan of Digital Media
	xix.	Sophistication of Digital Crimes

	xx.	Emerging technologies
	xxi.	Limited Window of Opportunity to Collection of Potential Digital Evidence
	xxii.	The Anti-Forensics
	xxiii.	Acquisition of Information from Small-Scale Technological Devices
	xxiv.	Emerging Cloud Computing or Cloud Forensic Challenges
Legal Systems and/or Law Enforcement Challenges	vii.	Jurisdiction
	viii.	Prosecuting Digital Crimes (Legal Process)
	ix.	Admissibility of Digital Forensic Tools and Techniques
	x.	Insufficient Support for Legal Criminal or Civil Prosecution
	xi.	Ethical Issues
	xii.	Privacy
Personnel-related Challenges	vi.	Lack of Qualified digital forensic personnel (Training, Education and Certification)
	vii.	Semantic Disparities in Digital Forensics
	viii.	Lack of Unified formal Representation of Digital Forensic Domain Knowledge
	ix.	Lack of Forensic Knowledge Reuse among personnel
	x.	Forensic Investigator Licensing Requirements
Operational Challenges	vi.	Incidence detection, response and prevention
	vii.	Lack Of Standardised processes and procedures
	viii.	Significant Manual intervention and Analysis
	ix.	Digital Forensic Readiness Challenge in Organisations
	x.	Trust of Audit Trails

The various sub-categories of the challenges presented in each of the different rows of the taxonomy shown in Table 1, however, focuses more on areas that can, for example, be considered when developing new curriculums and education materials for different undergraduate programmes as well as research projects for postgraduate studies.

The sub-categories can also be useful when developing dynamic digital forensic tools that focus on addressing specific identified digital forensic challenges. Organising the taxonomy into categories and sub-categories was necessary to simplify the understanding of the taxonomy as well as to present specific finer details of the taxonomy.

Note still, from the taxonomy in Table 1, that the sub-categories of the challenges listed in column two were only selected as common examples to facilitate this study and should not be treated as an exhaustive list. Therefore, more specific sub-categories of the challenges to each named category can and should be added as the need arises in future.
The major categories of the various digital forensics challenges explored in this study (with their details as shown in Table 1) include: technical challenges; legal systems and/or law enforcement challenges; personnel-related challenges, and operational challenges.

For the purpose of this study, technical challenges include: encryption; vast volumes of data; incompatibility among heterogeneous forensic analysis tools; volatility of digital evidence; bandwidth restrictions; limited lifespan of digital media; sophistication of the digital crimes; emerging technologies and devices; limited window of opportunity to collect potential digital evidence; anti-forensics; acquisition of information from small-scale technological devices, and lastly the emerging cloud computing or cloud forensic challenges.

Legal systems and/or law enforcement challenges on the other hand focus on jurisdiction; prosecuting digital crimes (legal process); admissibility of digital forensic tools and techniques; insufficient support for legal criminal or civil prosecution; ethical issues, and privacy.
Personnel-related challenges concentrate on, the lack of qualified digital forensic personnel (training, education and certification); semantic disparities in digital forensics; lack of unified formal representation of

digital forensic domain knowledge; lack of forensic knowledge reuse among personnel, and the forensic investigator licensing requirements.

Finally, the taxonomy concludes with operational challenges that include: incidence detection, response and prevention; lack of standardised processes and procedures; significant manual intervention and analysis; digital forensic readiness challenge in organisations, and trust of audit trails.

In the sub-sections to follow the various categories, sub-categories of the challenges faced by digital forensics as identified in Table 1 are explained in more detail.

Technical challenges
Technical challenges can be described as those challenges that can be addressed with existing expertise, protocols and operations. Implementing solutions to any of the identified technical challenges often falls to someone with the authority to do so. Knowing that, digital forensics requires a well-balanced combination of technical skills and ethical conduct; some of the identified technical challenges faced by digital forensics are explained in the sub-sections to follow.

Encryption – With the advances in communication technologies such as the Internet, complex encryption products are now widely and easily accessible, presenting the digital forensic examiner with a significant challenge. Moreover, as encryption standards rise and the algorithms become more complex, it will become more difficult and more time-consuming for specialists to conduct cryptanalysis and then piece together encrypted files into meaningful information (13). Cryptanalysis is described as the science of 'code breaking,' in which an individual reconstructs the original plaintext message from an encrypted version (14) without having a valid decryption key.

There is currently no proven or fully known direct or standardised formula for conducting cryptanalysis. In most cases encrypted data is completely inaccessible without the decryption key. If the suspect refuses to give the key or pleads plausible deniability, the investigator will have to try other methods to acquire the key (15). Although it is now the law in the UK that any encryption key must be given to the police, this is not the case in other jurisdictions, and punishment for not surrendering such keys may be far less severe than the potential punishment for any crime committed (15).

Vast Volumes of data – There has been tremendous growth in the volume of persistent storage – disk storage – used in both personal and corporate systems (16). With the incredibly large volumes of data existing within applications such as Enterprise Resource Planning (ERP) and as mail systems become larger, the volume and amounts of material being generated are by far not human readable in a lifetime – let alone in the scope of a trial or litigation (17). This has implications not only for the procedures and techniques used by investigators for data acquisition and imaging, but also (and more importantly) for the way in which the digital forensic data is analysed.

Incompatibility among Heterogeneous Forensic Analysis Tools - Digital forensic tools generally differ in functionality, complexity and cost. Some tools are designed to serve a single purpose or provide unique information to examiners, while others offer a suite of functions (18). All the same, most of the existing forensic analysis tools consist of dissimilar elements or parts (design and algorithms) and are consequently unable to work together harmoniously. Besides, some of the tools unable to cope with the ever-increasing storage capacity of target devices. This implies that huge targets pose a challenge as they require more sophisticated analysis techniques that allow digital forensic investigators to perform forensic investigations much more efficiently (19) thus easing digital investigations.

Volatility of Digital Evidence - Digital evidence is, by its nature, fragile. Almost any activity performed on a device, whether inadvertently or intentionally (e.g. powering up or shutting down) can alter or destroy potential evidence (20). In addition, loss of battery power in portable devices, changes in magnetic fields, exposure to light, extremes in temperature and even rough handling can cause loss of data. Collecting volatile data therefore presents a serious challenge to digital forensic investigators, because doing so can change the state of the system (and the contents of the memory itself).

Bandwidth Restrictions - According to Taute et al. (21), bandwidth restrictions in networks can limit or slow down the digital evidence acquisition process. Since the suspect machine in any network is live and active, digital forensic investigators need to connect to the forensic agent installed on the machine via a network. Copying the data as potential digital evidence from the suspect machine to the forensic workstation might slow down the bandwidth, especially if there are many users utilising the bandwidth at that particular time. Large remote evidence acquisitions may also have to be done after hours to accommodate smaller bandwidth capacities, thus posing a challenge to investigators.

Limited Lifespan of Digital Media - While digital storage media facilitate storage of and easy access to electronic data, they do not provide long-term archival storage (22). This is because, at the core of every digital storage media lies "bit preservation" and the ability to monitor for "bit loss", hence, any bit deterioration can compromise digital data (23). The life span of some digital storage media is typically short and also well enough known for all to be aware of the risks when using them for preservation purposes (24). This poses a serious storage challenge. In fact, even with the emerging cloud computing, the cloud servers leverage on redundant digital storage media which ensures that, in the event of a hardware failure, the data continues to be accessible from another part of the cloud where it is stored safely.

Sophistication of the Digital Crimes - The increasing sophistication of cyber-crimes poses significant challenges to investigations and digital forensic investigators. According to a report by The Association of Chief Police Officers (ACPO) (25), investigators are routinely faced with the reality of sophisticated data encryption, as well as hacking tools and malicious software that may exist solely within memory. Criminals now use anti-forensic techniques that can require endless digital investigations in the case of an attack (26) making it even harder for investigators to get the much needed evidence.

Emerging Technologies - According to Sheward (27), new and evolving technologies create new digital forensic challenges for investigators. Working with a new file system, for example, or even just a new type of file, can require a change in approach or the development of a new technique. While these changes may require slight alterations to well-defined procedures, it is extremely rare to have to deal with a technology that gives a complete transition.

Limited Window of Opportunity for Collection of Potential Digital Evidence - During the collection of potential digital evidence it is important for digital forensic investigators to prioritise which data must be collected first. This becomes a challenge to investigators especially when they are time constrained or when the window of opportunity to collect the data is small (28) or the time to image a system is too short. Investigators must take the necessary steps to ensure that they are able to collect and preserve critical information during this window of opportunity and analyse the data in a method that maintains its integrity.

The Anti-forensics - According to Garfinkel (29), anti-forensics (AF) is a growing collection of tools and techniques that frustrate forensic tools, investigations and investigators. People use anti-forensics to demonstrate how vulnerable and unreliable computer data can be. In order to use evidence from a computer system in court, the prosecution must authenticate the evidence. This also means that the prosecution must be able to prove that the information presented as potential evidence in fact came from the suspect's computer and that it has remained unaltered. Anti-forensics makes it hard for examiners to detect that some kind of event has taken place and it disrupts the collection of information, thus increasing the time that an examiner needs to spend on a case and casting doubt on a forensic report or testimony (30).

Acquisition of Information from Small-scale Technological Devices - According to Bennett (31), unlike traditional computer forensics on a desktop or laptop computer – where the investigator would simply remove the hard drive, attach it to a write blocker device (thus allowing acquisition of information on a computer hard drive without creating the possibility of accidentally damaging the drive contents) and image the hard drive so as to fully analyse the data – the process to extract information from a mobile device is much more complicated. Moreover, with the continued growth of the mobile device market, the possibility of the use of such devices in criminal activity will continue to increase (32). There are currently numerous manufacturers and models of mobile devices on the market, which results in creating a huge diversity of potential problems and/or challenges to investigators. It becomes extremely difficult for an investigator to choose the proper forensics tools for seizing internal data from mobile devices (32).

Emerging Cloud or Cloud Forensic Challenges - Cloud computing has emerged as an important solution offering organisations a potentially cost effective model to ease their computing needs and accomplish business objectives. However, mixed in with the cloud cost effective opportunities are numerous challenges that need to be considered such as jurisdiction and cloud heterogeneity (33), prior to committing to a cloud service. According to Leslie et al. (34), other challenges faced by the cloud include: safeguarding data security, managing the contractual relationship, dealing with lock-in and managing the cloud. Numerous security challenges also exists e.g. data protection, user authentication, and data breach contingency planning that also need to be addressed.

Legal Systems and/or Law Enforcement Challenges

There is an increased awareness in the legal community of the need for digital forensic services to obtain successful prosecutions that could otherwise fail because of unsatisfactory equipment, procedures or presentation in court (35). Therefore, in the sub-sections to follow, we examine some of the legal systems and/or law enforcement challenges faced by digital forensics.

Jurisdiction - The increasing popularity of cloud computing has made conventional crime detection even more difficult. The very strengths of cloud computing, which allows anyone anywhere in the world to use publicly accessible software to process data stored in a virtual cyber-space location, could be put to devious use by criminals to store incriminating data on a server located beyond the jurisdiction of the courts of their country of residence, preferably in a State with no judicial cooperation treaty with that country (36). This makes court jurisdiction a challenge during prosecution.

Prosecuting Digital Crimes (Legal Process)- According to Lauren (37), prosecuting cyber-crime is no easy task, despite disparate laws. Even with today's forensic capabilities, legal inadequacies in various jurisdictions (not to mention uneven law enforcement and legal processes) make prosecution a very challenging task. This has created the need for new legislation that allows for digital evidence to be presented in any court of law or civil proceedings (38), as well as for the prosecution of digital crimes.

Current digital forensic investigations are based on the existing legal system or legal processes and supporting laws available. The infrastructure to investigate digital crimes is based on the prevailing cyber-laws, which makes it difficult to adopt specific digital forensic models to carry out digital investigations and prepare court admissible reports (38). Many digital forensic practitioners simply follow technical procedures and forget about the actual purpose and core concept of digital forensic investigation (39).

Admissibility of Digital Forensic Tools and Techniques - Given the enormous volumes of data currently handled by digital forensic investigators, the admissibility of digital forensic tools and techniques used to collect and analyse data is becoming a challenge. As with all other forensic disciplines, digital forensic techniques and tools must meet basic evidentiary and scientific standards to be allowed as evidence in legal proceedings (40). This also means that, the tools, techniques, processes and procedures should be capable of being proven correct through empirical testing. In the context of digital forensics, this means that the tools, techniques, processes and procedures used in the collection and analysis of digital evidence data must be validated and proven to meet scientific standards if the results from such applications are to be acceptable as potential evidence in criminal cases.

Insufficient Support for Legal Criminal or Civil Prosecution - According to Mercuri (41), digital forensic techniques may be unfairly applied in order to tip the scales of justice in the direction of prosecution. Burgess (42) also states that, in the field of digital forensics (as in the field of law) procedures in civil cases differ somewhat from those in criminal cases. The collection of data and presentation of evidence may be held to different standards, the process of data collection and imaging can be quite different, and the consequences of the case may have very different impacts.

Ethical Issues - According to Bassett et al. (35), there are many ethical dilemmas with which investigators must be prepared to face during an investigation. One of the most common ethical concerns is managing the discovery of confidential data that is irrelevant to the case at hand. The question of what to do with irrelevant information arises. The general code of ethics to follow is that such information must be ignored because it is not relevant to the investigation. However, it is not always easy to ignore such information and any secrets that may be uncovered can weigh heavily on the mind of the investigator. Other ethical concerns may include:

acknowledgement of errors by investigators on evidence data; bias during an investigation; maintaining control and responsibility for forensics equipment (35).

Privacy - Privacy issues usually arise in the case of an investigation. Privacy is very important to any organisation or victim. Though, in special cases the investigator may be required to share the data or compromise the client's privacy to get to the truth. It is possible that the victim organisation may lose trust in the forensic team if, for example, private information is exposed (43). In addition, disclosure of any of the client's information to the Internet community or the public by direct or indirect means can be a violation of privacy policies as well as the ethical code of conduct. Any type of electronic transaction that leads to disclosure of private information can also be taken as a violation of privacy policies and the code of ethics. Confidential information should, therefore, be kept private by any forensic investigator. The next section elaborates on the personnel-related challenges faced by digital forensics.

Personnel-related Challenges

As with any potential forensic evidence, testimony that clearly establishes that the potential digital evidence has been under the control of responsible personnel and well-trained digital forensic investigators is required to assure the court of the fact that the evidence is complete and has not been tampered with in any way. In the sub-sections to follow, therefore, some of the identified personnel-related challenges faced by digital forensics are explained in more details.

Lack of Qualified Digital Forensic Personnel (Training, Education and Certification)- According to Desai et al. (44), digital forensics (DF) has become an important field due to the increase in digital crimes. However, there is a shortage of trained digital forensic personnel in this field. The competition for employing digital forensic specialists in law enforcement is fierce. Qualified digital forensic experts are a challenge to find, even in the private sector. Even if technically proficient specialists are available, very few are trained or certified to deliver convincing, scientifically valid and expert witness testimony in a court of law or civil proceedings.

Semantic Disparities in Digital Forensics - Digital forensics is a growing field that is gaining popularity among many computer professionals, law enforcement agencies, forensic practitioners and other stakeholders who must always cooperate. Unfortunately, this has created an environment challenged with semantic disparities within the domain (45). Besides, cooperation between the computer professionals, law enforcement agencies and other forensic practitioners, presupposes the reconciliation of the semantic disparities that are bound to occur in the domain which is also a big challenge.

Lack of Unified Formal Representation of Digital Forensic Domain Knowledge - According to Hoss and Carver (2), there is currently no unified formal representation of digital forensic knowledge or standardised procedures for gathering and analysing knowledge. This lack of a unified representation inevitably results in incompatibility among digital forensic analysis tools. Errors in analysis and in the interpretation of potential digital evidence are more likely where there is no formalised or standardised procedure for collecting, preserving and analysing digital evidence (46).

Lack of Forensic Knowledge Reuse among Personnel - According to Bruschi et al. (47), when detectives perform investigations and manage a huge amount of information, they make use of specialised skills and analyse a wide knowledge base of potential evidence. Most of the work is not explicitly recorded and this hampers external reviews and training. Past experience may and should be used to train new personnel, to foster knowledge sharing and reuse among detective communities, and to expose collected information to quality assessment by third parties. Hoss and Carver (2) adds that the preparation of potential digital evidence may often be inadequate to support legal action in court and/or civil prosecution, because the potential evidence and procedures utilised to extract the digital evidence did not adhere to the acceptable legal practices.

Forensic Investigator Licensing Requirements - In a paper by Schwerha (48), there has been a push in the United States to require digital forensic professionals to become licensed as private investigators. However, there are many reasons why digital forensic professionals should not be required to license as private investigators. Such requirement of licensure will limit the field unnecessarily as there are too many potential jurisdictions worldwide to allow the average practitioner to be licensed in every jurisdiction. Moreover,

requiring digital forensic professionals to become licensed private investigators will create a big challenge to most average investigators worldwide. The requirement to be a licensed private investigator has little or no connection to the skill set that is necessary to be a high-quality digital forensics professional (48). In the next section the operational challenges faced by digital forensics are discussed.

Operational Challenges

According to Whitehead (49), digital crimes (perhaps more than any other type of crime), can be international in their operational scope. There is a need for basic guidelines for the evidence collection process to be established worldwide. This ranges from broad principles that apply to nearly every investigation, through organisational practices so that a minimum standard of planning, performance, monitoring, recording and reporting is maintained, to recommended processes, procedures, software and hardware solutions. In this sub-section of the paper we explain in more details, some of the identified operational challenges faced by digital forensics.

Incidence Detection, Response and Prevention - Conventional IT environments with on-premises data processing mostly rely on an internal security incident management process that uses monitoring, log file analyses, intrusion detection systems, as well as data loss prevention (DLP) to detect intruders, attacks and data loss. According to Beham (50), detecting security incidents is often a challenge especially for cloud users. Moreover, incident response is needed because attacks frequently compromise personal and business data. It is critically important to respond quickly and efficiently when security breaches occur, so as to minimise the loss or theft of information and disruption of services caused by incidents (51).

Lack of Standardised Processes and Procedures - The lack of standardisation in digital forensics seriously hinders the investigation process (52) and makes it difficult to produce legally admissible digital evidence. There is currently no standardised digital forensic investigation process model for recovering potential digital evidence. According to Köhn et al. (5), the number of digital forensic models that exist has added to the complexity of the field. This has, therefore, led to a call for standardisation (4) so as to facilitate the investigation process. Recent research has also urged the need for new forensic techniques and tools that will be able to successfully investigate anti-forensics methods (53).

Significant Manual Intervention and Analysis - In most cases a physical hard drive image will have to be manually inspected and analysed. This process may be simple in a single drive, single partition, or a completely allocated disk drive. However, the process becomes complex and poses a challenge with multi-volume Redundant Array of Independent Disks (RAID) configurations (54). According to Ayers (55), digital forensic analysis is a very complex undertaking. Thus, whenever the process is under manual control, mistakes will be made and bias could be introduced, even inadvertently.

Digital Forensic Readiness Challenge in Organisations - According to Mohay (16), forensic readiness is the extent to which computer systems or computer networks record activities and data in such a manner that the records are sufficient in their extent for subsequent forensic purposes, and the records are acceptable in terms of their perceived authenticity as evidence in subsequent forensic investigations. However, Cobb (56) states that digital forensic readiness sounds like a daunting challenge to most organisations.

With the advances in cloud computing, organisations have been forced to change the way they plan, develop and enact their IT strategies. According to Reilly et al., (57) cloud computing has not been thoroughly considered in terms of its forensic readiness. Hence, there is a definite need to consider current best practices to include, for example, certain aspects of digital forensic readiness in the existing practices to address the challenges brought about by lack of forensics readiness in organisations. Barske et al. (58) also adds that, although the need for digital forensics and digital evidence in organisations has been explored (as has been the need for digital forensic readiness within organisations); decision makers still need to understand what is needed within their organisations to ensure digital forensic readiness.

Trust and Audit Trails - The goal of digital forensics is to examine digital media in a forensically sound manner but with additional guidelines and trusted procedures designed to create legal audit trails. The proof of clear and original audit trails play a key role in the user accountability and digital forensics. However, it is possible that an intruder may edit or delete the audit trail on a computer, especially weakly-protected personal computers (59). Sophisticated rootkits that dynamically modify kernels of running systems to hide what is

happening, or even to produce false results are also on the increase. The next section presents a critical evaluation of the proposed taxonomy of challenges for digital forensics.

Critical Evaluation of the Proposed Taxonomy of Challenges for Digital Forensics

The taxonomy presented in this paper is a new contribution in the DF domain. The scope of the taxonomy is defined by the categories of the digital forensic challenges identified in Table 1. The main categories of the challenges as depicted in the taxonomy are technical challenges; legal systems and/or law enforcement challenges; personnel-related challenges, and operational challenges. These categories are further explained in terms of their scope. The sub-categories identified in the taxonomy include examples where applicable. The reader is again reminded that most of the sub-categories identified in the taxonomy were selected as common examples to facilitate this study and do not by any means constitute an exhaustive list.

The proposed taxonomy can be used in the digital forensic domain, for example, to explicitly describe processes and procedures that focus on addressing individual challenges. Moreover, the taxonomy in this paper can also help to map and categorise different digital forensic challenges, as well as create a common platform to share information in the digital forensic domain.

For the sake of training, education and certification, the sub-categories of the digital forensic challenges identified in the taxonomy can be used to give direction to institutions of higher learning, especially when developing curriculums and education material for different undergraduate programmes as well as research projects for postgraduate study. Such areas will help to produce programmes for specialists and generalists for the larger digital forensic industry. The taxonomy can also present new research opportunities to students – especially for those interested in how to resolve specific identified digital forensic challenges.

Developers of digital forensic tools can, further, use the taxonomy to fine-tune digital forensic tools to cover as many sub-categories of challenges as possible in the case of digital forensic investigations. Developers will also find the taxonomy in this paper useful, especially when considering new digital forensic tools and techniques for addressing specific challenges of interest in the digital forensic domain. The proposed taxonomy can also be used to facilitate the assessment of existing or new tools to fully examine the extent to which it addresses the specific identified digital forensic challenges.

Individuals should also be able to use the proposed taxonomy to carefully and accurately identify and classify – with less effort – the different challenges faced by digital forensics. Without such taxonomy it would be hard and time consuming for anyone to be sure of the existence of certain specific challenges that they would want to explore further.

Finally, the taxonomy presented in this paper has been designed in such a way as to accommodate new categories of challenges and sub-categories that may emerge as a result of technological change or domain evolution. It should be possible for individuals to add new categories and sub-categories of the challenges, including potential modifications in any of the aforementioned categories or sub-categories. To the best of the authors' knowledge, there exists no other work of this kind in the domain of digital forensics; therefore, this is a novel contribution towards advancing the digital forensic domain.

Conclusions

The problem addressed in this paper involved the vast number of challenges faced by digital forensics. Despite numerous researchers and practitioners having studied and analysed various known digital forensic challenges for the last decade, there still exists a need for a formal classification of these challenges. This paper, therefore, presents a taxonomy of the various challenges faced by digital forensics to date. The taxonomy classifies the large number of digital forensic challenges into 4 well-defined and easily understood categories.

With the continued developments and research in digital forensics, the taxonomy can be of value to tools developers in assessing the extent to which existing and new digital forensic tools can address the identified challenges. Institutions of higher education can furthermore benefit from the taxonomy when developing educational material for different undergraduate programmes as well as research projects for postgraduate studies. The taxonomy in this paper can easily be expanded to include additional categories and sub-categories of challenges that may crop up in the future.

Finally, as part of future work, the authors are now engaged in a research project to try and develop specifications and ontologies that create a unified formal representation of the digital forensic domain knowledge and information even more as a way towards resolving existing endemic disparities in digital forensics. However, much research still needs to be carried out so as to provide directions on how to address many of the challenges faced by digital forensics. More research also needs to be conducted to improve the taxonomy proposed in this paper and spark further discussion on the development of new digital forensic taxonomies.

Acknowledgements
The authors wish to thank the members of the Information and Computer Security Architecture (ICSA) research group, Department of Computer Science, University of Pretoria and Kabarak University, for their support throughout the process of writing this paper.

References

1. Webb, K.K. Predicting Processor Performance, Issues in Information Systems 2004; 5 (1):340-346.
2. Hoss, A.M. and Carver, D.L. Weaving Ontologies to Support Digital Forensic Analysis, ISI 2009; Richardson, TX, USA.
3. Kara L. N., Brian, H. and Matt, B. Digital Forensics: Defining a Research Agenda. Proceedings of the 42nd Hawaii International Conference on System Sciences 2009; 1-6.
4. ISO/IEC 27043. Information technology - Security techniques - Digital evidence investigation principles and processes (Draft). Available at: http://www.iso27001security.com/html/27043.html [Accessed September 17, 2013].
5. Köhn, M., Eloff, J.H.P., and Olivier M.S. Framework for a Digital Forensic Investigation, in H.S. Venter, J.H.P. Eloff, L. Labuschagne and M.M. Eloff (Eds), proceedings of the ISSA 2006 from Insight to Foresight Conference, Sandton, South Africa.
6. Altschaffel, R., Kiltz, S., and Dittmann, J. From the Computer Incident Taxonomy to a Computer Forensic Examination Taxonomy. Proceedings of the Fifth International Conference on IT Security Incident Management and IT Forensics.
7. Hoefer, C.N. and Karagiannis, G. Taxonomy of cloud computing services. Proceedings of the IEEE GLOBECOM workshop on enabling the future service-oriented internet 2010; 1345-1350.
8. Strauch, S., Kopp, O., Leymann, F. and Unger, T. A Taxonomy for Cloud Data Hosting Solutions, Ninth IEEE International Conference on Dependable, Autonomic and Secure Computing 2011.
9. Lupiana, D., O'Driscoll, C. and Mtenzi, F. Taxonomy for Ubiquitous Computing Environments, First International Conference on Networked Digital Technologies 2009;469-475.
10. Sansurooah, K. Taxonomy of computer forensics methodologies and procedures for digital evidence seizure. Originally published in the Proceedings of the 4th Australian Digital Forensics Conference (Security Research Institute Conferences) 2006. Edith Cowan University, Perth, Western Australia.
11. Sriram, R. Digital Forensic Research: Current State-of-the-Art. CSI Transactions on ICT 2013; 1(1):91-114. Available at: http://securecyberspace.org/yahoo_site_admin/assets/docs/df-survey.334154504.pdf [Accessed June 22, 2013].
12. Garfinkel, S. Digital forensics research: The next 10 years. Digital Investigation 2010, 7:S64-S73.
13. Gallegos, F. Computer Forensics: An Overview. Information Systems Audit and Control Association (ISCA) 2005, vol. 6, Available at: http://www.isaca.org/Journal/Past-Issues/2005/Volume-6/Documents/jpdf0506-Computer-Forensics-An.pdf [Accessed February 18, 2013].
14. Thinkquest. Cryptanalysis: Introduction. Available at: http://library.thinkquest.org/27993/crypto/classic/analysis1.shtml [Accessed April 8, 2013].
15. Lowman, S. The Effect of File and Disk Encryption on Computer Forensics. Available at: http://lowmanio.co.uk/share/The%20Effect%20of%20File%20and%20Disk%20Encryption%20on%20Computer%20Forensics.pdf [Accessed February 21, 2013].
16. Mohay, G. Technical Challenges and Directions for Digital Forensics, Proceedings of the First International Workshop on Systematic Approaches to Digital Forensic Engineering, 2005:155-161.
17. Libby, D.A. Distributed Computer Forensics: Challenges and Possible Solutions, Available at: http://selil.com/archives/2668 [Accessed February 16, 2013].
18. Arthur, K.K., and Hein S.V. An Investigation into Computer Forensic Tools. Proceedings of the ISSA conference 2004. Midrand, South Africa.

19. Richard, G.G. and Roussev, V. Digital Forensics Tools - The Next Generation, Idea Group Inc, 2006:76-91.
20. DOJ. Volatility of digital evidence, Available at: http://www.policeone.com/police-products/investigation/tips/1655664-Volatility-of-digital-evidence/ [Accessed February 18, 2013].
21. Taute, B., Grobler, M. and Nare, S. Forensic Challenges for Handling Incidents and Crime in Cyberspace, Available at: http://researchspace.csir.co.za/dspace/bitstream/10204/3756/1/Taute_d1_2009.pdf [Accessed February 18, 2013].
22. Conserve O Gram. Digital Storage Media, National Service Park 2010, Number 22/5.
23. Reed, T. Time vs Technology and the Frailty of Digital Media, Available at: http://filmcourage.com/content/time-vs-technology-and-the-frailty-of-digital-media [Accessed August 15, 2013].
24. Harvey, R. Preserving Digital Materials - Google Books. Available at: http://books.google.co.za/books?id=Z_8glIHqKgQC&pg=PA128&lpg=PA128&dq=Limited+lifespan+of+digital+media&source=bl&ots=Qf3rNzycwR&sig=PtQPJhmT6dlifT-dPDGDAzfYCmI&hl=en&sa=X&ei=Dz8iUebCMcmwhAe4hYDIBQ&ved=0CEoQ6AEwBQ#v=onepage&q=Limited%20lifespan%20of%20digital%20media&f=false [Accessed February 18, 2013].
25. ACPO. Good Practice Guide for Computer-Based Electronic Evidence. Available at: http://www.7safe.com/electronic_evidence/ACPO_guidelines_computer_evidence.pdf [Accessed February 16, 2013].
26. Eroraha, I. Real-World Computer Forensics Challenges Facing Cyber Investigators, Computer Forensics Show 2010.
27. Sheward, M. Rock Solid: Will Digital Forensics Crack SSD's? Available at: http://resources.infosecinstitute.com/ssd-forensics/ [Accessed February 18, 2013].
28. Elancheran, A. Computer Forensics, Available at: http://uwcisa.uwaterloo.ca/Biblio2/Topic/ACC626%20Computer%20Forensics%20A%20Elancheran.pdf [Accessed February 18, 2013].
29. Garfinkel, S. Anti-Forensics: Techniques, Detection and Countermeasures, 2nd International Conference on i-Warfare and Security, 20087: 77-84.
30. Liu, V. and Brown, F. Bleeding-Edge Anti-Forensics, Infosec World Conference & Expo 2006, MIS Training Institute.
31. Bennett, W.D. The Challenges Facing Computer Forensics Investigators in Obtaining Information from Mobile Devices for Use in Criminal Investigations, Available at: http://articles.forensicfocus.com/2011/08/22/the-challenges-facing-computer-forensics-investigators-in-obtaining-information-from-mobile-devices-for-use-in-criminal-investigations/ [Accessed February 16, 2013].
32. Yates, M. Practical Investigations of Digital Forensics Tools for Mobile Devices, Proceedings of the Information Security Curriculum Development Conference, 2010:156-162.
33. Ferguson, R.I. Challenges in Digital Forensic Research. Available at: http://scone.cs.st-andrews.ac.uk/cybersecurity/slides/Ferguson-DigitalForensicsResearchChallenges.pdf [Accessed June 20, 2013].
34. Leslie, W., Will, V. and Edgar, A.W. Meeting the Challenges of Cloud Computing. Available at: http://www.accenture.com/us-en/outlook/Pages/outlook-online-2011-challenges-cloud-computing.aspx [Accessed June 20, 2013].
35. Bassett, R., Bass, L. and O'Brien, P. Computer Forensics: An Essential Ingredient for Cyber Security. Journal of Information Science and Technology 2006: 22-32.
36. Vaciago, G. Cloud Computing and Data Jurisdiction: A New Challenge for Digital Forensics. Proceedings of the third International Conference on Technical and Legal Aspects of the e-Society, CYBERLAWS 2012.
37. Lauren, M. Info-security - Cybercrime Knows No Borders. Available at: http://www.infosecurity-magazine.com/view/18074/cybercrime-knows-no-borders-/ [Accessed February 16, 2013].
38. Khan, A., Uffe, K.W. and Nasrullah, M. Digital Forensics and Crime Investigation: Legal Issues in Prosecution at National Level, Fifth International Workshop on Systematic Approaches to Digital Forensic Engineering, 2010:133- 140.
39. Ieong, R.S.C. FORZA-Digital Forensics investigation framework that incorporate legal issues, Digital Investigation: The International Journal of Digital Forensics & Incident Response 2006;3:29-36.

40. Craiger, P., Swauger, J., Marberry, C. and Hendricks, C. Validation of Digital Forensics Tools. Digital Crime and Forensic Science in Cyberspace, edited by Panagiotis Kanellis, Evangelos Kiountouzis, Nicholas Kolokotronis, and Drakoulis Martakos© 2006, Idea Group Inc.
41. Mercuri, R. Criminal Defense Challenges in Computer Forensics, In Proceedings of the Digital Forensics and Cyber Crime Conference, ICDF2C 2009, Albany, NY, USA.
42. Burgess, S. Computer Forensics - Criminal vs Civil: What's the Difference? Available at: http://www.burgessforensics.com/Civ_Criminal.php [Accessed February 23, 2013].
43. Anon. Computer Forensics Privacy Issues. Available at: http://www.computerforensics1.com/privacy-computer-forensic.html [Accessed February 23, 2013].
44. Desai, A.M., Fitzgerald, D. and Hoanca, B. Offering a Digital Forensics Course in Anchorage, Alaska. Information Systems Education Journal 2009, 7(35). http://isedj.org/7/35/. ISSN: 1545-679X. (A preliminary version appears in The Proceedings of ISECON 2006: §5114. ISSN: 1542-7382).
45. Karie, N.M. and Venter, H.S. Significance of Semantic Reconciliation in Digital Forensics. In the proceedings of the Digital Forensics, Security and Law conference, 2013:71-80, Richmond, Virginia USA.
46. Chaikin, D. Network investigations of cyber-attacks: the limits of digital evidence, Crime Law Soc. Change 2006, 46: 239-256.
47. Bruschi, D., Martignoni, L. and Monga, M. How to Reuse Knowledge about Forensic Investigations, in 'Proceedings of Digital Forensic Research Workshop 2004'. Baltimore, MD, USA.
48. Schwerha, J.J. Why computer forensic professionals shouldn't be required to have private investigator licenses, Digital Investigation: The International Journal of Digital Forensics & Incident Response 2008, 5(1-2):71-72.
49. Whitehead, A. Weakness in Computer Forensics. Available at: http://free-backup.info/weaknesse-in-computer-forensics.html [Accessed February 23, 2013].
50. Beham, G. Incident Detection and Cloud Forensics – Security at a Glance, Available at: http://ipbr.wordpress.com/2012/08/30/incident-detection-and-cloud-forensics/ [Accessed February 16, 2013].
51. Cichonski, P., Millar, T., Grance, T. and Scarfone, K. Computer Security Incident Handling Guide 2012, Revision 2, NIST Special Publication 800-61.
52. Leigland, R. and Krings, A.W. A Formalization of Digital Forensics, International Journal of Digital Evidence 2004, 3(2):1-32.
53. Alharbi, S., Weber-Jahnke, J., and Traore, I. The Proactive and Reactive Digital Forensics Investigation Process: A Systematic Literature Review. International Journal of Security and Its Applications 2011 October, 5(4):59-71.
54. King, G.L. Forensics Plan Guide – Forensic Investigation Plan Cookbook 2006, SANS Institute, Computer Forensics and Incidence Response
55. Ayers, D. A second generation computer forensic analysis system. Digital Investigation: The International Journal of Digital Forensics & Incident Response 2009, 6:S34-S42.
56. Cobb, M. Digital forensic investigation procedure: Form a computer forensics policy, Available at: http://www.computerweekly.com/tip/Digital-forensic-investigation-procedure-Form-a-computer-forensics-policy [Accessed February 18, 2013].
57. Reilly, D., Wren, C. and Berry, T. Cloud Computing: Pros and Cons for Computer Forensic Investigations, International Journal of Multimedia and Image Processing (IJMIP) 2011 March; 1(1).
58. Barske, D., Stander, A. and Jordaan, J. A Digital Forensic Readiness Framework for South African SME's, Proceedings of ISSA Conference 2010.
59. Yong, G. Digital Forensics: Research Challenges and Open Problems. Available at: http://itsecurity.uiowa.edu/securityday/documents/guan.pdf [Accessed June 21, 2013].