CONTRIBUTIONS TO CYBER-FORENSICS: PROCESSES AND E-MAIL ANALYSIS

DJAMEL BENREDJEM

A THESIS
IN
THE DEPARTMENT
OF
ELECTRICAL AND COMPUTER ENGINEERING

PRESENTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF MASTER OF APPLIED SCIENCE IN ELECTRICAL AND
COMPUTER ENGINEERING
CONCORDIA UNIVERSITY
MONTRÉAL, QUÉBEC, CANADA

JULY 2007

Your file Votre référence
ISBN: 978-0-494-34431-6
Our file Notre référence
ISBN: 978-0-494-34431-6

CONCORDIA UNIVERSITY

School of Graduate Studies

This is to certify that the thesis prepared

By: **Djamel Benredjem**

Entitled: **Contributions to Cyber-Forensics: Processes and E-Mail Analysis**

and submitted in partial fulfillment of the requirements for the degree of

Master of Applied Science in Electrical and Computer Engineering

complies with the regulations of this University and meets the accepted standards with respect to originality and quality.

Signed by the final examining committee:

_____ Chair
Dr. Rabin Rault

_____ External Examiner
Dr. Khaled Galal

_____ Examiner
Dr. Otmane Ait Mohamed

_____ Examiner
Dr. Hakim Lounis

_____ Examiner
Dr. Prabir Bhattacharya

_____ Supervisor
Dr. Mourad Debbabi

Approved by_____
Chair of Department or Graduate Program Director

_____

Dr. Nabil Esmail, Dean
Faculty of Engineering and Computer Science

Date: _____ 2007

Abstract

Contributions to Cyber-Forensics: Processes and E-Mail Analysis

Djamel Benredjem

The primary intent of this thesis is to contribute to cyber forensics by addressing two important aspects that are processes and analysis techniques. In this thesis, first, the existing proposals of cyber forensic processes are compiled and evaluated according to a priori, well-defined criteria. Beside this evaluation, a new digital forensic process is developed that uses the high potential of the previous proposals and addresses their identified deficiencies. In addition, a new technique of e-mail evidence investigation is presented to be used in cyber forensics analysis. The proposed technique extensively uses and exploits data mining techniques for answering several important questions that are extremely relevant in a forensic setting, such as e-mail clustering, classification, social network computation and the author identification of anonymous e-mails. Finally, towards this objective, the architecture and the design of an environment are also elaborated that implement the aforementioned technique.

To my wife and lovely daughters

Acknowledgments

I am very grateful to all the people who have provided valuable contributions to this successful research. I would first like to thank my supervisor, Professor Dr. Mourad Debbabi, to whom, I owe a great debt of gratitude. His advices helped to shape and carry out this work. I feel privileged working under his guidance and receiving all support to complete this research. My appreciation goes to Dr. Rachida Dssouli, the Director of the Concordia Institute for Information Systems Engineering (CIISE) and other departmental staff. I would like also express my gratitude to the members of the evaluation committee for their acceptance of this examination task as well as for their precious comments and feedback. This research is part of a major initiative on cyber forensics analysis. The initiative is funded by a grant from PROMPT Quebec in collaboration with Bell Canada and the Computer Security Laboratory (CSL) of Concordia University. My special thanks go to all the members of this project for their support and insightful and valuable scientific comments. I express my gratitude to my CSL colleagues especially those involved in cyber forensics research, namely A. R. Arasteh, A. Sakha, A. Fry, M. Hedjazi, A. Szporer, S. Hasfi, and F. Iqbal. I appreciate Pr. H. Lounis, Dr. R. Hadjidj and Dr. A. Bedrouni for their valuable support and contribution. I also thank S. Ray for his friendship and support.

I have a deep and caring thought to my defunct parents who had strived for my success. Enormous appreciation goes to my wife for being my best support in the crucial moments, for her patience and understanding all along. I feel a great love towards my daughters who always smile, stimulate and encourage me. Lastly, I remember my brothers and sisters, family (especially to Khalida and Farid Ghanem), friends and all who inspire me throughout my career.

Djamel Benredjem on June 30th 2007

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Over the last few decades, the concept of Internet, web services, information systems, wireless networks, hand-held devices and other emergent technologies have rapidly evolved and penetrated almost every aspect of our lives. The developing array of Internet and other interactive services represents an extraordinary advance in the availability of informational resources across every sector of society. Consequently, governments, corporations and institution are vitally dependent on information, computers and network technologies. The widespread Internet connections and the availability of diverse information provide good opportunities for communication, business and learning. Unfortunately, along with its advantages, the power and pitfalls of Internet are often exploited in various ways for criminal purposes. Now-a-days, cyber-crime has become the major concern facing institutions, individuals and the worldwide online economy. Attacks on military, financial, communication and transportation systems pose a serious threat to the critical infrastructure and hence to the national security. In this context, security is turned to be one of the most important challenges that have ever faced in computing research. The security design continues

to fall short of protecting individuals, corporations and organizations against the exposure to various threats that may result in severe effects such as losses or danger of human lives, financial losses, personnel information theft, denial of service, etc.

These attacks raise major concerns from the law enforcement standpoint. Due to the borderless nature of cyber attacks and lack of supporting evidence, many criminals and offenders have been able to walk away. Therefore, cyber-forensics has emerged as a new discipline in order to tackle cyber-crime and deter criminals by providing scientifically proven methods and techniques to gather, process, interpret, and use digital evidence to bring a conclusive description of cyber-crime activities, and above all in proper way to meet the legal issues.

Accordingly, there is a desideratum that consists of designing and implementing a framework that incorporates:

- Well-defined processes that guide the cyber investigator in verifying cyber incidents, collecting clues and evidence, imaging the storage media, managing the chain of custody, analyzing the evidence and reporting on a cyber-crime.
- Well-established techniques to conduct a thorough analysis of digital evidence in order to determine the facts and their chronology and extract irrefutable proofs to back a cyber-crime investigation.

In this respect, the primary objective of this thesis is to contribute to cyber-forensics analysis by improving the state of the art in terms of forensic processes and digital forensic analysis techniques.

1.2 Objectives

More precisely, the objectives of this research may be depicted as follows:

- A survey on the state of the art forensic processes.
- Carrying out a comparative study of these proposals.
- Elaboration of a unified cyber-forensic process that may leverage the potential of existing processes and their comparison.
- Studying the state of the art on e-mail analysis.
- Developing digital forensic analysis techniques by proposing an innovative yet practical approach for email analysis.
- Design and implementation of e-mail forensics techniques.

So, as a whole, the intent is to leverage useful and valuable contributions to two important aspects of cyber-forensics that are processes and forensic analysis. It is important to mention that the research reported in this thesis, is a part of a major research project that is being executed at the Computer Security Laboratory in collaboration with Bell Canada under the PROMPT Quebec research partnership program. Accordingly, my research in this thesis, is turned into practical and efficient tools that are implemented in this unified framework of the aforementioned project.

1.3 Thesis Organization

The rest of this thesis is organized as follows: Chapter 2 is dedicated to a presentation of cyber-forensic processes as well as standalone evaluation. Chapter 3 is an elaboration of the comparative study of the state of the art forensic processes. Chapter 4 is devoted to a presentation of our proposed unified digital forensic process. Chapter 5 illustrates an innovative and practical forensic analysis technique of email. Finally, the conclusion is outlined in chapter 6 with valuable remarks on this research and a discussion of future work.

Chapter 2

Cyber Forensic Processes

2.1 Introduction

A cyber-forensic process is an application that may structure a digital forensic investigation. It consists of a precisely defined sequence of phases including their associated sub-phases, inputs, outputs, requirements, orders, standards and considerations that together constitute a framework for cyber-forensic investigation. The primary intention is to execute these phases during an investigation such as identification of the incident, incident response, crime scene preservation, acquisition, analysis and reporting in a forensically consolidated way. Despite the importance of these process, only a handful number of documents seems to be published on digital forensic processes as compared to the publications on software engineering processes.

In this chapter, we elaborate the most prominent forensic processes in a literature review. For each process, we recall the motivations and detail the underlying phases of the process. Besides, we also provide an evaluation to highlight its potential and weaknesses.

Before digging into the inner details of each process, we first depict the typical forensic process requirements:

- *Generality:* This requirement stipulates that the process needs to be general enough to be applicable to a large variety of digital forensic cases.

- *Specificity:* This requirement stipulates that the process needs to be specific enough in order to precisely guide the forensic investigator during the exercise of his/her activities.

- *Lawfulness:* This requirement stipulates that the process should take into account the realization of the court of law considerations such as chain of custody, traceability, authorizations, and the sound acquisition of evidence.

- *Iterativeness:* This requirement stipulates that the process should have provisions to repeat a certain number of steps as many times as required for the proper execution of the investigation.

- *Feedback:* This requirement stipulates that the process should allow for a phase to yield its output as input to another (potentially previous) phase.

- *Computer Support:* This requirement stipulates that the process should be enforced at the tools level. In other words the process should be implemented into a tool that will guide the investigators through the process phases.

- *Pertinence:* This requirement stipulated that the process should have provisions to obligate the investigator to collect and analyze only pertinent parts of the evidence while not missing any precious piece of it.

When we compiled these processes, we went through all the published references that we have been aware of. We did not exclude any published process. The studied processes are:

- *DFRWS Process:* The Digital Forensics Research Workshop produced a process [35, 45] that sets out the steps for digital forensic analysis. The proposal is not

meant to be a final comprehensive process, but rather a solid basis for future research on this issue.

- *EEDI Process:* The End-to-End Digital Investigation (EEDI) process [43] is a very structured approach for conducting complex digital investigations using the framework developed by the Digital Forensics Research Workshop (DFRWS). The EEDI process allows the investigators to use a very structured investigation technique that combines computer technology with traditional investigative methods.

- *Department of Justice Process:* This forensic process [47], as issued by the US Department of Justice in 2001, is a very detailed process that starts from the first responder of the crime scene and ends to the court-room. This proposal constitutes detailed guidelines to digital crime scene investigation.

- *Reith Process:* This forensic process [28] is an abstract model that is applicable to any technology or type of cyber crime. It is intended that the model can be used as a basis for developing more detailed methods for specific types of investigation with potentially new technologies.

- *Mandia Process:* This forensic process [29] is meant to meet the needs of organization and/or individuals who must respond to computer security incidents.

- *Carrier and Spafford Process:* In this forensic process [4], the authors relies on the physical crime scene model to derive their own digital investigation process.

- *Ò Ciardhuain Process:* This forensic process [1] combines previous existing processes and adds new activities that were not present before. It captures the full range of the investigation, even before processing the evidence.

- *Casey and Palmer Process:* This proposed process [7] is structured to encourage a complete and rigorous investigation that ensures proper evidence handling, and reduces the chances of mistakes that might be created by re-conceived theories and other potential pitfalls.

- *Beebe and Clark Process:* This forensic process [3] aims at simplifying the complex process of digital forensic investigation while providing a mechanism for including layers of detail that are needed by its users.

In what follows, we present in details each of these processes.

2.2 DFRWS Process

The Digital Forensics Research Workshop (DFRWS) [35, 45] framework is the result of a fruitful collaboration between researchers, investigators and practitioners in digital forensic science. The intent of this framework is to define a process that incorporates scientific methods into the evolving discipline of digital forensic Science. The proposal is structured into several classes, which in turn contain various elements that are essential to cyber-forensic investigations.

2.2.1 Process Description

During the workshop, a set of classes has been developed, as follows:

- Identification
- Preservation
- Collection
- Examination

- Analysis
- Presentation

In what follows, we further detail each of these classes. We will mark a phase by a "*" to indicate that it is required.

Identification

In this class, the investigator is notified of a possible incident. Data, items and different components are then associated with the allegation or incident. The identification class contains the following elements:

- **Event/Crime detection:* It implies the discovery of a direct evidence of an event that indicates a potential cyber criminal activity.
- *Resolve signature:* It implies the use of signature analysis and mapping by an automated system to detect an event, such as: an intrusion detection system or an anti-virus software program.
- *Profile detection:* It implies the match of a particular profile instead of just an explicit signature. Signatures usually apply to individual events. Events, on the other hand, come together into an attack scenario, or attack profile. The profile is composed of different events, a pattern of behavior, or a pattern of specific results of an attack. The profile detection also relies upon an automated event detection system.
- *Anomalous detection:* It implies the detection of unusual behavior patterns that fall outside of the observed norm. It also needs to use of a detection system.
- *Complaints:* It relies on actual observers and their reporting of a potential criminal event.

- *System monitoring:* It requires a security monitoring infrastructure that may incorporate an anti-virus, a firewall, or an intrusion detection system. This element may be used together with another element of this class.

- *Audit analysis:* It involves the recording and analysis of various audit logs that are produced by different sources, targets and intermediate devices.

Preservation

This class refers to the integrity or state of the case evidence. The preservation class considers the following elements:

- **Case management:* It deals with the management and documentation of the case investigation. Elements such as notes, process controls, quality control and procedural issues should be documented along with the actions of the investigators and digital forensic examiners.

- *Imaging technologies:* It refers to the technology, used for imaging computer media. It includes the technology that is used to create an image of computer media and the technology used to extract such items as logs from a device. In this case, the log might be extracted from a bitstream image or it might be read out of the device to a peripheral as a result of a command issued by the investigator.

- **Chain of custody:* It refers to the process of monitoring each access to the item of evidence from the time it is collected until the time it is used in a court of law. This ensures a limited and strictly controlled access to the evidence and a better preservation of its integrity.

- **Time synchronization:* It refers to the normalization and the synchronization of all evidence pieces to the same timeline. This evidence could either be temporally adjusted to a common device, or a common time zone. Universal Time (UT) or Greenwich Mean Time (GMT) are examples of a common time zone. The investigator should take note of the difference of a particular piece of digital evidence from the pre-determined standard time.

Collection

This class refers to the methods by which the investigators and forensic examiners acquire evidence in a digital environment. The preservation class is considered as an element of this class. The specific elements of this class are:

- **Approved methods:* The training and qualifications of the forensic examiners and investigators as well as the techniques used by them to collect evidence should be approved by a court of law.

- *Approved software:* This step refers to the software products that are used to collect the evidence. To be admissible in an investigation, any software program has to be completely identical to a software that went through the whole process of court testing and that was already approved. If it is not the case, the program in question should undergo court-testing before being used.

- *Approved hardware:* Similar to the methods and the software products, the hardware devices, used in the investigative process, must be exactly the same as the devices that have survived to the court challenges.

- **Legal authority:* A complete legal authorization (e.g. search warrant, policy) is necessary in order to be able to access any evidence and before extracting

information from computer media. Any evidence seized illegally would become inadmissible in court.

- *Lossless compression:* This step refers to the eventual compression techniques that are parts of encryption, backup or digital signature software used to collect and/or preserve the evidence. The software program that uses compression has to be absolutely lossless, and proved to be not altering the evidence on which it is used.

- *Sampling:* Any sampling technique that is used in the collection phase must be proven to preserve the integrity of the evidence. Sampling methods should be valid and all the conclusions that can be drawn from the sample have to be clearly defined.

- *Data reduction:* Any technique or program that is used to reduce data should be used on a copy of the evidence, to ensure that the integrity is preserved. It must be proven that these techniques and programs will produce results that are valid and repeatable, and do not alter the collected evidence.

- *Recovery techniques:* The hardware, software and methods that are used by the forensic examiner to recover and extract evidence should be approved.

Examination

This class includes the tools and techniques that are used to examine the evidence. This class concentrates on evidence discovery and extraction and leaves the conclusions for the analysis class. The preservation class still exists in this part of the investigation. It is made of the following specific elements:

- **Traceability:* For the conclusions to be deemed correct, valid and reliable at the time of the investigation, investigators have to rely on a traceable and continuous

chain of evidence. By doing so, every piece of the evidence is documented and the circumstances of its collection are known and verifiable.

- *Validation techniques:* This step refers to the techniques that are used to support the evidence. Usually, the latter is corroborated by other pertinent pieces of evidence. If the technical validity of a digital evidence is determined, then it may stand on its own.

- *Filtering techniques:* If filtering techniques are used, they have to be identified and described by the investigator. Filtering is used, among other things, to extract evidence from a gross data collection, which is a file or files containing data obtained from a digital source that contain individual evidentiary data. Filtering techniques have to be properly understood by the investigator and clearly defined.

- *Pattern matching:* This step focuses on the methods, used to detect potential events relying on some pre-determined signatures or patterns. When the signatures or patterns are unclear, vague or demonstrate a large number of false positives or negatives, the evidence and the conclusions drawn from them will be doubtful and challenged in court.

- *Hidden data discovery:* This step refers to finding evidence, which is hidden on computer media, using encryption and steganography techniques. This step also includes the deleted data that may be recovered forensically.

- *Hidden data extraction:* This element refers to the extraction of hidden evidence from a gross data collection.

Analysis

After the evidence has been found, collected, identified and extracted from a gross data collection, it has to be analyzed using different techniques. These techniques are required to be valid, because they will influence directly the conclusions obtained from the evidence and the credibility of the evidence chain constructed from there.

Presentation

In this final class, the investigators will present all the conclusions of the investigation. These conclusions have to be expressed in a structured, clear, concise and objective report. Figure 1 depicts the six classes of this process, while Figure 2 depicts the overall steps of the process.

2.2.2 Evaluation

DFRWS is a very useful efforts towards a more practical and comprehensive digital forensic process. The authors devoted great effort to identify and structure the activities involved into a digital investigation. Furthermore, the process proposal is detailed and accounted for most of the forensic tasks. However, as the process is purposed for general cases, it does not elaborate on the legal steps required for the authorization class based on a particular type of case or legal system. Furthermore, no provisions are given for the incident reporting step in the identification class. Also, it does not account for the packaging, transportation and storage of the evidence. Moreover, the analysis phase needs to be more detailed. It is also important to mention that the process does not make explicit the need for iterating over some previous phases or providing them with feedback from subsequent steps. Actually, The process is proposed as a linear proposal of phases. However, at some points during the analysis phase, the investigator might require to go back to previous phases, for instance to

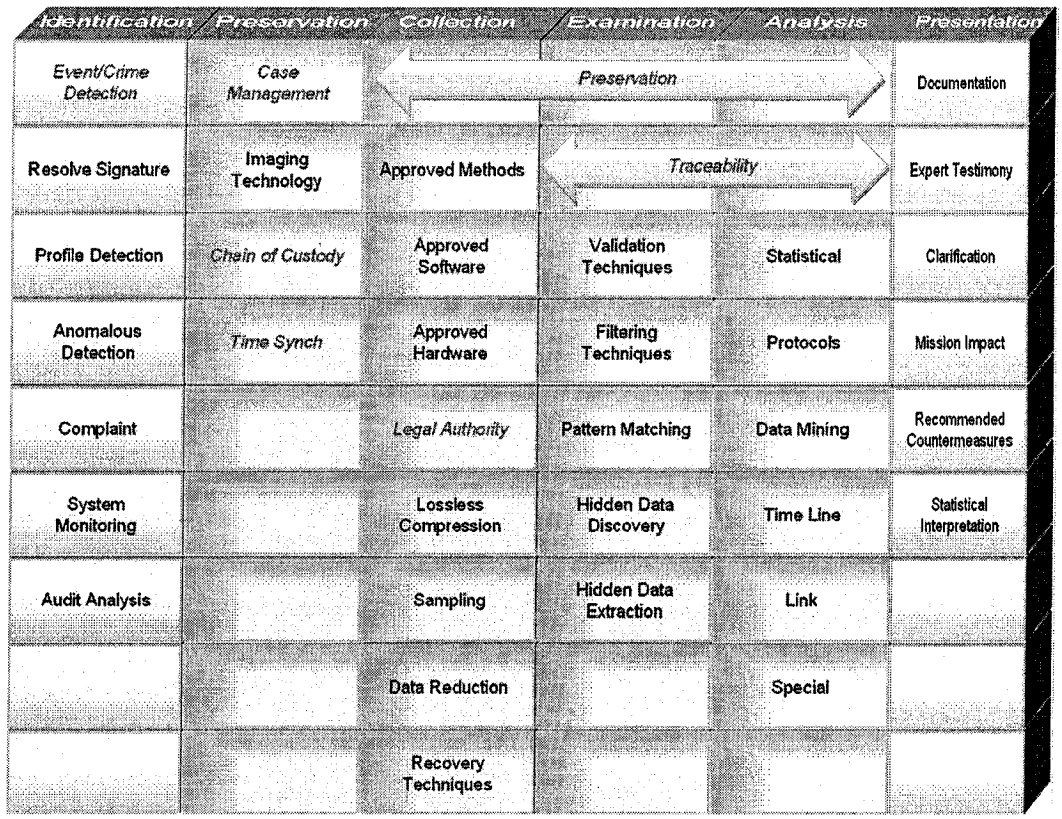

Figure 1: DFRWS Forensics Process

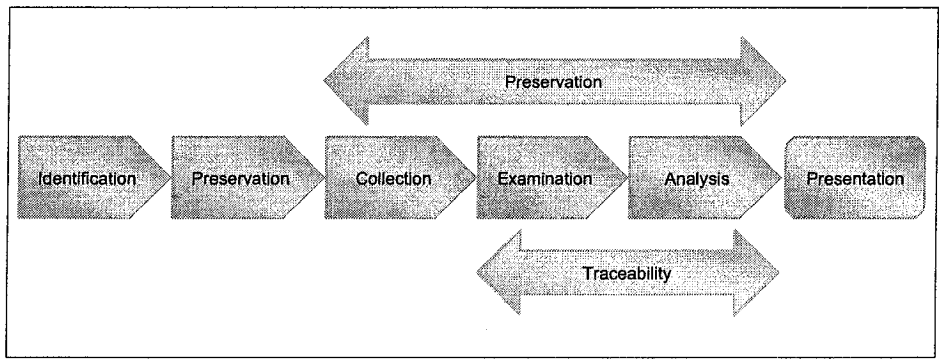

Figure 2: DFRWS Forensics Process

collect omore pieces of evidence. Therefore a sound digital forensic process should account for the iterative nature of forensic investigations.

2.3 EEDI Process

The End-to-End digital investigation process (EEDI) [43] is built on top of the DFRWS process. It proposes a refined and structured approach to conduct complex digital investigations. It can be seen as a collection of steps that should be included in the collection, examination and analysis classes of the DFRWS Framework.

2.3.1 Process Description

The proposed steps of EEDI are as follows:

- Collecting evidence
- Analysis of individual events
- Preliminary correlation
- Event normalizing
- Event de-confliction
- Second level correlation
- Timeline analysis
- Chain of evidence construction
- Corroboration

In what follows, we detail each step of the aforementioned list.

Collecting Evidence

When an incident occurs, it leads to a compromise (e.g. unauthorized access, modification of system, data destruction). As soon as the incident is identified, the evidence collection should begin. Each type of evidence has a particular method for its collection. The following evidence is important in this step:

- Image of affected computers
- Logs of intermediate devices, especially those on the Internet
- Logs of affected computers
- Logs and data from intrusion detection systems, firewalls, etc.

Analysis of Individual Events

Generally, the alert or incident is composed of one or more individual events, which are reported in different logs from different devices. The investigator should, before proceeding further, examine all these events, assess their importance with respect to the ongoing investigation and identify how they might connect to each other.

Preliminary Correlation

Once the individual events are examined, they need to be linked into a chain of evidence. This step of the investigation will first help in understanding what happened on the underlying systems or devices. Second, it helps to judge the order of occurrence of events.

Event Normalizing

When events are reported from multiple sources, they might have different syntax. Prior to the analysis, the events need to be normalized i.e. brought to a common

syntax where they are comparable. This will ease subsequent analysis. In addition, the duplications, if any, should be removed.

Event De-Confliction

If the events are reported multiple times from different sources, the EEDI process should view the group of events as a single event instead of multiple events. Whenever events are conflicting, an analysis should be undertaken in order to resolve the underlying conflicts. This process is referred to de-confliction.

Second Level Correlation

This step is a continuation of the preliminary correlation, except, at this point, the events have been refined through normalization and/or de-confliction, which should enable for more correlation.

Timeline Analysis

All the normalized and de-conflicted events are used to build a timeline. The whole event analysis, correlation, de-confliction and timeline analysis are iterative, meaning that as the investigation unfolds and new evidence is recovered, the iterative process should constantly be updated.

Chain of evidence construction

Once the timeline of events is set up, the construction of a coherent chain of evidence may begin. The main goal of this step is to come up with a chain of evidence where each link is supported by one or more pieces of evidence, and will lead to the next link in the chain. Nevertheless, a completely defined chain of evidence rarely happens because there are often gaps in the evidence collection due to missing data and logs.

Corroboration

In this step, the investigator tries to backup and support each evidence with another, independent, evidence or events. As stated in [43], the best evidence is the one that has been developed digitally and corroborated through traditional investigation or vice versa.

Figure 3 depicts the overall steps of the EEDI process.

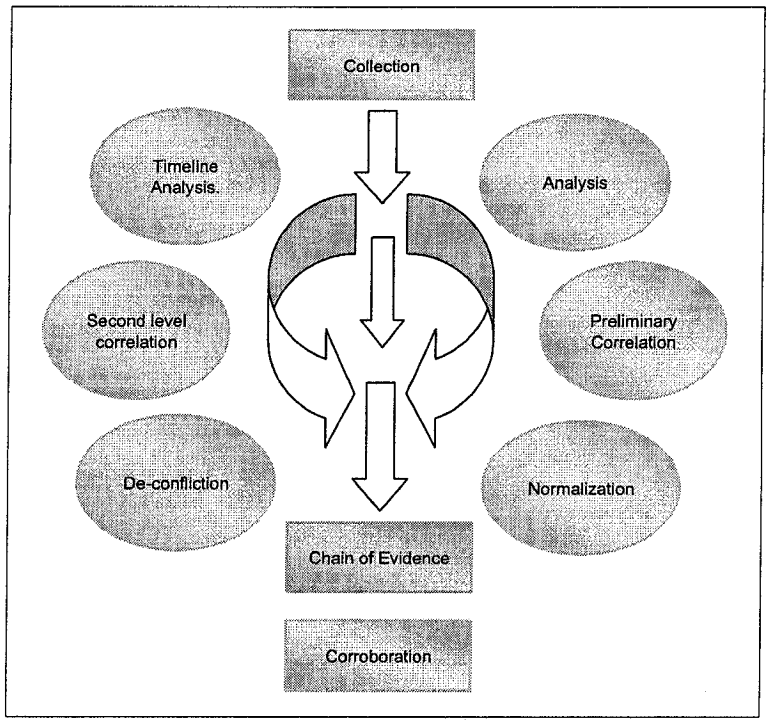

Figure 3: EDDI Forensic Process

2.3.2 Digital Investigation Process Language

A very useful and important feature of EEDI process is the proposition of a language called the Digital Investigation Process Language (DIPL). DIPL is the first known

modeling language for digital forensic science. It is derived from a LISP dialect, and it permits the description of an investigation in formal functional terms.

2.3.3 Analysis

EDDI is a significant refinement and extension of the DFRWS process. It provides more structure and details valuable guidelines in a digital forensic investigation. Besides, the idea of introducing a dedicated language, such as DIPL, is a major contribution since it allows for the formal specification of cyber investigation activities. Such an information may be extremely relevant for planning investigations, assessing productivity (comparing what has been planned with what has been done), admissibility analysis, investigation mining, analytics and statistics, etc. Nevertheless, EEDI does not consider the verification of the authenticity of the events. This is of paramount importance while fake events, generated by the attacker or the events, such as: the log files are manipulated by the attacker to mislead the investigation. The event de-confliction phase should also involve the event verification step. Furthermore, the model assumes that in the beginning of the investigation, all the pertinent and necessary events are known. However, as the investigation proceeds, more evidence will be required. However, since the iterations of the proposed model are during the analysis phases, it will become impossible to acquire new evidence.

2.4 Department of Justice Process

The U.S Department of Justice issued detailed guidelines to digital crime scene investigation [47]. These guidelines constitute an investigation process that starts with the first responder of crime scene and ends at the court-room.

2.4.1 Process Description

The main phased of this investigation are as follows: following:

- Securing and evaluating the scene
- Documenting the scene
- Evidence collection
- Packaging, transportation and storage
- Examination
- Analysis
- Reporting

In what follows, we detail each of these phases.

Securing and Evaluating the Scene

During this phase, the first responder has the duty to take all the measures that are needed to secure the crime scene. This is like cordoning a conventional crime-scene. In other words, he has to ensure the safety of every person at the scene, and most importantly protect the integrity and state of all the evidence.

Documenting the Scene

In any investigation, it is of paramount importance to document every detail of the crime scene and each action taken by the investigators throughout the investigation. The documentation includes pictures, notes, sketches, etc. It is also important to precisely record the location, time and condition of computers, storage media, other electronic devices and conventional evidence.

Evidence Collection

Computer evidence must be handled with a very special care during the collection phase to preserve its physical integrity as well as the data, it contains. Special considerations should be given to the data that might get damaged or altered from electromagnetic fields. The investigator needs to take appropriate measures in that regard. It should be noticed that the search for evidence and its collection at an electronic crime scene may require a search warrant.

Packaging, Transportation and Storage

Each electronic device should be handled with a special care and the investigators should follow a strict protocol, such as: how to package, transport and store the evidence. Many factors like humidity, temperature and static electricity can modify the original state of the evidence. The whole process should be well-documented to maintain the chain of custody.

Examination

Every digital crime should be handled according to its specificities. In a published guideline [47], the U.S Department of Justice states different categories of digital crimes, and what part of the evidence, the investigator should focus on examining for each crime type.

Analysis

Once the evidence is examined, all the pieces that are extracted will be subjected to extensive analysis in order to draw conclusions out of what was collected.

Reporting

The whole investigation process, from the arrival at the crime scene to the conclusions of the analysis, as well as the chain of custody, is described to a report that will then be presented in a court of law. Figure 4 depicts all the steps of the Department of Justice process.

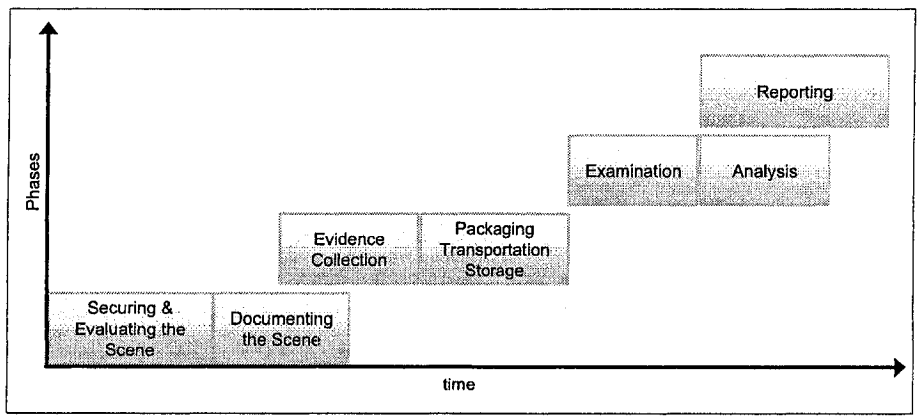

Figure 4: Department of Justice Process

2.4.2 Analysis

The Department of Justice process has very efficient features such as a dedicated phase to the crime scene investigation which is lacking in the previously presented processes. In addition, it has a great support for different types of incidents. Moreover, the proposed process is the only one that has incorporated crime and evidence categorization and in this way the process has provided more detailed sub-steps based on the type of the crime and evidence at hand. In addition, the packaging, transportation and storage are introduced as sub-steps of the collection phase that is an indication of the fact that the process approximates a common investigation scenario.

Though being one of the most detailed proposed forensics processes, the process

does not have a planing and follow-up phases. Also the process does not provide sufficient guidelines for evidence analysis. Lastly, the process does not support feedback and phase iteration, which are really needed in real-life scenarios.

2.5 Reith Process

Reith et al. researchers at the U.S. Air Force, proposed an abstract model for forensic investigation in [28]. This model was developed as a result of a comparative study of previous forensic processes. As stated by the authors, the process was meant to be:

- A consistent and standardized framework for digital forensic tool development
- Applicable to future digital technologies
- Usable by judicial members to relate technology with non-technical observers
- A model for incorporating non-digital electronic technologies

2.5.1 Process Description

The proposed abstract model includes the following steps:

- Identification
- Preparation
- Approach strategy
- Preservation
- Collection
- Examination

- Analysis
- Presentation
- Returning evidence

In the following, we detail each of these steps.

Identification

It refers to the detection of occurrence for an incident and the identification of its type.

Preparation

It refers to the organization of all what will be needed to undertake the investigation, such as tools (e.g. software, hardware), search warrants, techniques, authorization, etc.

Approach Strategy

It consists of setting up an appropriate approach depending on the nature of the cyber-crime incident. This approach should cater for maximizing the collection of evidence and reducing the impact on victims.

Preservation

It refers to the protection of the state and integrity of the crime scene and the physical and digital evidence that it contains.

Collection

It refers to the use of proper procedures and the gathering of all the evidence, found on the scene. The digital evidence should be duplicated to preserve the state of the original version.

Examination

It refers to the identification and the localization of highly relevant evidence together with the preparation of a detailed documentation to be used by the analysis phase.

Analysis

In refers to the determination of significance, reconstruction of data fragments and the drawing of conclusions, based on the found and reconstructed evidence.

Presentation

It refers to the production of a report that summarizes and explains the conclusions of the investigation.

Returning Evidence

This is to ensure that the digital and physical properties are returned to the owners. The criminal evidence should be kept for further analysis or use in a court of law.. Figure 5 depicts the overall steps of the process.

2.5.2 Analysis

Reith and al. [28] proposal is a major attempt to introduce a general framework that is easily understandable and applicable to different types of digital investigation. In addition, this framework has introduced, for the first time, the concept of investigation

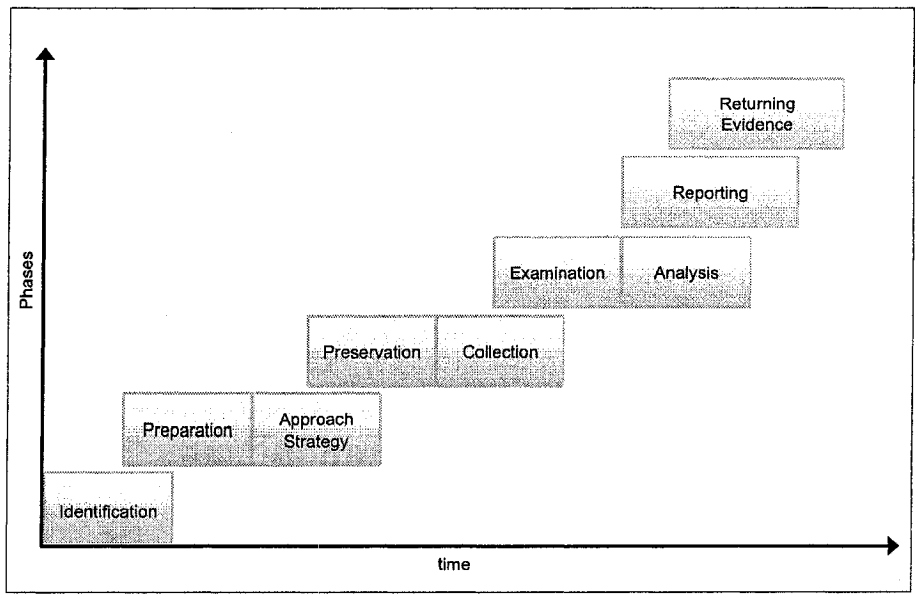

Figure 5: Reith Forensics Process

planning. Moreover, the abstractness of the process make the incorporation of non non-digital technologies possible. Nevertheless, the model is not detailed enough, which may result into difficulties while used in real-life forensic investigation cases. In addition, the model does not explicitly support the chain of custody. Moreover, the model does not account for the iterative nature of some forensics investigation. The proposed model describes a linear process. However, at some points during the analysis phase, the investigator might require the collection of more pieces of evidence.

2.6 Mandia Process

In 2003, Mandia et al. [29]introduced a new forensic investigation process with the belief that it meets the needs of any organization or individual in responding to computer security incidents. The authors also argue that law enforcement personnel

or hired investigators may use the proposed process, understand all of its phases even if what they have to perform, in terms of actions, is just a portion of the entire process.

2.6.1 Process Description

The process phases are as follows:

- Pre-incident preparation
- Incident detection
- Initial response
- Formulation of a response strategy
- Data collection
- Data Analysis
- Reporting
- Resolution

In what follows, we detail each of these phases.

Pre-incident Preparation

Preparation leads to successful incident response and constitutes a proactive measure. Ideally, preparation will involve not just obtaining the tools and developing techniques to respond to incidents, but also taking actions on the systems and networks that might be subjected to an incident, which will make them part of the investigation. At the organizational level, a few strategies is needed to be prepared for a potential attack. This includes the following:

- Implementing host-based security measures
- Implementing network-based security measures
- Training end-users
- Deploying intrusion detection system (IDS)
- Creating strong access control policies
- Performing timely vulnerability assessments
- Ensuring that backups are performed on a regular basis

Incident Detection

If an organization fails in detecting incidents, it cannot succeed in efficiently responding to them. Computer security incidents are normally identified when someone suspects that an unauthorized, unacceptable, or unlawful event has occurred involving the organization's computer networks or data-processing equipment. When the incident is detected, it is very important to record every pertinent fact such as: current time and date, who/what reported the incident, nature of the incident, when the incident occurred, hardware/software involved, etc.

Initial Response

It involves determining the type of incident that has occurred and assessing the impact of the incident. The idea is to gather enough information to begin the next phase, which is developing a response strategy. The other purpose of the initial response phase is to document the steps that must be taken.

Formulation of a Response Strategy

The goal of this phase is to determine the most appropriate response strategy, given the circumstances of the incident. The strategy should take into consideration the political, technical, legal, and business factors that surround the incident. The final solution depends on the objectives of the group or individual that has the responsibility of strategy selection.

Data Collection

This refers to the the accumulation of facts and clues that should be considered during the forensic analysis. The data is collected on the basis of the conclusions. While collecting data, it must be handled in a manner that protects its integrity.

Data Analysis

This includes reviewing log files, system configuration files, trust relationships, web browser history files, email messages and their attachment, installed applications and graphic files. Software analysis is performed as well as a review of time/date stamps, keyword searches, etc. Forensic analysis also includes performing more low-level tasks, such as looking through information that have been logically deleted from the system to determine if deleted files, slack space, or free space contain data fragments or entire files that may be useful to the investigation.

Reporting

The challenge of this step is to create reports that accurately describe the details of an incident, that are understandable to decision makers, that may withstand the barrage of legal scrutiny, and that are produced in a timely manner.

Resolution

The goal of this phase is to provide protection and correction solutions against the security vulnerabilities that have lead to the compromise of a system. These includes, deployment of security appliances such as IDSs, firewalls and restoring the system to its secure operational state. Figure 6 depicts the steps of the Mandia process.

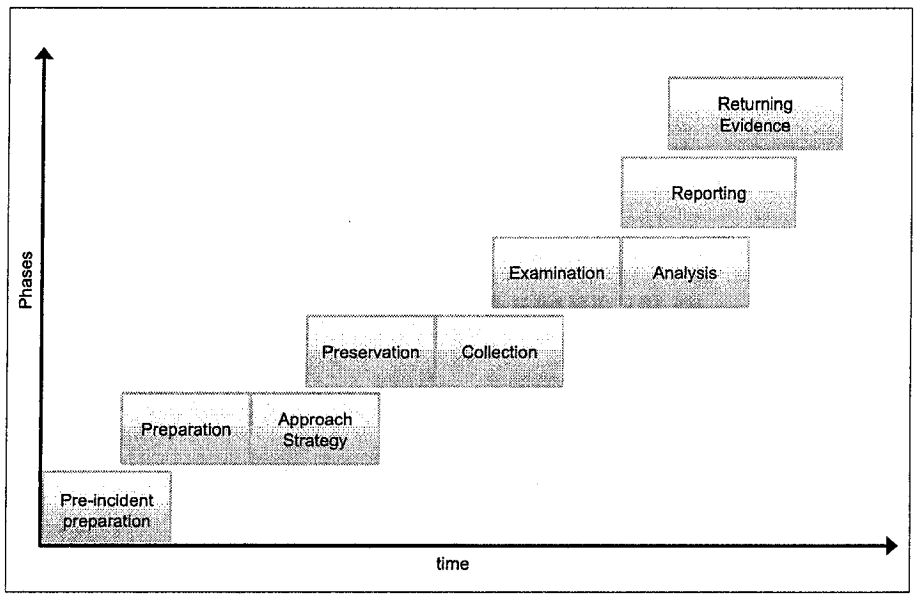

Figure 6: Mandia Forensics Process

2.6.2 Analysis

In compare to the rest of the introduced processes, the Mandia process leverages the advantage of supporting the necessary iterations among the steps though it is not explicitly stated. In addition, the process description contains enough guidelines in the form of questions and checklists that are useful to choose the correct set of actions in each step according to the case. On the other hand, although the process defines a solid and abstract framework for digital investigation, it is not specific enough to

subdivide each phase into more specific steps to guide the investigator through each phase during the investigation. Besides, the process is really less explicit if not silent on several technological aspects.

2.7 Carrier and Spafford Process

The main idea of the Carrier and Spafford process [4] is to rely on the the physical crime scene to elaborate a practical and useful digital forensic process. To this end, they consider the digital crime scene as an entity that has its own witnesses and evidence.

2.7.1 Process Description

The main steps of the the Carrier and Spafford process are:

- Readiness
- Deployment
- Physical crime scene investigation
  - Preservation of the physical scene
  - Survey for the physical scene
  - Documentation of the evidence and the scene
  - Search for physical evidence
  - Physical crime scene investigation
  - Presentation of the complete theory
- Digital crime scene investigation

- Preservation
- Survey
- Documentation
- Search and collection
- Reconstruction
- Presentation

- Review

In what follows, we detail each the aforementioned steps.

Readiness

It is an ongoing phase that is not attached to a particular incident or crime. The purpose of this phase is to assure that the operations and infrastructures are able to support any future investigation. This encompasses operation and infrastructure readiness:

- *Operation readiness:* It is the activity where the people that are involved in the investigation process receives the necessary training and equipments that allow them to conduct efficient and rigorous digital investigations.
- Infrastructure readiness: During this activity, it is determined if there is actual data that exists for future investigation to take place. As Carrier and Spafford specified, this only applies to those who maintain the environment that could be the target of a crime.

Deployment

During this phase, a method is supplied for the detection and confirmation of an incident. This phase consists of the two following sub-phases:

- *Detection and notification:* This refers to the the detection of an incident together with the notification of the appropriate authorities and officials in order to trigger the investigation.

- *Confirmation and authorization:* This refers to the confirmation of the incident and the receipt of the needed authorization to investigate the case and its crime scene.

Physical Crime scene investigation

In this step, the main intent is to collect and analyze the physical evidence and try to reconstruct the actions as precisely as possible that took place during the incident. This is structured according to the following sub-phases:

- *Preservation of the physical scene:* This sub-phase refers to securing of the crime scene, preventing or limiting the entrance/exit of individuals to/from the crime scene as well as helping the ones that deserve assistance by identifying witnesses and detaining any potential suspects.

- *Survey of the physical scene:* At this level, the investigator needs to get an idea on how to handle the physical crime scene and what skills are needed. The investigator and the first responder will take a walk around the crime scene to identify obvious as well as fragile pieces of physical evidence, and subsequently develop an initial theory about the crime.

- *Documentation of the evidence and the scene:* This involves the recording of any detail that may help the investigators in the analysis and reconstruction. Photographs, sketches and videos of the crime scene should be taken together with any evidence found. In the case of a digital incident, it is really important to record the state of the computer.

- *Search for physical evidence:* This involves an in-depth search and collection on the scene for additional physical evidence. Each type of evidence has its own specific collection procedures to follow. When it comes to a digital incident, this step may imply to looking for additional media and digital devices at the crime scene. The physical evidence that is collected from the scene is then sent to the laboratories for analysis and the results are used in the reconstruction phase.

- *Physical crime scene investigation:* During this phase, the investigators elaborate a theory for the incident using the analysis results from the collected physical an digital evidence and the photographs and sketches of the crime scene. With a digital incident, the results of the digital crime scene investigation are correlated with physical evidence to link individuals to the digital events.

- *Presentation of the complete theory:* At this level, the investigators present the evidence and the theory that have been developed from the physical crime scene to a court of law or corporate management.

Digital crime scene investigation

The following sub-phases approach to the computer system and search it for evidence from a crime scene.

- *Preservation:* This involves securing the access to the digital crime scene and preserving the state of the digital evidence. This may include isolating the system from the network, collecting volatile data that could be lost and identifying any suspicious process that is running on the system. Also the suspect users should be investigated who have logged into the system during the incident. It also caters for the creation of duplicates of the evidence from the original versions.

- *Survey:* This involves finding the pieces of digital evidence for the given class of crime. The main purpose of this sub-phase is to find obvious pieces of evidence. It will demonstrate to the investigator the skill level of the suspect and the requirement of analysis techniques.

- *Documentation:* This sub-phase involves documenting Each piece of digital evidence that is found on the crime scene in an accurate and perspicuous way as it has been found.

- *Search and collection:* This sub-phase involves a thorough analysis of the system for digital evidence. This sub-phase uses the results of the survey phase to focus on additional analysis types.

- *Reconstruction:* This sub-phase involves putting the pieces of the digital puzzle together. The digital evidence is classified and is assessed to determine the amount of trust that can be placed on it and also to reconstruct the timeline of events.

- *Presentation:* This sub-phase involves the presentation of the digital evidence that was found to the physical crime scene investigation team. The physical crime scene investigators integrate the results from each of the digital crime

scenes. This phase documents and presents the findings of a specific digital crime scene to other investigators.

Review

This phase involves reviewing the entire investigation to identify areas of improvement. This includes reviewing the quality of each of the physical and digital investigations and evidence that contributed to solve the case. Figure 7 depicts the steps of the Carrier and Spafford process.

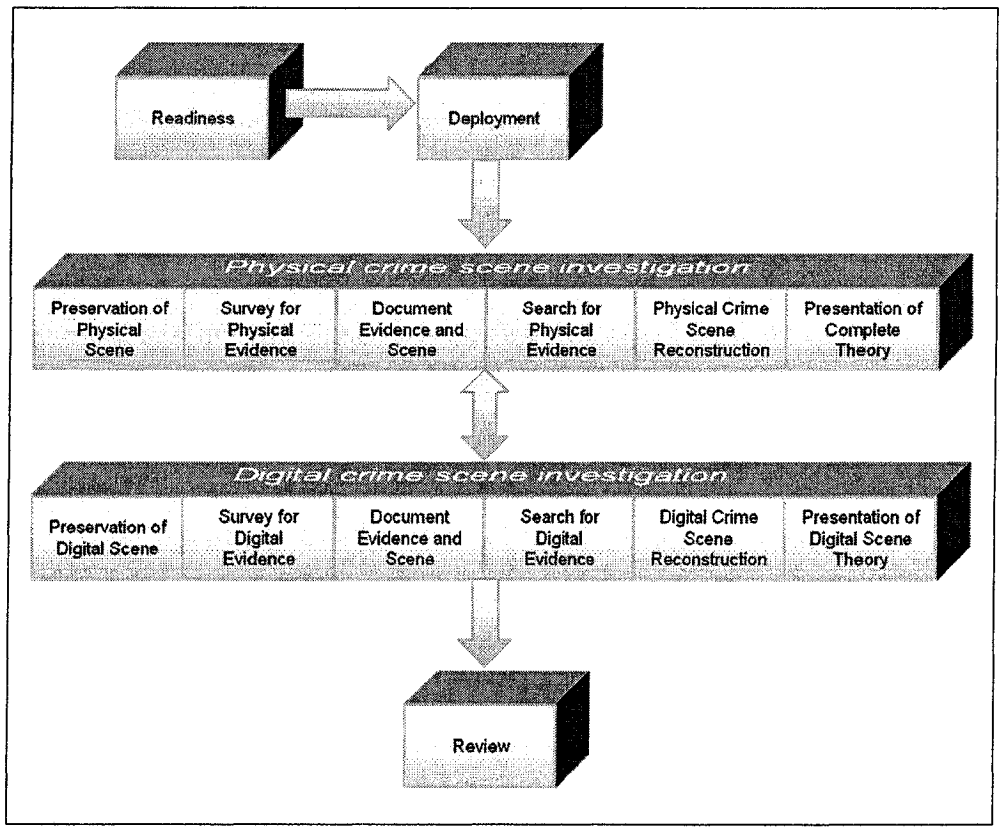

Figure 7: Carrier and Spafford Forensics Process

2.7.2 Analysis

One of the major contributions of the Carrier and Spafford process is the definition of a digital investigation model that combines the physical and digital crime scenes in the investigation process. Another important innovation of this model is the definition of a review phase to asses the quality of the investigation. This is an effective measure to improve the quality of future investigations. Moreover, this model deals with the infrastructure readiness in the investigation process to ensure that enough evidence exists for a proper and effective investigation to take place. Nevertheless, it should mentioned that the Carrier and Spafford process has some limitations. For instance, it does not support the feedback and iteration concepts that are needed in real-life investigations. A crime investigation is an iterative process during which analysis of existing evidence leads to or requires the collection of new pieces of evidence. Therefore, the process should support iteration between the analysis phase and collection phase and in some cases of the authorization phase.

2.8 Ó Ciardhuáin Process

In 2004, Séamus Ó Ciardhuáin proposed a new a new cyber-forensic investigation process [1] that combines the existing models and adds new activities that were not addressed in the previous models. This model captures the full range of the investigation activities in an elaborated way including those that are pertinent before evidence processing.

2.8.1 Process Description

The model's steps are as follows:

- Awareness

- Authorization
- Planning
- Notification
- Search and evidence identification
- Collection of evidence
- Transport of evidence
- Storage of evidence
- Examination of evidence
- Hypothesis
- Presentation of hypothesis
- Proof/defense of hypothesis
- Dissemination of information

Awareness

This refers to the creation of an awareness that an investigation is actually needed. Such an awareness is generally created by events external to the organization, which will carry out the investigation, e.g. a crime is reported to the police or an auditor is requested to perform an audit. It may also result from internal events, e.g. an intrusion detection system alerts to a system administrator that a system's security mechanism has been compromised. This step is important to ensure that the correct approach is taken to conduct an investigation in a particular context.

Authorization

Obtaining the needed authorization to conduct an investigation may be very complex and require an interaction between external and internal entities to the organization. Authorization varies depending on the type of investigation. Sometimes, a system administrator may require only a simple verbal approval from company management to carry out a detailed investigation of the company's computer systems. In other contexts, law enforcement agencies usually require a formal legal authorization (e.g. court orders or warrants) that sets out, in precise terms, what is permitted in an investigation.

Planning

Investigation planning is strongly influenced by information from both; inside and outside the investigating organization. From outside, the plans will be influenced by regulations and legislation, which set the general context of the investigation and that are not under the control of the investigators. There will also be information collected by the investigators from other external sources. Within the organization, there will be the organizational strategies, policies, and information about previous investigations. The planning activity may rise a need to backtrack and obtain further authorization, for example when the scope of the investigation is found to be larger than what has been originally planned for.

Notification

This refers to informing the subject of an investigation or other concerned parties that the investigation is taking place. This activity may be inappropriate in some investigations while required for others.

Search and Identification of Evidence

This phase refers to the localization of the evidence together with its identification.

Collection of Evidence

In this phase, the investigating organization takes possession of the evidence in a form that can be preserved and analyzed, e.g. imaging of hard disks or seizure of entire computers. Any error or poor practice, at this stage, may result into a useless evidence, particularly in investigations that are subject to strict legal requirements.

Transport of Evidence

Evidence must be transported to a suitable location for later examination. This could be simply the physical transfer of seized computers to a safe location. However, it could also be the transmission of data through networks. The means of transport used have to preserve the integrity of the evidence.

Storage of Evidence

The collected evidence is often required to be stored because examination cannot take place immediately on the crime scene. Always, the integrity of the evidence should be preserved.

Examination of Evidence

This phase involves the use of a potentially large number of techniques to find and interpret significant data. It may require the repair of damaged data in ways that preserve its integrity. Sometimes, when there are very large volumes of data to be examined, automated techniques to support the investigator are required.

Hypothesis

Based on the examination of the evidence, the investigators have to construct a hypothesis on the occurrence of incident. Backtracking from this activity to the examination activity is to be expected, as the investigators develop a greater understanding of the events, which led to the investigation in the first place.

Presentation of Hypothesis

The hypothesis must be presented to persons other than the investigators. In a police investigation, the hypothesis will be placed before a jury, while an internal company investigation will place the hypothesis before management for a decision/action to be taken.

Proof/Defense of Hypothesis

Generally, the hypothesis will not go unchallenged. A contrary hypothesis and supporting evidence will be placed before a jury. The investigators will have to prove the validity of their hypothesis and defend it against criticism and challenge. Successful challenges will probably result in backtracking to the earlier stages to obtain and examine more evidence, and construct a better hypothesis.

Dissemination

This phase refers to the partial/total dissemination of the results of the investigation. Some information may be made available only within the investigating organization, while other information may be more widely disseminated. Policies and procedures will normally determine the details of the dissemination activity. The collection and maintenance of this information are the key aspects of supporting the work of investigators and are likely to be fruitful areas for the development of advanced applications

incorporating techniques such as data mining and expert systems. Figure 8 depicts the overall steps of the Ó Ciardhuáin process.

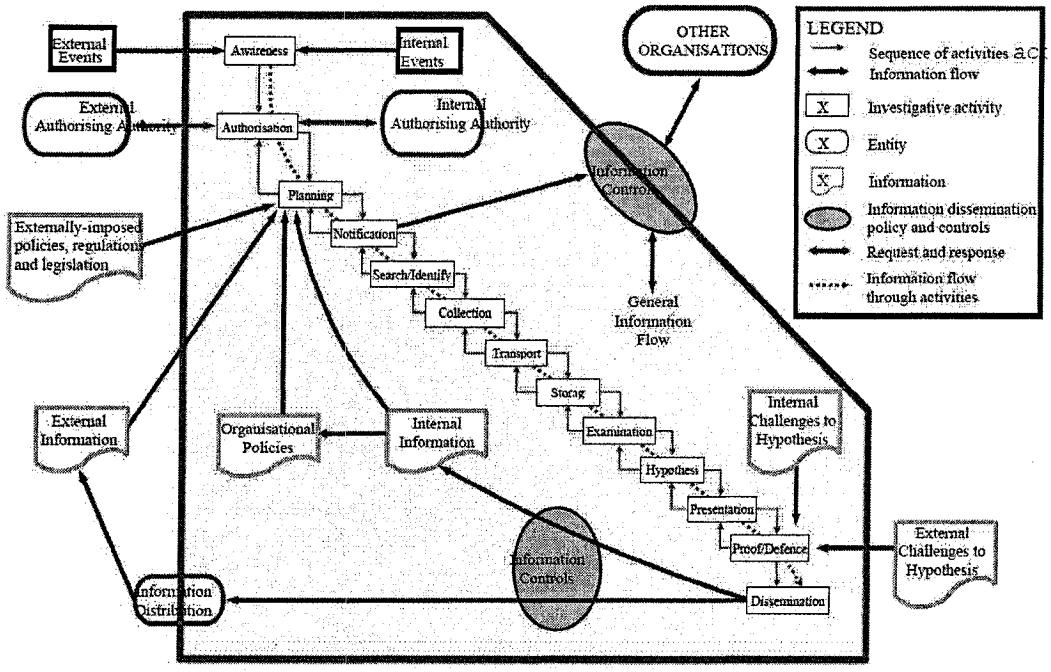

Figure 8: Ó Ciardhuáin Forensic Process

2.8.2 Analysis

The proposed process is a great proposition that included, for the first time, investigation planning as well as information dissemination among the model phases. In addition, this model is tightly pertinent to real-life cases by accounting for hypothesis proof and defence. One additional feature of this process is the explicit iteration through the backtracking of information that flows from each step to the previous one. As of the limitations of this process, it is important to mention that since the process follows a waterfall-like model, it suffers from the deficiencies of this model. In spite

of the fact that the process contains authorization and planning phases, at the beginning of the process model, it not always easy/possible to predict on all the needed resources (e.g. personnel, authorizations). Besides, the waterfall model supposes that after one phase finishes, the next phase begins. In some real-life scenarios the investigation does not have a linear model. For instance, while some actions related to the analysis phase are performed, new evidence might be needed and therefore the underlying authorizations become required. It not possible to account for all these in the initial planning phase. Furthermore, due to the nature of the waterfall model, false assumptions are detected very late such as during the hypothesis proof/defense. However, if these flaws are detected early during the investigation, the cost and even the result of the investigation may dramatically change.

2.9 Casey and Palmer Process

The investigative process, presented by Casey and Palmer in [7], is structured to encourage a complete, rigorous investigation, ensure proper evidence handling, and reduce the chance of mistakes created by pre-conceived theories and other potential pitfalls. This model provides investigators and examiners with a logical flow of events that, taken together, seek to provide:

- Acceptance: The steps and methods have earned professional consensus.
- Reliability: The methods employed can be proven (trusted) to support findings.
- Repeatability: The process can be applied by all, independent of time and place.
- Integrity: The state of evidence is proven (trusted) to be unaltered.
- Cause and effect: The process caters for logical connection between suspected individuals, events, and exhibits.

- Documentation: The process caters for recordings that are essential for testimonial evidence (expert testimony).

2.9.1 Process Description

The process phases are as follows:

- Incident alerts or accusation
- Assessment of worth
- Incident/crime scene protocols
- Identification or seizure
- Preservation
- Recovery
- Harvesting
- Reduction
- Organization and search
- Analysis
- Reporting
- Persuasion and testimony

Incident Alerts or Accusation

This step can be signaled by an alarm from an intrusion detection system, a system administrator reviewing firewall logs, curious log entries on a server, or some combination of indicators from multiple security sensors installed on networks and hosts.

This initial step could also be triggered by events in more traditional law enforcement settings like citizens reporting possible criminal activity. When presented with an accusation or automated incident alert, it is necessary to consider the source and reliability of the information. Also, some initial fact gathering is usually necessary before launching a full-blown investigation.

Assessment of Worth

The handling of this step in the process varies with the associated investigative environment. Applied in law enforcement environments, all suspected criminal activity must be investigated. In civil, business, and military operations, suspicious activity will be investigated but policy and continuity of operations often replaces legalities as the primary concern. Regardless of environment, a form of triage is performed at this step in the process.

Incident/Crime Scene Protocols

Protocols, practices, and procedures are employed to minimize the chance of errors, oversights, or injuries. Whoever is responsible for securing a crime scene, whether first responders or digital evidence examiners, should be trained to follow accepted protocols. The output of this step is a secure scene where all the contents are mapped and recorded, with accompanying photographs and basic diagrams to document important areas and items.

Identification or Seizure

Once the scene is secured, potential evidence of an alleged crime or incident must be seized. Clear procedures and understanding of necessary legal criteria are essential before activity can proceed successfully. The goal here for trained and experienced

investigators is not to seize everything at a scene (physical or virtual) but to make informed, reasoned decisions about what to seize and be prepared to document and justify their actions.

Preservation

Working from the known inventory of confiscated or seized components investigators must make sure that potentially volatile items remain unchanged. First, the original material is catalogued and stored in a proper environmentally controlled location, in an unmodified state. Second, an exact copy of the original material that will be scrutinized as the investigation continues.

Recovery

Prior to performing a full analysis of preserved sources of digital evidence, it is necessary to extract data that have been deleted, hidden, camouflaged, or that are otherwise unavailable for viewing using the native operating system and resident file system. In some instances, it may also be necessary to reconstitute data fragments to recover an item. Whenever feasible, this process is performed on copies of original digital evidence from the preservation step.

Harvesting

This stage in the process is where the actual reasoned scrutiny begins, where concrete facts begin to take shape that support or falsify hypotheses built by the investigative team. The investigator will look for categories of data that can be harvested for later analysis (groupings of data with certain class characteristics that, from experience or training, seem or are known to be related to the major facts of the case or incident).

Reduction

It involves activities that help eliminating or targeting specific items in the collected data as potentially germane to an investigation. The decision to eliminate or retain is made based on external data attributes such as hashing or checksums, type of data, etc.

Organization and Search

To facilitate a thorough analysis, it is advisable to organize the reduced set of material from the previous step, grouping, tagging, or otherwise placing them into meaningful units. It may be advantageous at this stage to actually group certain files physically to accelerate the analysis stage.

Analysis

It involves the detailed scrutiny of data identified by the preceding activities. The techniques employed here will tend to review and study of specific, internal attributes of the data such as text and narrative meaning of readable data, or specific format of binary audio and video data items. Additionally, classes and individual characteristics found in this step are used to establish links, determine the source of items, and ultimately to locate the offender.

Reporting

Final reports should contain important details from each step, including reference to protocols followed and methods used to seize, document, collect, preserve, recover, reconstruct, organize, and search key evidence. The majority of the report generally deals with the analysis leading to each conclusion and descriptions of the supporting evidence and analysis.

Persuasion and Testimony

In some cases, it is necessary to present the findings in a report and address related questions before decision makers can reach a conclusion. A significant amount of effort is required to prepare for questioning and to convey technical issues in a clear manner. This step includes techniques and methods used to help the analyst and/or domain expert translate technological and engineering details into understandable narrative for discussion with decision makers. Figure 9 depicts the overall steps of the Casey and Palmer forensic process.

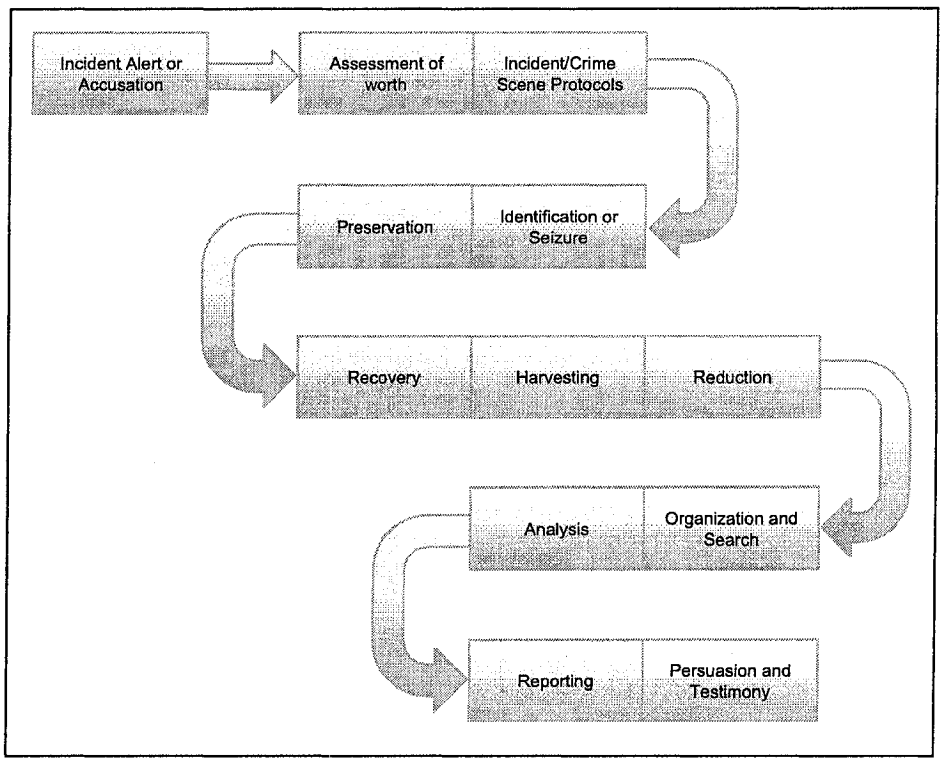

Figure 9: Casey and Palmer Forensic Process

2.9.2 Analysis

This process leverages an edge over the rest of the processes in two areas. First, it introduces the persuasion and testimony phase, which is lacking in the rest of digital forensic process models. In fact, successful completion of this phase is of paramount importance to the success of the case. Any deficiency in this phase will directly lead to the complete failure of the case if presented to a court of law. Second, the process has particularly subdivided the collection and analysis phases into sub-phases, which will cause the investigation to follow a more manageable and robust step-by-step flow. As for the limitations of this process, it is important to mention that the porposed model, does not support the iterative nature of forensic investigations. As it has been noted before, during the forensic investigation, as the analysis phase proceeds, there might be a need for new pieces of evidence, changes in the plan or authorizations. However, the process fails to provide some evolutionary or feedback based mechanism to support this iterative nature of forensic investigation. In addition, the process does not have any provision for the preparation phase, which is of crucial importance to any forensic investigation. Without, equipments, training and sufficient auditing data, many investigations will fail simply due to the lack of enough supporting evidence.

2.10 Beebe and Clark Prcess

Beebe and Clark proposed in [3] a framework that simplifies the complex process of digital forensic investigation, yet provides a mechanism for including layers of detail that are needed by its users. Their primary goal was to ensure the framework's expansion capability while integrating previous frameworks and models. The process consists of phases and principles. Phases and sub-phases are the steps that the investigator should take during the investigation in a pre-specified order. On the other

hand, principles are a set of objectives that should be achieved during the whole investigation or throughout a subset of phases.

2.10.1 Process Description

The proposed process is made out of the following steps:

- Preparation
- Incident response
- Data collection
- Data analysis
- Presentation of findings
- Incident Closure

The principles that are defined over all the investigations are:

- Evidence preservation
- Documentation
- Proper investigative authority
- Sensitivity and classification
- Investigative priority
- Information flow and controls
- Case management
- Process improvement feedback

In what follows, we will detail each of these steps.

Preparation

Thoughtful preparation can improve the quality and availability of the digital evidence collected, while minimizing organizational cost and burden. Preparation activities include, but are not limited to:

- Assess risk considering vulnerabilities, threats, loss/exposure, etc.
- Develop an information retention plan for both pre and post-incident
- Develop an incident response plan, including policies, procedures, personnel assignments, and technical requirements definition
- Develop technical capabilities (e.g. response toolkits)
- Train Personnel
- Prepare host and network devices
- Develop evidence preservation and handling procedures
- Develop legal activities coordination plan for both pre/post-incident

Incident Response

This phase consists of the detection of initial, pre-investigation response to a suspected computer crime related incident, such as: a breach of computer security, the use of a computer to view contraband material, etc. The purpose of this phase is to detect, validate, assess, and determine a response strategy for the suspected security incident.

Data collection

Data and information required to validate an incident and determine its impact that will be initially collected during the incident response phase. Once a decision has been

made to investigate the legal or administrative actions, the formal data collection phase ensues. The purpose of that phase is to collect digital evidence in support of the response strategy and investigative plan.

Data Analysis

It is the most complex and time consuming phase in a digital investigation process. The goal of this phase is confirmation analysis (to confirm or refute allegations of suspicious activity) and/or event reconstruction (answer "who, what, where, when, why and how" type questions). Collected data is surveyed, extracted, and reconstructed during the data analysis phase.

Presentation of Findings

The aim of this phase is to communicate relevant findings to a variety of audiences, including management, technical personnel, legal personnel, and law enforcement. The presentation(s) are intended to provide both succinct and detailed confirmatory and event reconstruction information regarding the data examined in the data analysis phase.

Incident closure

It focuses on closure of the investigation. However, it is important to not only close out this investigation and act upon decisions related to it, but also to attempt to preserve knowledge gained to enhance subsequent investigations. Figure 10 depicts the overall steps of the Beebe and Clark process.

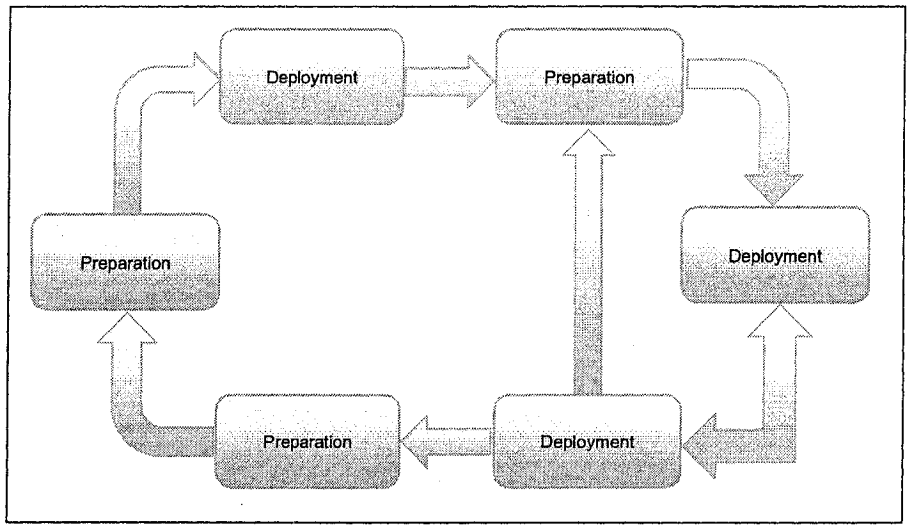

Figure 10: Beebe and Clark Forensic Process

2.10.2 Analysis

The process enhances the previous process models, as described above, in many directions. First, it introduces the nice concept of forensic principle. As Beebe and Clark have mentioned, the process of the forensic investigation is an objective-based process rather than the execution of a sequence of tasks in a pre-specified order. This is due to the specificity and uniqueness of each investigation that necessitate a non-checklist and abstract framework satisfying the needs of particular judicial systems and digital crime scenes. Second, the Beebe and Clark process introduces the phase of incident closure that contains the actions, which should be taken before closing the case in order to review the quality of the investigation, document the effective techniques, dispose of some pieces of evidence and apply the necessary corrective measures. Although the process introduces some new phases, it still lacks the planning phase. Moreover, the process does not support the iterative nature of forensic investigations. Although the analysis phase has been described to be iterative, no considerations are

provided for cases analysis that may require new pieces of evidence.

Chapter 3

Comparative Study of Digital Forensic Processes

This chapter is entirely dedicated to comparing the previously evaluated models with respect to a variety of metrics. The basic idea is to identify the nice features and strong points of theses processes as well as their deficiencies, limitations and flaws. The intent is to design a new forensic process by leveraging those proposed nice features, while avoiding the current flaws and fixing the corresponding deficiencies. On the other hand, such approach would also provide appropriate guidelines for both researchers and users to adopt adequate process models under different conditions. Indeed, digital forensics process models thoroughly examined in the previous chapter incorporate procedures and techniques designed to effectively tackle different situations related to illegal activities commonly referred to as cyber crime. They cover the most widely accepted tools in providing a consistent and standardized framework that supports all stages of a digital investigation.

Tailored to specific tasks in the practice of digital forensics, these models exhibit many differences in terms of definition, scope, and concepts. As stated above, this

study introduces a methodological approach designed to provide the ability to conduct a comprehensive comparative evaluation of different models set to drive digital investigative processes. Such evaluation framework refers to a set of thoroughly selected metrics representing requirements and critical features that characterize the digital forensics models. Building a well-defined, robust, and comprehensive comparative framework is thus essential to highlight commonalities and differences, emphasize weaknesses and strengths in existing digital forensics models, and finally address appropriate research directions. Indeed, the proposed comparative approach stems primarily from our own experience with process modeling technologies. It is also the result of both a comprehensive literature review and brainstorming sessions. The resulting requirements and features are broadly classified into seven categories below. Figure 3 depicts the overall criteria on which we do the comparison of the proposed processes. In what follows, we discuss each of these criteria.

Predisposition: This criterium reveals the impact of extra investigation process actions on conducting a full high-performance investigation in different environments and organizations. Such actions are not part of the cyber forensic investigation process. However, they determine the natural aptitude of a system to be investigated and an investigator to act correctly in specific circumstances. In turn, predisposition incorporates such requirements and features as:

- *Hardening*: This feature represents all specific procedures that system administrators and incident response teams take as a measure intended to protect a given system and handle reported incidents. Indeed, an organized and controlled environment is susceptible of helping an investigator to effectively achieve good results.

- *Training*: This is a vital requirement for investigators to focus on acquiring sufficient knowledge and basic skills to enable them to adequately perform specific

	Predisposition	Functionality						Admissibility	Architecture	Usability	Quality Assurance
		Collection	Examination	Analysis	Reporting	Testimony	Case Closure				
DFRWS	Identification Class elements	Collection Class elements	Examination Class elements	Analysis Class elements	Presentation Class elements	Expert Testimony	- Mission Impact -Recommended Countermeasures - Statistical Interpretation	- Preservation class - Legal authorization - Traceability	- Linear - Evolving	- General - Tool support (DIFL)	None
EEDI	Based on DFRWS	Collecting evidence	Based on DFRWS	- All process steps	Based on DFRWS	Based on DFRWS	Based on DFRWS	Based on DFRWS	- Partially Iterative	- Specific - DIFL language	None
DOJ	- Personnel training - Investigative Tools and Equipment	Evidence collection	Forensic examination (not detailed)	Analysis (not detailed)	Examination report	Testimony on examination tasks	None	- Search warrant - Evidence preservation - Chain of Custody	None	- General - Prescriptive - No tool	None
Reith et al.	Identification	Data collection	Examination step	Analysis	Presentation	None	Return evidence	- Search warrant - Authorizations - preservation	- Evolving	- General - No tool	None
Mandia et al.	Pre-incident preparation	Data collection	None	Analysis	Reporting to decision makers	None	Resolution	- Chain of custody - Authentication - Integrity - Accreditation	- Parallelism - Partially Iterative	- General - Prescriptive - No tool	None
Carrier & Spafford	Readiness	- Survey phases - Search and collation phases (Physical/digital)	None	Reconstruction	Reporting between investigators	None	- Review phases	- Search warrant - Chain of custody - Integrity	- Partially Iterative (data collection – Reconstruction) - Parallelism (physical/digital)	- General - Prescriptive - No tool	None
Seamus	Awareness	- Search and identification of evidence - Collection	Examination	Hypothesis	Presentation	Proof/Defense	Dissemination	- Authorization - Evidence Integrity - No chain of custody	- Iterative	- General - Prescriptive - No tool	None
Casey & Palmer	Accusation or incident alarm	- Identification or seizure - Recovery	- Harvesting - Reduction - organization and search	Analysis	Reporting	Testimony	None	- Legal criteria - Preservation	Parallelism (Case management)	- General - Prescriptive - No tool	None
Beebe & Clark	Preparation	Data collection	None	Data analysis	Presentation of findings	None	Incident closure	- Legal authority - Preservation - Chain of Custody	- Parallelism (Principles apply to all phases) .	- General - Prescriptive - No tool	None

Figure 11: Process Comparison Table

tasks regarding digital forensic investigation.

- *Tools and equipment*: An important feature regarding digital forensic models lies in the need to provide investigators with appropriate, updated, and improved techniques, software and hardware required to properly gather, analyze, and handle evidence.

- *Social networks*: Perhaps one of the most interesting and useful features of any digital investigation is the social network that an investigator can use to pursue particular actions. Indeed, it is highly recommended for an investigator to develop and maintain a strong collaboration with national and international specialized agencies. Such collaboration provides the ability to deal with different aspects and issues inherent to the transborder nature of cyber crime.

Functionality: Digital forensics investigation processes and tasks are required to provide powerful and effective functionalities in the course of challenging activities performed to identify, gather, analyze, and present evidence. Indeed, the purpose of a process model of digital investigation is to allow an investigator to perform a variety of functions in accordance with established standards and legal requirements. In this way, the investigator can produce outputs that meet the special challenge for the admissibility of digital forensic evidence in a court of law. In this respect, the core functions of a cyber forensic model can be clearly identified as:

- *Collection*: Collection or acquisition of data from a source is the most significant function of a digital forensics investigation. While considered as a vulnerable process, it allows investigators and examiners to first identify and then acquire relevant data that may contain evidence, and finally guarantee its integrity at a site of incident.

- *Examination*: Once the necessary data has been collected and transported to a digital forensic laboratory, the next step lies in the examination of digital evidence. Using accepted forensic procedures, examination is thus a critical function designed to provide the ability to identify, recover, and extract digital evidence.

- *Analysis*: Analysis is an essential function through which a forensic investigator derives useful information. It refers to the process of interpreting the extracted data to reconstruct a chain of evidence and draw appropriate conclusions.

- *Presentation*: Presentation of evidence determines the outcome of a digital forensic investigation. It refers to the important phase where the investigator assumes the arduous task of completely and accurately reporting his or her findings and the results of the digital evidence analysis. It also encompasses the prosecutor's presentation of the digital evidence and testimony during trial in court. In this respect, a well-documented case is much more likely to be admissible. Hence, it is essential that the investigator keeps a complete, accurate, and comprehensive documentation of all activities and actions he or she performed during the investigation.

Case Closure: The final phase of a digital forensic investigation may include different steps designed to focus, when appropriate, on closure of the case under investigation. In this context, the challenge facing the investigator stems from the strategic need to enhance subsequent investigations based on knowledge and experience acquired. Hence, the final phase may well incorporate different activities designed to enable the investigator to:

- Preserve all information related to the case.

- Identify lessons to be learned.

- Submit appropriate recommendations.
- Undertake all actions designed to efficiently disseminate critical information pertaining to problems encountered and best practices to authorized personnel and agencies.

Admissibility: Admissibility of digital evidence in court is both a challenging legal issue and an important requirement. Indeed, the ultimate purpose of conducting a digital forensic investigation is to offer consistent evidence. In other terms, each process, task, and action must be organized and performed to ensure and demonstrate the validity, integrity, and reliability of digital evidence.

- *Authorization*: A legal authorization, typically in the form of a search warrant, may be required prior to conducting a digital forensic investigation. Once the authorization is granted, an investigator can have access to or seize a device or system, collect evidence, or conduct interviews. In the presence of critical systems, additional permission may, however, be sought before a full analysis can be conducted.
- *Preservation*: Preservation involves all measures designed to ensure the accuracy and integrity of digital data and evidence across the whole investigative process.
- *Chain of custody*: Mistracking key data and evidence would compromise the outcome of a digital forensics investigation. In this respect, establishing and maintaining a proper and continuous chain of custody represents a significant step towards preserving the integrity of digital data and validating how evidence has been gathered, tracked and protected.
- *Relevance*: A legal requirement governing the admissibility of extracted and analyzed digital evidence lies in the relevance of such evidence to the case under investigation.

- *Accreditation*: National and international efforts have specifically addressed the issue regarding digital evidence. These efforts are directed towards achieving the harmonization of methods and practices. In 1997, the G-8 countries have initiated an action plan recommending that nations "develop and employ compatible forensic standards for retrieving and authenticating electronic data for use in criminal investigations and prosecutions." Tasked to develop such standard, the International Organization on Computer Evidence (IOCE) later proposed different guidelines regarding best practice, training, and quality assurance and laboratory management. In addition, different non-profit organizations have been involved in similar efforts and finally established programs "designed to provide a toolkit of templates and reference documents that can be utilized by laboratories to develop and implement a quality system meeting the requirements of accreditation." In this context, the concept of accreditation has emerged as a vital requirement for laboratories engaged in activities related to digital forensics.

Usability: Determines the degree of easy to use of the process and how well investigators can use the process to perform investigation's tasks with effectiveness, efficiency and satisfaction.

- *General*: The process should accommodate different realistic situations and different contexts.
- *Prescriptive*: The process should provide and indicate the appropriate and detailed procedures, guidelines, rules and techniques, which would lead to the desired goal.
- *Tools support*: Indicates whether the process is backed by tools that implement the functionalities and provide a convenient method of viewing and updating

information on ongoing investigation.

Architecture: The structure of the digital forensics process should fit the purpose. It allows traceability from the high level steps down to the underlying technology.

- *Iteration*: An important feature of a cyber forensic investigation lies in the continual need to recollect data, refine examination, and reconstruct the chain of evidence. In order to address this requirement, a comprehensive digital forensics model must incorporate interacting phases designed to enable iteration between different processes.

- *Evolution*: Evolution refers to the attribute of a model of digital investigation that is abstract enough, open, adapted to new technologies, and applicable to different digital cases and environments. In other terms, a robust digital forensics model is designed to provide a consistent and standardized framework that supports all phases of an investigation, regardless of the current technologies, the specific type of cyber crime, and the environment.

- *Parallelism*: Parallelism is another important feature of a digital forensic model architecture designed to provide mechanisms for parallel execution of complex investigations, while maintaining a predefined synchronization to achieve consistent results.

Quality assurance: It is commonly acknowledged that digital evidence is both fragile and vulnerable. Often described as "invisible and easily altered or destroyed", digital evidence requires "precautions to prevent alteration, special tools and equipment, specialized training, and expert testimony". In this respect, the idea of introducing best practices was first discussed in 1997 at the meeting of the International Organization on Computer Evidence (IOCE) held in The Hague, The Netherlands. While

there is a clear consensus within the area of digital forensic investigation - a field of growing complexity - regarding the meaning and importance of best practices, efforts are, on the other hand, directed towards bringing together different concepts to provide an adequate framework of best practices. As a result, quality assurance has emerged as a vital ingredient to meet the combined requirements of control, accuracy, and reliability of digital evidence. A quality assurance program incorporates different mechanisms designed to provide the ability to continuously monitor, control and, when appropriate, improve quality. Quality is planned, created, improved, and finally delivered through a systematic effort. In order to create quality, law enforcement and forensic organizations and agencies must established and maintain a "quality system". Using the graphical model illustrated in Figure 12. A quality system can thus be defined as "a set of interdependent processes that function harmoniously, using various resources, to achieve the objectives related to quality". According to the ISO 8402 (1994), a quality system is "organizational structure, procedures, processes and resources needed to implement quality management". The concept of quality extends the administrative mission to include such activities as the measurement and control of quality at every phase from preparation and case closure, through evidence examination, analysis, and presentation. Figure 12 depicts the overall criteria of the process.

As already stated above, the analysis presented in Chapter 2 triggered the motivation to compare prevalent digital forensic investigation models. In substance, these models exhibit distinct characteristics and inherent advantages and disadvantages, despite their striking similarities. Efforts to conduct a comprehensive comparison of existing models of digital investigation were partly dedicated to drafting an initial list of requirements and attribute features. This list has been constantly refined, extended, validated, and finally adopted. Based on these requirements and attribute

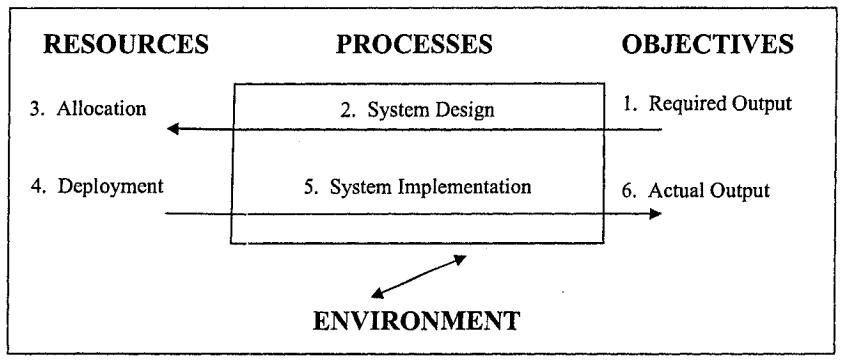

Figure 12: Graphical Model of a System (Willborn and Cheng, 1994)

features, Table 3 provides a comparison of the activities, tasks, and processes corresponding to each framework of digital investigation. At first glance, it can be seen that each model adopts and incorporates a set of pre-investigative measures or tasks, such as crime detection, personnel training, investigative tools and equipment, procedures, policies, incident response planning, etc. While they are not investigative by nature, these activities are commonly considered to greatly influence the outcome of a digital investigation, in terms of success or failure.

A common inherent feature characterizing these frameworks lies in the presence of all basic functionalities that form the core of a cyber forensic process model: Collection, examination, analysis, and reporting. It reveals a general trend towards building and implementing standardized frameworks of investigation designed to support and deal with all types of cyber crime. Thus, activities, tasks, and processes pertaining to specific types of crime are incorporated into the same phase or sub-phase to support investigators in charge of minor internal incidents or law enforcement agents in search of digital evidence destined to a court of law. In this respect, it can be argued that such an approach would create considerable confusion that compromises the admissibility of digital evidence. In addition, admissibility represents another distinguished characteristic pertaining to every model of digital investigation. Indeed,

admissibility - as a legal issue - addresses the challenging question of whether evidence resulting from a digital forensic investigation can be introduced at all to a court of law. Thus, appropriate actions are performed at each step of the investigation process in pursuit of the ultimate goal of securing the legal admissibility of digital evidence. Hence, admissibility-related components such as legal authorization, search warrant, evidence preservation and integrity, and chain of custody are incorporated in almost every proposed framework to address legal issues.

While the DOJ-based framework is unique in providing features tailored to support specific types of investigative actions, the remaining digital forensics models of investigation can be properly described as general. This attribute feature refers to the degree to which an investigative platform is able to expand its usefulness. Indeed, general models of digital investigation provide - through phases consisting of multiple processes - the ability to perform different types of investigative actions and processes: Forensic readiness process, incident response process, digital investigation process, and digital forensic investigation process. Differences in the definition and understanding of these concepts contribute significantly to the reigning of confusion in this new area of research. In seeking to identify these ill-defined terms, it may be possible to promote a common understanding of important concepts and thus enhance communication amongst researchers. Indeed, forensic readiness may be defined as the ability of an organization to maximize its potential to use digital evidence, whilst minimizing the cost of an investigation [40]. Evidence may thus be collected well in advance of a crime to help a given organization in deterring computer crime and responding to successful attacks. In addition, forensic readiness-related actions may also include a monitoring scheme to detect and deter major incidents, a policy for securing storage and handling potential evidence, a staff training program in incident awareness, and finally a set of evidence collection requirements [40]. On the

other hand, the incident response process refers to decisions, procedure elements, and actions related to computer security incident handling within an organization. It is generally performed by a Computer Security Incident Response Team (CSIRT) engaged in handling security violations and breaches intended to disrupt confidentiality, integrity, and availability of information systems. Depending on the nature of the incident, members of the team may be required to communicate with and assist law enforcement agents, particularly in case cyber crime perpetrators are prosecuted [44].

Furthermore, the digital investigation process may be defined as a set of actions or procedure elements intended to address "questions about digital states and events" [2]. This type of investigation focuses on corporate security incidents. Its goal is solely to investigate a given incident and present findings to managers and decision-makers. This investigation is more often administrative in nature. The final decision would, in this context, be oriented towards fixing the vulnerability of an information system, punishing the offender, etc. Finally, the digital forensic investigation process includes all procedure elements and appropriate actions undertaken in response to a violation of the law. The ultimate goal of such actions is to present admissible digital evidence to a court of law. The tasks pertaining to different steps are not, in this case, homogeneous and fail to impose a clear separation between roles attributed to different actors (system administration, security officer, computer security incident response team, digital forensics examiner, digital forensics analyst, digital forensics and legal prosecutor) involved in a digital forensic investigation process [17].

Regarding these definitions, the objectives are different from one model to another, whereas the tasks executed in different steps remain the same. As a result of this multi-views approach adopted by almost all the models, the phases may be not homogenous, for instance a phase can be viewed from a preventative information security perspective where there is little need for digital evidence, a second phase viewed from

a business perspective where collecting appropriate digital evidence would be beneficial and a third phase viewed from law enforcement view requires an investigation conducted in a systematic, formalized and legal manner to ensure the admissibility of the evidence [40]. Testimony may be required to explain the conduct of the investigation. Thereof, there is no unique objective of the process and no clear separation between roles of the different persons involved in a digital forensic investigation process (system administration, security officer, CSIRT, digital forensics examiner,digital forensics analyst, digital forensics and legal prosecutor) [17]. The examples that illustrate this confusion are the examination and analysis phases, reporting and testimony. The examination phase is mainly technical, its action are performed by the digital forensic examiner. Analysis differs from examination in that it refers to the interpretation of examination recovered data and putting it in a logical and useful format [47], it is performed by the digital forensic investigator. Some models such as Mandia et al. [29], Carrier [4], and Beebe [3] simply ignored this step. They transferred operations to the analysis step with no regard to the roles of the investigative members team. The reporting and testimony are two other confused tasks. Most of the models, for instance Reith et al. [28], Mandia et al. [29], Carrier [4], Casey [7], and Beebe [3] provide for reporting and presentation to decision makers or between investigators team and no provisions to the court presentation. Except DFRWS, DOJ and Seamus models, the other do not treat the testimony issue. Therefore, from the enforcement law view those models do not fit the purpose of complete digital forensic investigation process. Furthermore, it is appropriate to note that the digital forensic frameworks presented herein give little consideration to important features such as architecture,usability and quality issues.

Although the proposed digital forensics process models have evolved to become increasingly enhanced since the first proposed framework DFRWS, they still lacking

in many ways. This fact incites us to propose a new cyber forensic investigation process model to set right the shortcomings observed along the previous analysis and comparative study. The new cyber forensic investigation process model should be seen from a law enforcement perspective and with the major objective: present admissible finding to a court of law. The other views can be considered as sub-perspective with limited objectives.

Chapter 4

Towards a Unified Cyber Forensic Process

4.1 Motivations

In this chapter, the primary objective is to propose a unified digital forensic process built on top of the literature review and comparative analysis presented in Chapters 2 and 3. The intent is to present a new process designed to:

- Build on top of the commonalities proper to the existing state of the art processes.
- Incorporate the most adequate features characterizing the reviewed processes.
- Address deficiencies that have been identified in the proposed processes.
- Propose new ideas highly suitable for digital forensic process.

The research reported in this chapter represents an attempt to contribute to the state of the art cyber-forensic processes rather than to propose a comprehensive process designed to subsume or even outperform the existing processes.

In this respect, the characteristics of the resulting cyber- forensic process can be summarized as as follows:

Principles versus Phases: A great idea inspired from the Beebe and Clark Prcess [3] lies un the need to structure a process into principles and phases. Phases are steps that an investigator is required to take in some pre-specified order during the investigation. On the other hand, principles are a set of specific objectives that an investigator is required to progressively achiev throughout the whole investigation.

Description Language: A key attribute, proper to End-to-End digital investigation process (EEDI) [43], consists in using of a language for the description of planned and executed forensic activities. To this end, the unified cuber-forensic process adopts an XML-based format to describe, present, and further manipulate digital forensic activities. The choice of XML is, from the process standpoint, motivated by convenience.

Sophisticated Control and Iterativeness: The proposed process accounts for the iterative nature of forensic investigations. Actually, it is endowed with a control structure that permits feedback from phases together with the iteration of some of the steps when needed.

Computer Support: An important feature of the proposed process is communally known as computer support. It means that the process has been embedded into a forensic analysis software environment. It is directed towards guiding the investigator through the process phases when conducting a cyber crime investigation.

4.2 Principles

As proposed in several reviewed processes, a principle represents a fundamental attribute or requirement of a digital forensic process that cannot be uniquely associated with a single phase. While it has to valid throughout the entire investigation, a principle crosses more than one phase of the process. The principles of the proposed

Unified Cyber-forensic Process are thus laid down and thoroughly explained in what follows.

4.2.1 Evidence Preservation

Evidence preservation stipulates that all activity related to digital investigation must be performed in a manner that guarantees safeguarding, accuracy, integrity, and reliability of the evidence.Indeed, digital evidence is fragile, vulnerable, often invisible, and easily altered or destroyed. Thus, in every phase, the investigator is required to collect, handle, transfer, and examine the evidence with precautions to prevent alteration. For instance, during the acquisition phase, the evidence preservation principle is essential. It stipulates that appropriate measures for storing and transporting digital evidence are required to maintain its integrity in accordance with evidence preservation standards. During the examination phase, this principle implies the obligation for the the investigation team to work only with copies of the evidence so as to keep the originals intact. It is very important to ensure that any changes to the data are thoroughly documented and actually performed without any modification. During the closure phase, a proper disposal of the evidence must be carried out. While digital evidence legally authorized for destruction must be destroyed, authorized evidence is returned to the owner and the remainder data is preserved and properly archived.

4.2.2 Chain of Custody

The chain of custody principle refers to reliability and authentication of evidence. All actions directed to monitor each access to the evidentiary data from the time it is collected until the time it is used in a court of law must be clearly identified and well-documented . Accordingly, during all the investigation phases, this principle imposes

a limited and strictly controlled access to the evidence from the time it is seized, transformed, and transported, until the closure of the case. The latter phase leads to archiving, disposal or return of the evidence in question. Any improper handling of the chain of custody may result into allegations of tampering or misconduct, which would compromise the case in question and finally lead to the acquittal of cyber criminals.

4.2.3 Documentation

The documentation principle stipulates that all activities executed within a digital forensic investigation process activities must be fully documented. For instance, during the verification and initial response phase, detailed notes must be maintained together with pictures, videos, and sketches. It is of paramount importance to document all information from the initial advisory that an incident may have occurred to the formulation of the investigation plan. Further more, any event pertaining to "who, what, where, when, why, and how" type questions, witness statements, and damage information, including direct costs and personnel time, must be documented [3]. During the acquisition phase, it is necessary to document in details all actions undertaken and all techniques and methodologies used by the investigative team . During the examination phase, the investigative team must use adequate tools to document all actions performed together with results derived from these actions. In deed, the results obtained must be accurately documented as they will form the basis for the investigative report that will be submitted later to the prosecuting authority. During the analysis phase, documentation is very prevalent. During the reporting and presentation phases, documentation pertaining to the previous phases is put together and compiled into an easy to understand document, which will be used to present the sum of the investigation to the appropriate authorities. During the closure phase,

it is important to maintain detailed documentation regarding the review and results of the investigation, , and all preventive and corrective measures. It is essential to document all details regarding the proper disposal of evidence. It is during this phase that all documentation collected and created during the whole process is properly archived to mark the end of the investigation.

4.2.4 Complete Authorization

The complete authorization principle aims to ensuring that the investigation is conducted under proper authority in respect of full legality. For instance, during the verification and incident response phase, once the nature of the incident and the type of the required investigation are properly determined, proper investigative authorities (e.g. warrants) are obtained in order to allow the investigative team to tackle the next phases. During the acquisition phase, actual seizing of evidence occurs and intrusive actions are consequently performed. In this context, it is essential that the investigation team confirms the obtaining of appropriate warrants from the previous phase prior to undertaking any action. During the closure phase, this principle is less prevalent as all authorities to conduct the investigation have been already obtained.

4.2.5 Information Disclosure

The information disclosure principle is integrated to ensure that the information collected during the investigation is only provided to legal and authorized principals. For instance, during the verification and incident phase, this principle is critical as any violation could tip off a perpetrator who may consequently, destroy evidence and compromise the investigation. It is essential that an evaluation be made as to who will be made aware of the situation and impending investigation. During the examination phase through to the presentation of findings, it is important that the

sensitivity of any information discovered or revealed during the course of the investigation be managed appropriately. It is conceivable that, initially, a piece of evidence may be assigned a sensitivity level, other elements may come to light and modify the sensitivity level. In certain cases, it may be necessary to restrict access to specific information. During the closure phase, this principle requires that appropriate levels of security be assigned to all documentation. In addition, it is also required that all elements be treated in accordance with policy regarding sensitive information. During the archiving process, elements are subject to different archiving standards based on their security classifications.

4.2.6 Investigative Priority

The investigative priority principle deals with priority aspects of an investigation. For instance, during the planning phase, the investigation plan is elaborated based on directives and priorities received from higher authorities. As a result, the investigation plan makes the investigative priorities clear for the next phases. During the acquisition phase, whenever the resources are not enough to conduct all of the evidence collection, the investigative team must develop an action plan in accordance with the established priorities. During the examination phase, the team must also examine evidence in accordance with priorities. Since this process is inherently iterative and possesses feedback, prioritized actions serve as feedback and could, as a result, be used when it is required to return and collect additional evidence.

4.2.7 Quality Assurance

The quality assurance principle is integrated to ensure the high quality of all investigative activities. It stipulates that digital evidence requires "precautions to prevent

alteration or destruction". Thus, it is mandatory to provide special tools and equipment, appropriate procedures, and specialized training in order to meet the combined requirements of control, accuracy, and reliability of digital evidence. All phases, from preparation to closure, need quality assurance mechanisms to monitor executed actions, control, and improve their outputs.

4.2.8 Accreditation

The Accreditation principle refers to the scientific and proven methods and techniques used to gather, process and interpret digital evidence during criminal investigations and prosecutions. It stipulates that harmonized methods and practices and standardized equipment must be used to perform a digital investigation. For Instance, during the acquisition and analysis phases all the methods and equipment used to collect and examine the digital evidence must be accredited by a court of law and/or admitted by the law enforcement agency.

4.2.9 Iterativeness and Control

The iterativeness and control principle stipulates that in all the process phases, and throughout the investigation, feedback must be taken from previous phases according to predicted controls. In addition, an iteration over some phase or sub-phase must when needed proceed according to the iterativeness permitted in the process.

4.3 Phases

In this section, we will detail each of the processes by presenting a description, underlying objective, inputs and outputs and a set of activities that should be carried within this phase.

4.3.1 Preparation

Description: This phase encompasses all the preparatory activities that will facilitate the successful execution of the investigation.

Objectives: The primary intent of this step is to prepare the needed initial authorization, infrastructure and equipments and also to establish the legal and policy settings.

Input: In terms of input, we cite the mandate to prepare for potential cyber-crime investigations in a particular geographic area and under some jurisdiction. We are also given the type or list of potential organizations where these investigations are going to take place together with underlying policies.

Output: In terms of output, all needed initial authorizations are granted, together with a clear understanding of the underlying laws and policies that are proper to the investigation.

Activities: The planned activities are:

1. *Acquire and verify the primary authorization:* During this step, the investigator needs to acquire the first authorization needed to start working on the case and to access the crime scene. The authorization should be written and signed by the person/s of concern (the investigator's supervisor, the owner of the company, etc.).

2. *Identify relevant laws:* Depending on the location of the crime scene, the investigator may need to review the legislation and laws of that specific area, to make sure that he/she does not violate any jurisdiction.

3. *Identify relevant company policies:* If the investigator is investigating a crime in a company, before starting any action in the case, he/she needs to be familiar with all of the company policies.

4. *Acquire the necessary tools and techniques:* Depending on the crime, the investigator will need different tools and techniques to perform his/her investigation. He/she should prepare all the tools and techniques that he/she generally needs and look for some the specific ones that he/she may need for that type of investigation. The required list of tools and techniques will be updated at the planning phase, when the investigator has a better idea of the type of crime he/she is dealing with and the requirements for that specific crime.

4.3.2 Verification and Initial Response

Description: This phase encompasses incident verification and the elaboration/implementation of an initial response to consider or eradicate a cause of the incident. It serves as a precursor to the entire investigation. Moreover, it allows the investigator to rapidly develop an appraisal of the situation in order to determine whether an incident is worth for investigation. If so, the crime scene is needed to be secured.

Objectives: The primary intent of this phase is to verify whether an incident is worth for investigation has really occurred and to secure the corresponding crime scene. Besides, the objective is also to formulate and implement a response strategy to contain or eradicate the cause of the incident.

Input: This phase requires to input a reason to initiate a verification of an information system or one of its elements. The input will be generally an allegation of an suspected, abnormal or illegal event. Such events may include a company's security policy violation, a detected security breach, or a suspected criminal activity.

Output: As stated earlier, this phase will determine if the allegations are founded. If so, the outputs to this phase are the acquisition of proper legal/administrative authorities to conduct an investigation, the production of the investigative plan and a secure crime scene for the collection of evidence phase.

Activities: The planned activities are:

1. *Identify witnesses:* When the investigator arrives at the scene, he/she needs to identify every person present and to ask them to stay for questioning about the incident and whatever they may know.

2. *Document the scene:* An organized step by step scene documentation is one of the important stages in a proper processing of an investigation. The final result of a properly documented crime-scene offers a finished product to others that allow to either reconstructing the scene, the chain of events in an incident and/or a court room presentation. In documenting the scene, there are actually three functions or methods, used to properly document the crime scene. Those methods consist of written notes, which will ultimately be used in constructing a final report, crime scene photographs, and diagrams or sketches. Consistency between each of these functions is paramount.

3. *Physically secure the scene:* It is important that the first investigator on the crime scene properly protects the evidence. The entire investigation hinges on that first person being able to properly identify, isolate, and secure the scene. The scene should be secured by establishing a restricted perimeter. This is done by using some type of rope or barrier. The purpose of securing the scene is to restrict access and prevent evidence destruction.

4. *Digitally secure the scene:* The investigator needs to secure the digital crime scene and prevent data from changing. To do this, he needs to suspend suspect processes, and acquire volatile and non-volatile data, and make sure that the integrity of the evidence is protected.

5. *Confirm that the incident occurred:* The investigator verifies the systems and looks for direct evidence that corroborates the existence of an incident. The

investigator will then confirm the occurrence of the incident and the need to actually investigate it.

6. *Interview witnesses:* Each person, present at the scene, should undergo a questioning session with the investigator. In this step, the investigator tries to learn the most about the incident circumstances and nature. Afterwards, he/she will use this knowledge to conduct interviews and gather, as much as possible, useful and important information about the incident.

7. *Identify existing security procedures:* In this step, the investigator verifies if any security procedures were implemented by the company where the incident occurred. He/she should also acquire a documentation of the actions taken during the occurrence of the incident and adjust his/her first response according to those actions.

8. *Identify potential suspects and detain them if needed:* If the investigator has any reasonable doubt that a specific person might be involved in the cyber-crime incident, or he/she catches a suspect red handed, he/she has to detain him if he/she has the authority to do so. In addition, he/she has to call the competent authorities to come and pick up the suspect for further questioning.

9. *Collect volatile evidence:* In the same process of securing the digital crime scene, the investigator needs to collect the volatile data that could be lost and make sure that the collected items remain unchanged.

10. *Establish an initial list of systems to be investigated:* Once the investigator becomes familiar with the crime scene, he/she should make an initial list of the systems that he/she needs to investigate. This list is valuable in the planning of the investigation. Note that, as the investigation unfolds, the investigator may find out that new systems are required to be investigated.

4.3.3 Planning

Description: This phase leverages the output from the previous phases in order to elaborate a viable plan together with an investigation strategy that may lead to a successful investigation. The aforementioned plan will guide the investigator in the subsequent phases i.e. acquisition, analysis, reporting, presentation and closure.

Objectives: The main objectives is to categorize the cyber-crime incident according to an established prior taxonomy. This would help in planning for the needed authorizations, resources, infrastructure and equipment. In addition, the intent is to elaborate a course of actions to be executed in order to carry out the investigation. If the ultimate goal is prosecution, the elaborated course of actions should follow the needed standards to be admissible in a court of law.

Input: In terms of input, we have the output of the previous phases i.e. preparation, verification and initial response.

Output: In terms of output, this phase will produce information on the type of crime, systems to investigate, required authorization, required infrastructure and resources and a course of actions.

Activities: The planned activities are:

1. *Categorization of crime:* Before starting the collection of evidence, the investigator needs to classify each crime in a different category according to an a priori defined taxonomy. Each incident is categorized according to its severity and its importance.

2. *Identification of systems to investigate:* Once the verification and initial response phase is finished, the investigator should finalize his/her list of systems to investigate. Every damaged, infected or corrupted infrastructure and system should be placed on the list for further examination.

3. *Identification of the required authorization:* The authorization is required in every phase. In this particular phase, the investigator is going to plan for all the authorizations that he/she may need for his/her investigation, according to the crime category and to the systems he/she needs to investigate.

4. *Identification of required resources:* Each investigation has a different level of importance. Some investigations need larger infrastructure and resources than others depending on the importance and the severity of the incident. In this step, the investigator will evaluate and estimate the human and technical resources needed for the ongoing investigation, as well as, the cost and the amount of time that should be spent on the case.

5. *Preparation of the case plan:* Every little detail on how the acquisition, examination, analysis, reporting, presentation and closure are going to be executed is planned at this level. Once, it is detailed enough and the investigator has a clear idea of the actions to be taken, he/she can carry on with the rest of the investigation.

4.3.4 Acquisition

Description: In this phase, the investigator proceeds with the acquisition of the evidence in a form that can be preserved, transported, stored, duplicated and later analyzed. During the acquisition process, the investigator should follow strict procedures and respect approved rules and regulations.

Objectives: The main objectives of this phase are evidence collection, packaging, transportation, storage, preservation, compression, sampling, encryption and categorization in addition to system recovery activities, as needed.

Input: In terms of input for this phase, we cite proper legal authorizations to conduct

the acquisition activities in addition to the investigation plan and a secure crime scene.

Output: As of the output, we cite a collection of data that preserves the integrity of the evidence and that is extremely useful and pertinent for analysis purposes.

Activities: The planned activities are:

1. *Collection:* Computer evidence must be handled very carefully in a way that preserves its physical integrity and that of the data, it contains. Special considerations should be taken to the data that may get damaged or altered from electromagnetic fields. The investigator needs to attend there before such an incedent takes place.

2. *Packaging:* Each electronic device should be handled with care and the investigators should follow a strict protocol to package the evidence. Many factors like humidity,temperature, static electricity, etc, can modify the original state of the evidence.

3. *Transportation:* Evidence must be transported to a suitable location for later examination. This could be simply the physical transfer of seized computers to a safe location; however, it could also be the transmission of data through networks. The means of transport used has to preserve the integrity of the evidence.

4. *Storage:* The collected evidence, in most cases, are needed to be stored because examination cannot take place immediately. Each piece of evidence should be stored following a protocol that is specific to it, and in a way that will preserve the integrity of the evidence.

5. *Preservation of evidence (duplication):* Every investigator should know that the original evidence has to be preserved always and any analysis should be made

on a copy of that evidence. In this step, the investigator will duplicate the evidence needed for analysis and store the originals in a safe environment for proof of integrity.

6. *Creation of integrity proof:* Each evidence found should be well documented and preserved to be able to proof its original state in case of an inquiry by the court. The documentation will provide a good proof of integrity, which will make the evidence admissible in court.

7. *Compression:* This step explains eventual compression techniques used by encryption, backup or digital signature software used to collect and/or preserve evidence. The software program, that uses compression, has to be absolutely lossless, and proved to be not altering the evidence on which it is used.

8. *Sampling:* The sampling techniques used in the collection phase must be proven to preserve the integrity of the evidence. If in fact the sampling technique had an impact on the collected evidence, it has to be demonstrated clearly and unambiguously. Every sampling method should be valid and all the conclusions that can be drawn from the sample have to be defined clearly.

9. *Encryption:* Sensitive evidence should be protected through a cryptographic systems/application .

10. *Evidence Categorization:* The collected evidence should be separated in different groups to perform a better more specific analysis to each group. Some categories may include: volatile evidence, memory, etc.

11. *System recovery techniques:* In this step, the investigator will recover a locked system to a safe state. Once this system is safe, the investigator will be able to examine it and extract useful information from it.

4.3.5 Examination

Description: This phase is critical for the continuation of the investigation. It applies some processing to the acquired evidence for the purpose of extracting valuable information that will be subjected to extensive analysis.

Objectives: The intention is to extract relevant digital evidence from raw data that was collected in the previous phase. The extracted information should be documented in details in order to allow an appropriate exploitation.

Input: As input, we do have the data collected on the crime scene in the acquisition phase.

Output: This phases produces as output valuable and relevant digital evidence that is fully documented.

Activities: The planned activities are:

1. *Data reduction:* With the rising of computer capacities and the expansion of network communication, the investigator faces a huge volume of information to examine when he/she is conducting a digital investigation. It is very important to maintain the focus on useful data. While the content of data is known, some irrelevant data can be eliminated. For instance, the protected system files included in a working copy of an image of a computer media might have no relevance to the investigation and therefore should be filtered out. Another way to minimize the volume of data is to eliminate the information related to events that occurred in a certain period not relevant to the incident based on the file timestamps and/or MAC times.

2. *Evidence recovery:* The forensic examiner can use different approved methods, techniques, software and hardware to identify and recover every evidence and information contained in the system or the media investigated. The recovery

may include the deleted data, unallocated space, partition table, file system, and file slack, emails/attachments, log files, etc. The recovered evidence should be documented.

3. *Categorization of findings:* Based on the type of the evidence and its source, the forensic examiner may categorize the piece of evidence and answer two important questions: what is this evidence? and where did it come from? The categorization of findings may appear very significant to investigative analysis and reconstruction.

4.3.6 Analysis

Description: The collected data, in the acquisition phase, and identified /recovered information, in the examination phase, are rigourously analyzed and interpreted in ordered to bring together all pieces of the cyber crime puzzle, understand the events that occured, reconstruct the timeline, establish the facts and identify the suspects.
Objectives: The intention is to prove the allegations and reconstruct the complete facts by addressing the following inevitable questions: What happened (what)?, where did it happen (where)?, who did it (who)?, How it has been done(how)?, what is the motivation (why)?, and when did it happen (when)?
Input: As input, this phase takes the collected and examined digital evidence from the previous phases.
Output: As output, this phase may produce facts, timeline and proof statements.
Activities: The planned activities are:

1. *Timeline generation:* The examined pieces of evidence are used to create a sequence of events and set up a time of actions leading to the crime facts. This temporal aspect of evidence may help the investigator to associate cyber criminal actions happened to the principals involved.

2. *Chain of evidence construction:* Individual pieces of digital evidence may not be useful by their own but when they are integrated in a coherent chain of evidence, where each link is supported by one or more pieces of evidence and leads to the next link, they become very useful for the investigator to create a relational reconstruction and identify important connections between people and the investigated systems.

3. *Hypothesis:* It is very hard to get the entire picture of what occurred in a cyber crime incident. The investigator must establish a hypothesis of what happened based on the evidence that was left on the crime scene after subjecting it to examination, correlation and extensive analysis. This hypothesis should be strong enough to stand up against court scrutiny and challenges.

4. *Extracting hidden data:* Hidden data may reveal precious information about cyber criminals. Therefore, such data should be extracted. The functionality of hidden data may have significant implication on what happened. The analysis of protected, encrypted and compressed files as well as steganography can reveal the attempt to conceal relevant data.

5. *Low-level analysis:* The investigator should perform low-level (system level) tasks to recover valuable information that has been logically deleted from the system such as memory analysis to determine if it contains data fragments that may be useful to the ongoing investigation.

4.3.7 Reporting

Description: This phase describes, in details, the whole actions and activities (including the techniques and procedures), performed by the investigative team within a cyber forensic process. The first incident responders, the forensic examiners and the

investigators should create an accurate report at the end of their tasks investigation. All the reports together with the conclusions should be consolidates into a single document that will be handed to the investigation lead and/or to the prosecutor.

Objectives: The reporting step aims at supplying the leading investigation and the prosecutor with a clear, concise and understandable document about the underlying activities taken to address the crime-incident.

Input: As input, in this phase, we have the the notes, sketches, pictures and reports from the execution of the previous process phases.

Output: The output of the reporting phase is a comprehensive written report describing in details the steps, actions, and conclusions that the digital investigators have made.

Activities: The planned activities are:

1. *Activity Documenting:* Any task, activity or results should be documented. The description should be concise, clear and following an established reporting standard as much as possible.

2. *Findings description:* The conclusions of the examination and analysis activities should be communicated in more unambiguous way by making a comprehensive and clear description of the supporting evidence. The findings are described in more detail and then summarized as per the requirements of the intended people.

3. *Report Consolidation:* All the reports should be consolidated in a single and unique report that will be handed to the investigation lead or the prosecution team.

4.3.8 Presentation

Description: The presentation of evidence determines the outcome of a digital forensics investigation. In this phase, the forensic examiner investigator and the prosecutor will present the findings and supporting evidence and testify during the trial in court (assuming that the result of the investigation is complete). In this respect, a well-documented case is much more likely to be admissible. Hence, it is essential that the investigator keeps a complete, accurate, and comprehensive documentation of all activities and actions.

Objectives: The purpose of this phase is to present the findings of the investigation together with and admissible evidence to a top management or in a court of law.

Input: The presentation considers the result of the reporting phase as an input.

Output: The output of this phase is a presentation of the findings together with the supporting evidence to the target audience.

Activities: The planned activities are:

1. *Preparation:* This activity is about the preparation of the presentation of the investigation conclusions. Time should be allocated to presentation, scripting, rehearsal with the appropriate participants (e.g. attorney), identification of areas that need more clarifications, anticipation of audience questions, etc. Conclusions should come first in the presentation and they should be supported afterwards with a strong evidence.

2. *Displaying digital evidence:* The digital evidence should be displayed in the testimony in order to back up the claims made and also to provide the needed context for some facts. For a successful case presentation and testimony, time should be allocated to figure out what to show in terms of digital evidence and when to show it.

3. *Testimony:* In this activity, the investigator will present, in a perspicuous way, his/her case to a senior management or a court of law. He/She will proceed according to a prepared script and display the needed digital evidence to back up his/her claims.

4.3.9 Closure

Description: The final phase of a cyber forensic investigation process is obviously the closing of the case under investigation. Indeed, great consideration should be given to this step in order to capitalize on the knowledge and experience gained and to properly destroy, archive and return the evidence.

Objectives: The intent here is to proceed with the case closure by doing an appropriate disposal of the evidence and also to share knowledge and experience to enhance the execution of future investigations.

Input: As input in this phase, we have the output of all the previous phases in the form of reports, evidence, etc.

Output: The output of this phase is the evidence disposal together with a set of recommendations, and learned lessons to incorporate in the subsequent digital investigative activities.

Activities: The planned activities are:

1. *Evidence disposal*: The investigation team together with other parties such as the prosecution team should decide on which part of the evidence should be destroyed, archived and/or returned to legal owners.

2. *Lessons learned:* The lessons, taught during the investigation, should be used later in other cases in order to avoid errors, improve the execution of cyber forensic actions and enhance the performance.

3. *Submit recommendations:* Appropriate recommendations and best practices about critical information pertaining to encountered problems should be submitted to the concerned parties.

4. *Cross checking/referencing:* A solved case can be used to deal with unsolved similar cases.

5. *Profiling:* Undertake all actions to create profiles for investigators and digital crime cases.

4.4 Control and Iterativeness

The proposed process is endowed with a pre-defined control that will allow iteration over the process phases and also for feedback from phases. This feature may be revealed important and essential when conducting real-life investigations. Actually, as the investigation unfolds, there may be a need to do, for instance, more evidence acquisition, examination and/or analysis. In this type of situations, control becomes an important aspect in a digital forensic process. Figure 13 depicts the pre-defined control over the phases of the proposed process.

4.5 Description Language

A very important feature of the proposed process is the computer support together with the design of a dedicated description language that is meant for the specification of cyber forensic activities. The importance of this language stems from the following:

- To provide computer support for the investigators on the field.

- To provide the software capability to plan for investigations.

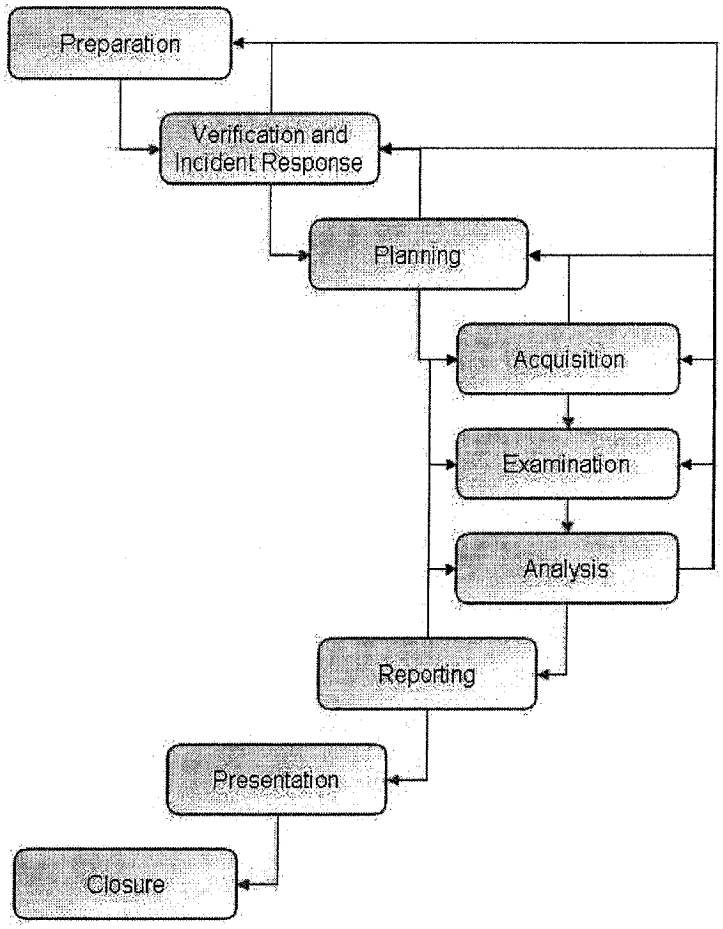

Figure 13: Control Aspects in the Proposed Process

- To provide the capability of quality assurance and reporting over planned and executed investigations.

The language comes in the form of XML. The choice of this format is motivated by the extensive availability of APIs that facilitate the visualization, parsing, modification, transmission, and manipulation of XML documents. Figure 14 presents a sample of the document format type.

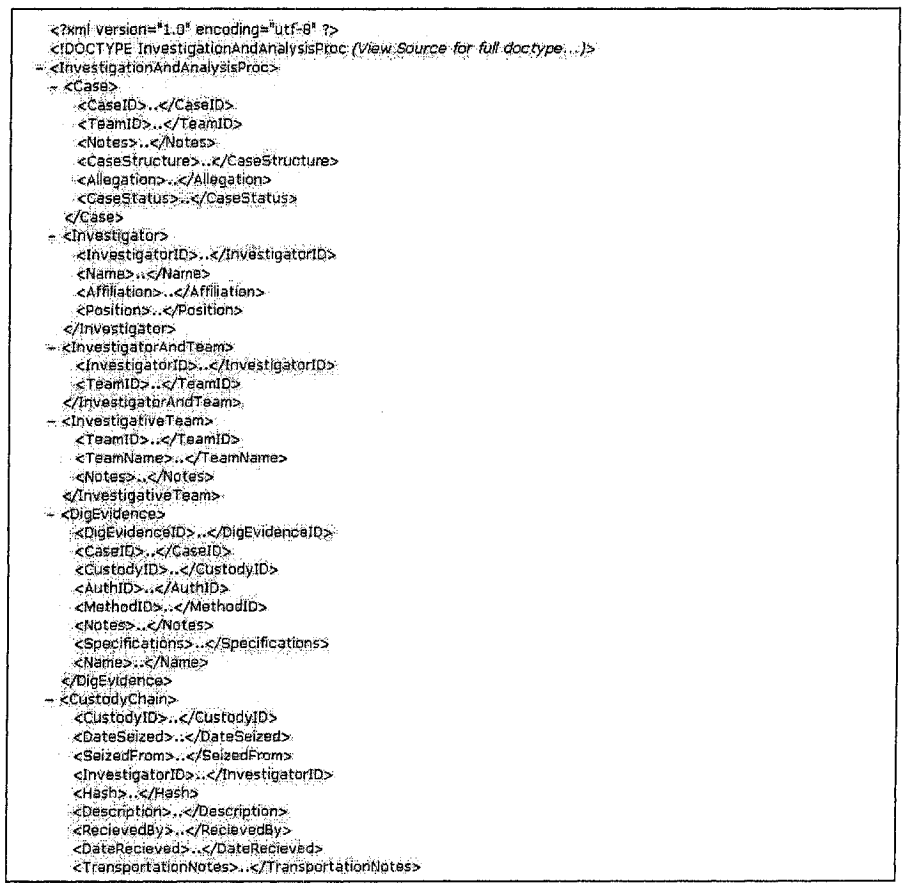

Figure 14: XML Forensic Description Language

Chapter 5

E-mail Analysis

5.1 Motivation

E-mail may be considered as one of the main vectors of Internet communication, and its traffic has been increasing tremendously since the advent of the World Wide Web. Nowadays, it is estimated that more then two trillion E-mail messages are sent per year either for business or personnel use.

Many companies and institutions rely on E-mail for business transaction. Since a long time, individuals have adopted E-mail as an alternative to regular mail to exchange and share information and news . With the increase in E-mail traffic, the exploitation of E-mail for illegitimate practices is also becoming a common practice. Some of the known examples are: sending spam, threats, hoaxes, bullying, harassment, racial vilification, viruses and worms. As an example, on the 22 of February 2006, twelve Nigerians were arrested in Amsterdam and the city of Zaandam in E-mail scam investigation. The arrests involved in this Internet E-mail scam are known as 419 frauds, where the recipients are coerced into investing in an artificial scheme. In this instance, most of those victims are found as US citizens, with a total of two

million dollars in fraudulent earnings. In more serious cases, E-mail is used to carrying out criminal activities such as drugs and human trafficking, plots, assassinations, child pornography, etc.

One of the major difficulties facing E-mail investigators is the large amount of E-mails to examine and the reliability of the information contained in these E-mails. E-mail, by its nature, is very easy to send. The "From" address header field can be easily forged as in the case of spam E-mails. The metadata containing the E-mail header and the path along which the message has traversed can also be forged or kept anonymous. An E-mail can also be routed through anonymous E-mail servers to hide any information about its origin. Sometimes the only useful information to help identify the author of an E-mail is the E-mail structure and the message it carries.

5.2 Objectives

As a result of the growing E-mail misuse, investigators need efficient automated methods and tools for analyzing E-mails. In this work, we propose to develop an E-mail analysis framework to assist an investigator and to gather clues and evidence in such an investigation. The framework needs to provide all sort of functionalities to handle an E-mail corpus in order to extract the required knowledge.

Simple functionalities, such as E-mail storing, editing, searching and queering are certainly needed, but not sufficient. We need to provide more advanced functionalities for a multi-stage analysis to help the investigator gain a profound insight in the E-mail corpus under investigation and achieve more ambitious goals. So our objective lays in the following four directions: E-mail statistic analysis, E-mail data mining, E-mail social networks and E-mail geographic localization.

Statistical analysis can without any doubt reveal some non straightforward information about E-mails. For instance, the distribution of E-mails per sender or receiver

in a specific period of time allows to identify periods of time with intense E-mail exchange and by the same suspicious activity partners. This in turn may help to restrict the extent of E-mails for thorough investigation. A more advanced use of statistics can also help compute profiles for E-mail authors to be used for authorship identification [46, 20]. E-mail mining, on the other hand, has received great interest in the context of computer forensics. Some applications showed big success in identifying or categorizing authors of anonymous E-mails [41]. Other applications demonstrated the power of machine learning tools for spam filtering and intrusion detection. E-mail social networks allow to model user E-mail flows and behaviors to detect misuses that manifest as abnormal E-mail behaviors [27]. Finally, the E-mail geographic localization capability, leverages a means by which an investigator can identify and view geographic locations associated with E-mail accounts and individuals.

5.3 E-mail Statistic Analysis

To compute statistics on an E-mail corpus, E-mails are first loaded from their raw files and preprocessed. The preprocessing of an E-mail consists in extracting its relevant information such as the sender and receiver, the subject, the message body, etc. We use the Java Mail API (JMA) for this reason. The extracted information of each E-mail is stored in a database according to the schema shown in figure 15.

The E-mail table collects information about the sender, receiver, and the date on which the E-mail was sent. In addition, the table archives E-mail domains and other information related to the clustering and classification of E-mails using data mining tools. When a single E-mail has many receivers, the table inserts a row for each receiver. This is because, conceptually, there is no difference between sending an E-mail to many people at once and sending it to each one individually.

Each E-mail is associated with a folder name attribute for a virtual classification

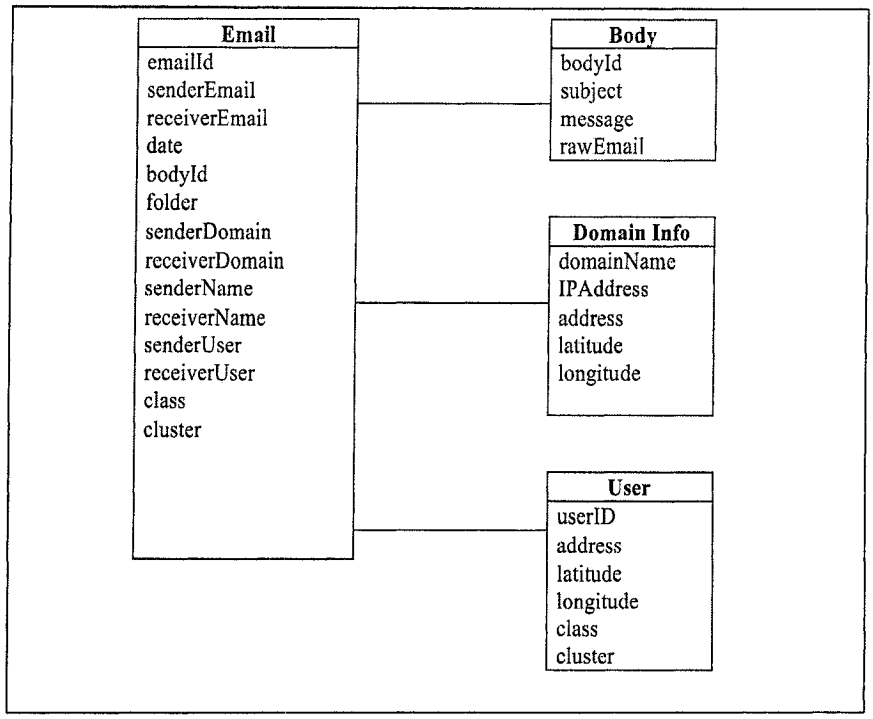

Figure 15: E-mail database schema

of E-mails in folders. The body table stores information of the E-mail body composed of the subject, the E-mail message and the E-mail itself in its raw format. The domain table stores information about domain addresses and their geographic coordinates. We use some Internet services to identify the physical address of an E-mail server and its geographic location in terms of latitude and longitude. This information is later used to provide a geographical localization of E-mail information on a map. The user table stores information about E-mail users and their physical addresses.

In the current state of our framework, we compute mainly the statistical distributions of E-mails. However the framework offers the possibility to dynamically add further statistical operations by just specifying appropriate SQL statements.

The following distributions on E-mails are computed:

- E-mails per sender,

- E-mails receiver,
- E-mails per sender domain,
- E-mails per receiver domain.

Note that, we compute all the statistics on E-mails, restricted to a specific period of time and a specific list of users. Each statistic is associated with a chart in two or three dimensions (See Figure 16).

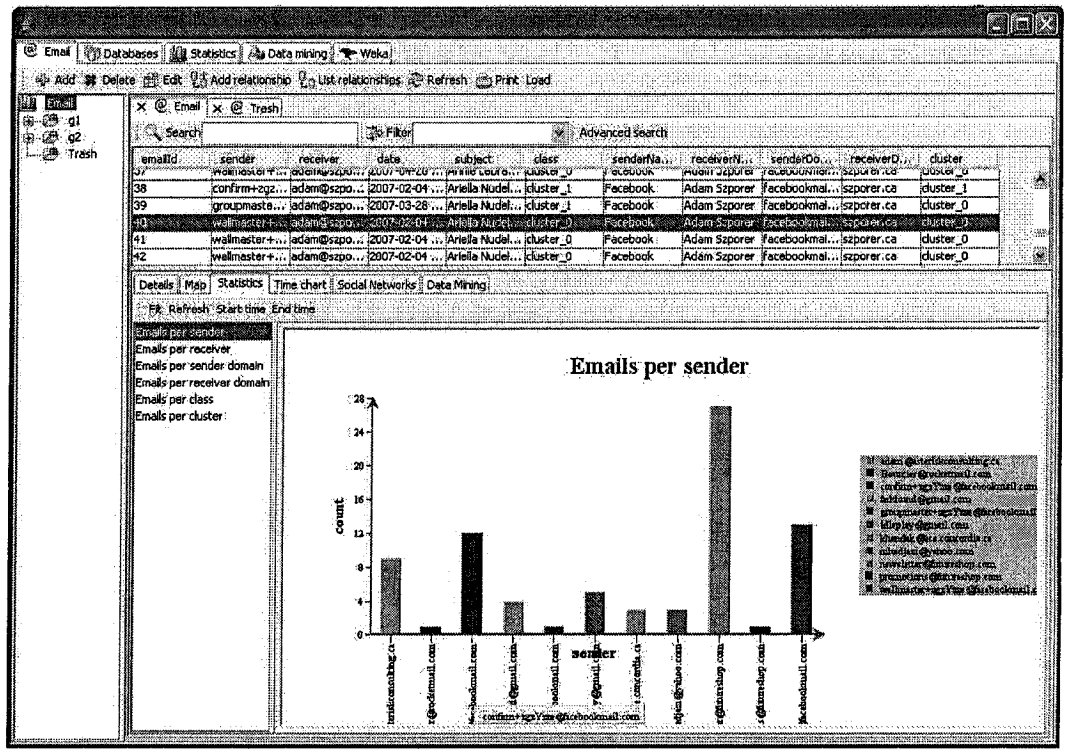

Figure 16: Statistics Viewer

5.4 E-mail Mining

5.4.1 E-mail Classification

The classification in data mining is achieved under supervised learning. The process starts by building a model to map instances from a data set to well known classes. Each instance is a vector of numerical attributes. Attributes can be nominal, but they are ultimately converted to numeric representations before processing.

In general, to build a classification model, we feed a learning algorithm with a set of already classified instances, called the *training set.* Learning algorithms are very numerous in the literature. Some common examples are: decision trees [22], Bayesian probabilistic approaches [30], support vector machines [23] and neural networks [39]. To test the performance of the developed model, we use a different set of classified instances, called the *testing set.* If the classification error of the model is judged acceptable, the model can be used to predict the class of non-classified instances. The accuracy of the built classifier depends largely on the size of the training set and its representation capability with respect to the instances to classify. Generally, the lager the size of the training set and the higher its representation capability, the better is its predicting power.

E-mail classification deals in general with the classification of E-mail message bodies and their subjects [48]. The message and the subject of an E-mail are converted to vector of *metrics*, called also *features.* A typical conversion, starts by tokenizing the message and the subject of an E-mail to a list of words, then a vector of words frequencies or counts is produced. Depending on the purpose of the classification, the words are stemmed to root, and a list of stop words (words to ignore) is used. Sometimes thresholds are used to fix a limit above which, word frequencies and counts are considered. In other situations, a totally different set of metrics is used. In our

framework, we are concerned with two types of classifications. In the first one, we classify E-mails per topics while in the second one, we classify E-mails per authors for authorship analysis.

Classification of E-mails per Topics

The possibility to automatically classify E-mails per topics is an important issue in an E-mail investigation. For instance, if the investigation is about drug trafficking and we have a large number of E-mails to analyze, the possibility to isolate E-mails related to this matter would have a big impact on both effort and time saving. One could think about performing a simple search with the word "drug" or other related keywords to achieve this goal. However, in some situations this is not satisfactory. Very often, people use different names and some speech artifacts to hide information. So the proper way to isolate interesting E-mails is to use a classifier which has already been trained to identify drug related E-mails.

In our framework, topic classification is achieved using a near classical text mining procedure. We start the process by selecting a training set, consisting of and already classified as the collection of E-mails. Part of this set is used to form the testing set [38]. E-mails are first passed through a preprocessing phase to extract their constituents: the sender, the receiver, the subject and the message body. Each E-mail is transformed into a vector of attributes, composed of the sender, the receiver, and some features extracted from both the subject and the message body. The features we extract are mainly word frequencies. We tokenize the message and the subject of an E-mail in a list of stemmed words. Both punctuation and space characters are ignored. We also use a stop list to ignore some structural words. For the classification, we use a support vector machine learning algorithm [23].

5.4.2 E-mail Classification for Authorship Analysis

Authorship analysis is one of the most important issues in cyber-crime forensics. It consists in the examination of the characteristics of a written text in an attempt to draw some conclusions about its author. In the case of authorship identification, we try to determine the likelihood that a specific individual is the author of a piece of text by examining some of his writings. For this purpose, we need to be able to construct a profile for a suspected author by capturing some features of his/her writing style in the form of metrics. In general, the main issues in authorship studies are:

- Which types of writing-style features are effective for the identification of authors?
- Which classification techniques are effective for identifying the authorship of E-mails?

Authorship identification in data mining is considered as a classification problem. It requires the identification of sets of relevant features of a large number of writings. Up to now, there is no consensus on a best set of features for a wide range of application domains.

In early studies, researchers analyzed word usage to identify authors. But since this feature depends mostly on the content (topic) of the written text, the effectiveness of this approach is limited. A generic authorship identification needs to be based on content free features. In [15] and [16], the author proposed to use sentence length and vocabulary richness. Burrows proposed a set of more than 50 high-frequency words in [18]. Mosteller and Wallace in [31] demonstrated the discriminating capability of function words ("if", "while", "then", etc) which was confirmed more recently in other works [37, 18, 9, 12]. Functions words are classified as syntactic features. They are content free but help to syntactically form sentences. Punctuation is another

syntactic feature used in authorship identification. Holmes, in [8], analyzed the use of some lexical features such as the use of two or three letter words and words beginning with a vowel. However, syntactic features are proven to be more interesting.

Recently, Corney, Vel, Anderson, and Mohay in [25] investigated some other features for authors with different education backgrounds. Koppel, Argamon, and Shimoni in [26] and Argamon, Saric, and Stein in [42] showed that there exists a difference in writing styles between males and females, especially in the use of pronouns and certain types of noun modifiers. In [33] and [34], the authors investigated new characteristics to be used in the case of E-mail authorship analysis. Other studies focused on comparing different types of writing style features and classification techniques.

Even if key-word-based features are not effective for author identification, in some situations, the proper use of these features may improve the performance of authorship identification in topic related to investigations, as they can express personal interests [6, 41].

Techniques for Authorship Identification

Techniques for authorship identification can be classified in two categories: statistic techniques and machine learning techniques. In early studies, most analytical tools used in authorship analysis were statistical methods. Statistical techniques precede machine learning techniques as proposed in a pioneering study of Mendenhall in [46] in 1887. The technique constructs histograms of word-length distribution for various authors and compare them to the distribution histogram of the text to classify. Farringdon in [20] used a cumulative sum statistics procedure. The technique consists in creating the cumulative sum of measured variable deviations and present it as a graph to compare between authors. Burrows in [5] employed principle component analysis on the frequency of function words.

In recent times, machine learning techniques are more investigated for authorship analysis. Machine learning methods achieve higher accuracy than statistical methods and can deal with a larger set of features. They are also more tolerant to noise and nonlinear interactions among features [10]. In 1996, Tweedie et al [13] were successful in using a neural networks, to attribute authorship. Neural networks were also used by Frank, Kowalczyk, and Ham in [21] to identify poets with an accuracy of up to 90% using letter-sequence features. Diederich et al., in [19], used Support Vector Machines(SVM) to identify with up to 80% accuracy the writings of seven authors from a set of 2,652 newspaper articles written by several authors. De Vel et al. in [34] used SVM to classify 150 E-mail documents from three authors with an accuracy of 80%. Stamatatos et al. in [11] studied the impact of the training data size on authorship-identification performance. They concluded that the classification performance is improved as the number of writings from each author increases, but drops as the number of authors increases.

Framework for Authorship Identification

Authorship identification techniques have achieved many successes in literary and forensic applications. On the other side, E-mail authorship studies are very limited.

E-mails have very special characteristics compared with conventional writings such as literary works and published articles. This makes the application of authorship techniques to E-mail not so straightforward. E-mails are short texts, sometimes less than few words. Ledger and Merriam in [14] established that authorship characteristics would not be significant below 500 words. Forsyth and Holmes in 1996 reported a difficulty in identifying the author of a text with less than 250 words. This is because the vocabulary used in short text is limited and features such as vocabulary richness

may not be apparent. Furthermore, authorship identification for conventional writing deals in general with no more than 10 authors. In E-mail investigations, E-mail users are numerous, but E-mails available from each user may be limited. This makes author identification a big challenge. However, E-mails have some special features that conventional writings do not have. Some of these features are: structural layout traits, unusual language usage, unusual content markers, sub-stylistic features, special greetings and signatures, etc. All these special characteristics can be used to construct an appropriate feature set. So an important step in our E-mail authorship project is the proper selection of the set of features to use.

In our work, we developed an approach for authorship identification where we considered three types of writing-style features: lexical, syntactic and structural. We used machine learning techniques to build classification models to identify authorship.

Used Features Based on the review of previous studies, we use three types of features: lexical, syntactic, and structural.

Lexical features include both character and word-based features [33, 38, 46, 34]. Syntactic features include function words and punctuation. These features are important because they capture author's practices related to sentence structures [33]. Structural features are related to text organization such as paragraph indentation, paragraph size and separation. We list in the following the set of features, we considered:

- Lexical features
  - Character-based features
    * Frequency of upper-case characters.
    * Frequency of upper-case characters at the beginning of sentences.
    * Frequency of white space characters: space, tab, new line

* Frequency of special characters

  ~ , @, #, $, %, ^,&, *, -, _, = ,+, >, <, [,], {, }, /, \, |

- Word-based features
  - Frequency of vocabulary words
  - Frequency of numbers
  - Average sentence length in terms of word
  - Percentage of different words
- Syntactic features
  - Frequency of function words (320 features)
  - Frequency of punctuation
- Structural Features
  - Frequency of blank lines
  - Average size of E-mails
  - Average number of words per paragraphs
  - Presence of a greeting

Authorship Identification Process Authorship identification is achieved using classification techniques provided by the Weka API publicly available at "http://www.cs.waikato.a Some classifiers known to yield good results in authorship identification are the ID3 decision tree [22], backpropagation neural networks [39], and SVM [32]. In our study, we used both SVMs, bayesian networks and the J48 decision tree classifier.

The process of authorship identification is based on the following steps (see Figure 17 for a graphical view of the process):

1. **E-mail Collection:** The E-mail investigator has to collect as set of E-mails written by suspicious known others.

2. **Feature Extraction:** E-mails are preprocessed to extract their message bodies and subjects, then the relevant features are extracted using appropriate text parsers. A profile is constructed for each E-mail. It consists of the set of features structured as a vector of numbers.

3. **Model Generation:**

   The constructed set of E-mails profiles is divided into two subsets. The first one (the training E-mails) is used to train the learning algorithm, the second one (the test E-mails) to evaluate the predicting power of the constructed model. The effectiveness of the constructed model is a function of its ability to correctly classify the test E-mails.

4. **Author Identification:**

   After the classification model is built, it is used to predict the authorship of suspected E-mails.

5.4.3 E-mail Clustering

Clustering is the process of grouping data in semantically close sets to achieve simplification by modeling data using its clusters. The notion of clustering stems from mathematical foundation, such as: statistics and numerical analysis. From a machine learning perspective, clustering is an unsupervised learning process used to retrieve hidden patterns [36]. In practice, clustering is used in data mining applications such as scientific data exploration, computational biology, text mining, marketing, medical diagnostics and many others. In the case of E-mail mining, we use clustering to infer

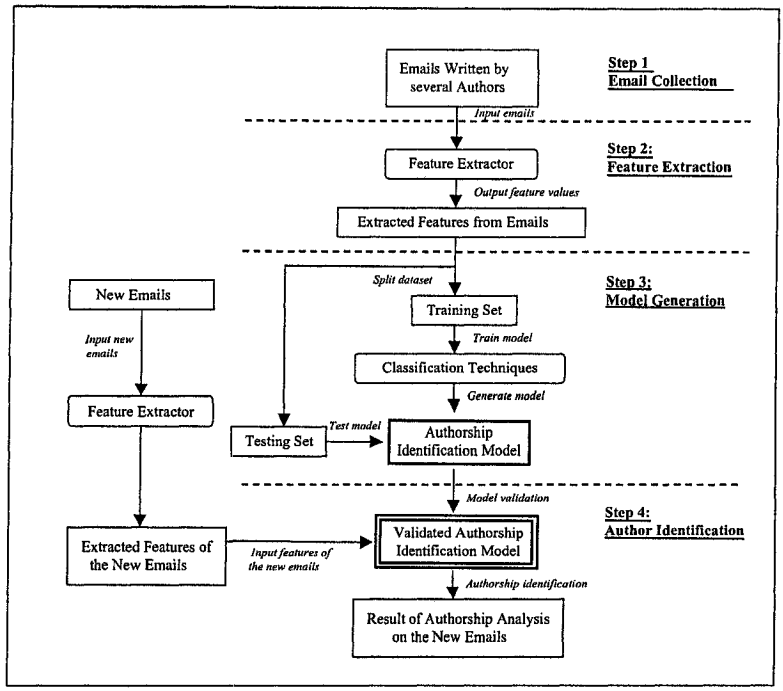

Figure 17: Authorship Identification Process

E-mail's discussion topics, user activities and authorship. We also use clustering to initiate the process of creating classifiers.

To cluster E-mails per discussion topic, E-mails are processed for feature extraction in exactly the same way as for the classification per topic (see section 5.4.1). After clustering, clusters are tagged with some statistics, namely the most and least frequent keywords and the most and least frequent senders and receivers. Most frequent keywords help to identify the topic of discussion, where as most frequent senders and receivers help to tag people with their topics of interest. For authorship identification, E-mails are processed alike the classification for authorship analysis (see section 5.4.2). After clustering, clusters are tagged with the most and least frequent senders and receivers. Since E-mails are clustered according to writing styles, most frequent senders may help to identify authors of anonymous E-mails. Another way of identifying authors is to cluster E-mails with both known and unknown authors.

If an E-mail with an unknown author appears in a cluster where a specific sender is the most frequent, this sender may be the author of the E-mail.

5.5 E-mail Social Networks

Social network analysis is the study of connections between people. E-mail social networks allow to model user E-mail flows and behaviors to detect misuses that manifest as abnormal E-mail behaviors [24]. The social network of an E-mail corpus can be depicted as a graph, where the nodes are senders and receivers and the edges represent the E-mail traffic.

The structure of a person's social network may convey a great deal of information about his/her behavior and interaction with people (friends, colleagues, family members, etc). For instance, we can know how often this person maintains distinct relationships between groups of people, how often, and how much. Do these people have few close friends, and do they have regular interactions? Are these interactions separated based on roles (work, friends, family, etc)? What types of content do they exchange? In an E-mail investigation, social networks can reveal interesting information about potential suspects, like identifying collaborators, suspicious E-mails and suspicious periods of times.

In our framework, we are mainly interested in visualizing E-mail social networks and computing some statistics on E-mail flows. We use three types of graph to depict social networks. The first one, called *user network*, is a graph where nodes are E-mail accounts and the edges represent the E-mail traffic flow (see figure 18). The traffic between individuals can be split according to the different classes of E-mails as computed during the last classification. The second graph, called *domain network*, is a graph where nodes are E-mail domains and the edges represent the E-mail traffic flow. The traffic between domains can also be split according to the different classes

of E-mails computed during the last classification. The last graph, called E-mail temporal network, is the user network augmented with time information about E-mails, to reflect how the E-mails exchange activity evolves with time.

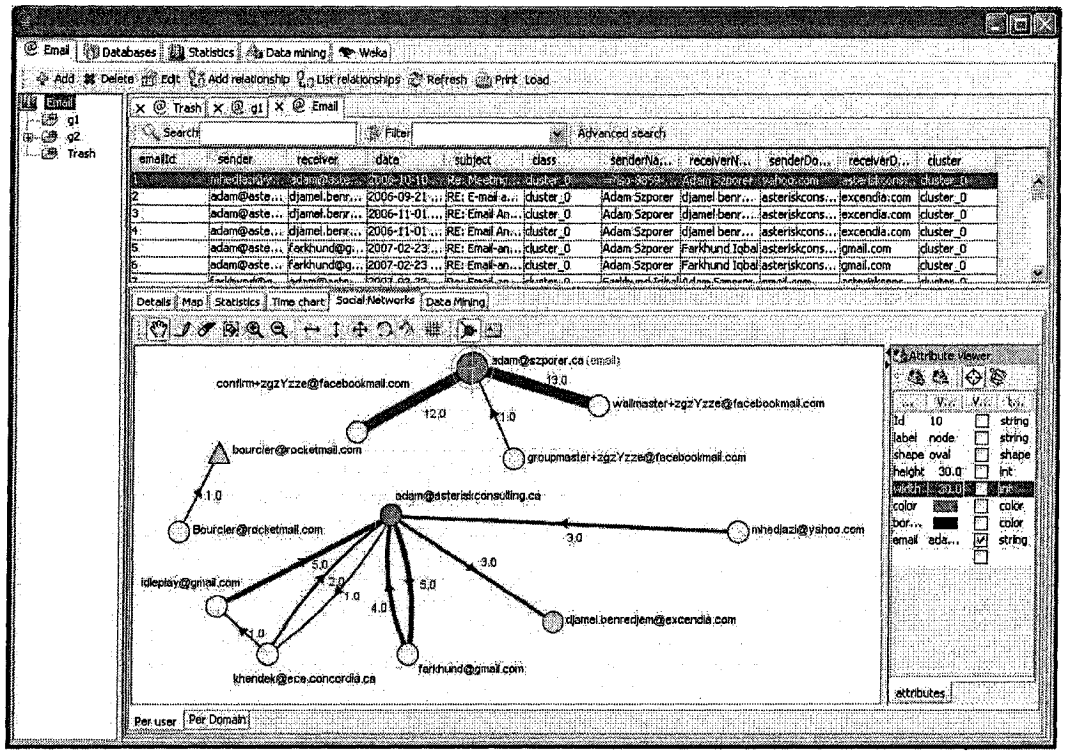

Figure 18: Social Network Viewer

5.5.1 E-mail Geographic Localization

In any investigation, localizing resources and individuals is an important issue. In this perspective, we propose a geographic framework in the form of an interactive map viewer where an investigator can visually localize and explore geographic sites of relevance to the investigation. Some information, related to investigated E-mails can also be rendered directly on the map viewer to guide the investigator. For instance, an E-mail is rendered on the map as an arrow between the sender and receiver E-mail

domains geographic locations. If the physical addresses of the sender and receiver are known, an arrow between these two locations is also drawn.

Other information such as the flows of E-mail between E-mail participants can be rendered directly on the map by appropriately labeling arrows that connect them (see figure 19).

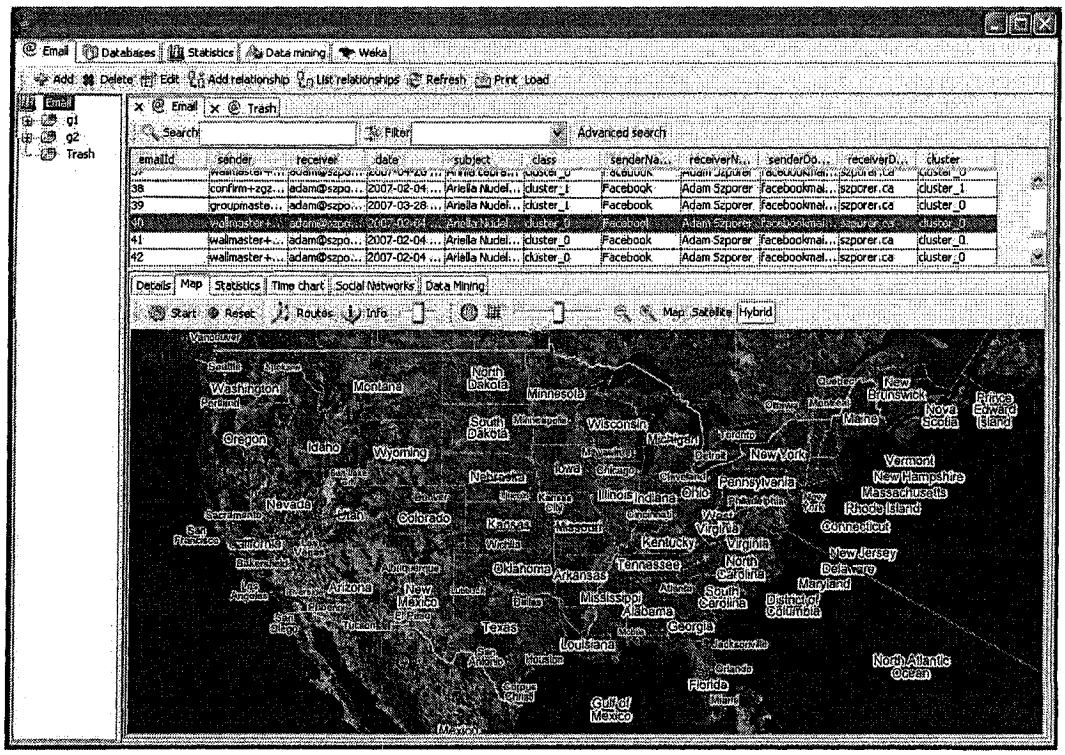

Figure 19: Map Viewer

Map Retrieval

Maps from Internet GIS system, such as: Google, are accessible using a simple browser. The map of region, at a specific zoom level, is recovered tile per tile from the server using simple http requests, then reconstructed locally and displayed using a Javascript code. The Mercator projection combined with some affine transformations allows to locate positions on the map and navigate thought it.

Google offers three different views of a map (see Figure 20):

1. The map view, where the map is rendered using simple drawing,
2. The satellite view with high quality satellite pictures,
3. The hybrid view where both satellite pictures and informative drawings are combined.

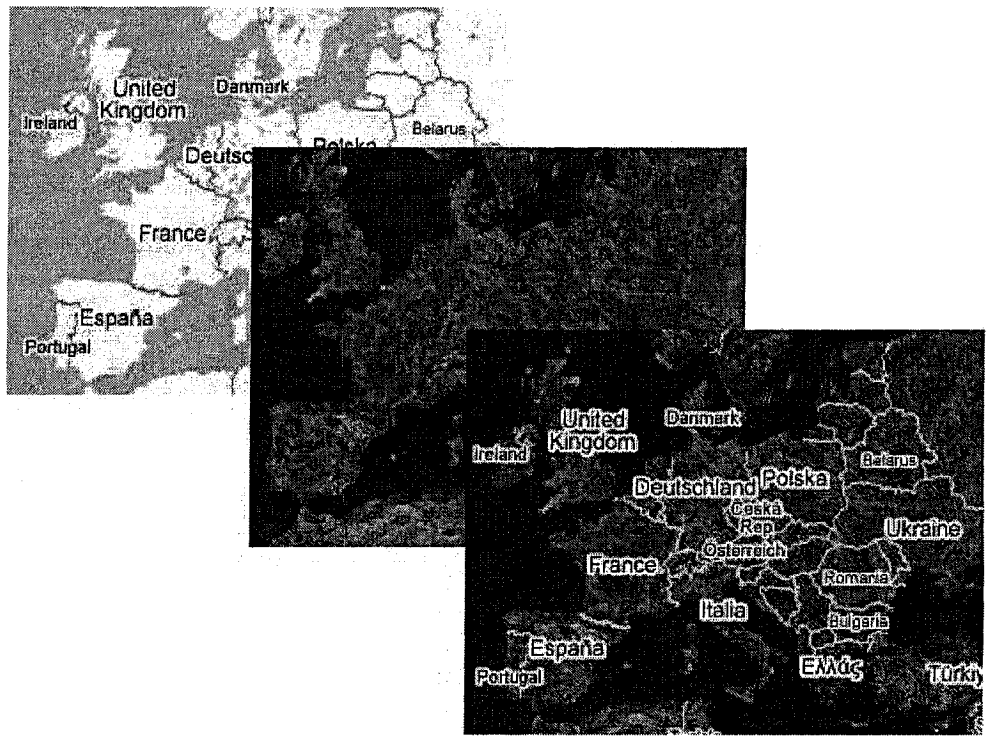

Figure 20: Map, Satellite and Hybrid Views of Maps

Our solution to the map retrieval issue was to design map retrieval and map reconstruction system in the same way as the browser works. To ensure that the solution works even offline, a cache management system has been elaborated to persist already explored maps on the local hard drive. This system can also be set to capture the map of a whole region in a disk and even the whole planet, at different zoom levels

and for the three different views (map, satellite hybrid). Note that, with this solution, we avoid disclosing any information to the outside world. In fact since the map is cached locally, all map manipulations such as localizing specific regions or drawing on top of the map is performed locally.

5.6 E-mail Forensic Analysis Framework

An appropriate computer software is required to assist the investigator to successfully carry out his/her inquiries. In this context, we developed a tool we called *E-mail Forensic Analysis Framework* (EFAF) as an integrated framework, where a security analyst may perform a variety of tasks related to E-mail analysis. Some of these tasks relate to the E-mail classification and clustering, authorship identification, and social networks computation. With functionalities implemented in EFAF, several forensic data sources can be presented together in a common interface. EFAF converts to distributed data sources relevant to a forensic investigation and creates a common easy-to-use view with some useful decision support functionalities for an E-mail investigation. These data sources can also be queried and explored using drill down capabilities to gain insight in the data and extract knowledge. These tools can then be used in a very simple and convincing way for data mining and statistical decision making.

EFAF can be used to provide plenty of analysis ranging from simple statistic models to sophisticated data mining models constructed using the tool Weka. Weka is a powerful open source data mining environment integrated in the EFAF module.

The functionalities of EFAF include:

1. The inspection of E-mail archives and related information by browsing through several data sources coming from different databases

2. Develop data mining models to help classify E-mails in well established categories, or cluster them according to unknown relationships

3. Compute dependencies between users and cliques of users using social networks

4. Search through E-mails using keywords

The EFAF system is programmed in Java and uses several Java technologies such as Java swing, Java mail and JDBC. Swing is used to build the graphical user interface and render information in several visual formats (trees, lists, pictures,...). Java mail API is able to parse E-mails in several file formats and extract relevant information. JDBC allows to store these information in any JDBC compliant database such as Oracle , MySql and SQLserver.

EFAF is composed of five sub modules, which can be used separately or together, to build and explore decision support models. These modules are:

1. Database browser,

2. Statistics explorer,

3. Data mining explorer,

4. Weka tool,

5. E-mail Explorer.

We start by giving a brief description of the first four modules then dig into the details of the E-mail explorer and the capabilities it provides for an E-mail investigator.

5.6.1 Inter Database Browser

As its name suggests, the inter database browser allows to interact with several heterogeneous databases (Oracle, Sybase SQL server..) from a single interface, with a drill down capability. This capability allows to navigate through the data in a database, as presented in tables or views, by using a tree structure. The navigation in a database is based on relationships between tables as represented by foreign keys. The inter database browser extracts the relationship information from the metadata saved in the database, then use it to construct a physical entity relationship diagram. To allow the navigation among several databases, the user can supply and save relationships between tables of different databases to the inter database browser. These inter database relationships are then used to create connections between the entity relationship diagrams of several databases and expose them as a single entity relationship diagram for all connected databases.

The main functionalities implemented in the database navigator are as follows:

1. Dynamic creation of connections to JDBC compliant databases
2. Presentation of data in tables and views with a drill down capability
3. Creation of relationships between different data bases to let the drill down capability among several databases
4. Issuing and saving SQL statements
5. Preparing data sets to create data mining models using the tool Weka

Dynamically Create Connections to JDBC Compliant Databases

This capability allows the user to view, at the same time, to the contents of several JDBC-compliant databases (see figure 21). The user may easily establish a connection

to a new local or remote database server. The specification of a connection requires to set the appropriate JDBC driver, the connection string which mainly indicates the address of the database server on the network, and finally an appropriate username and password. The user can also add JDBC drivers dynamically to the environment, drop drivers and database connection as needed.

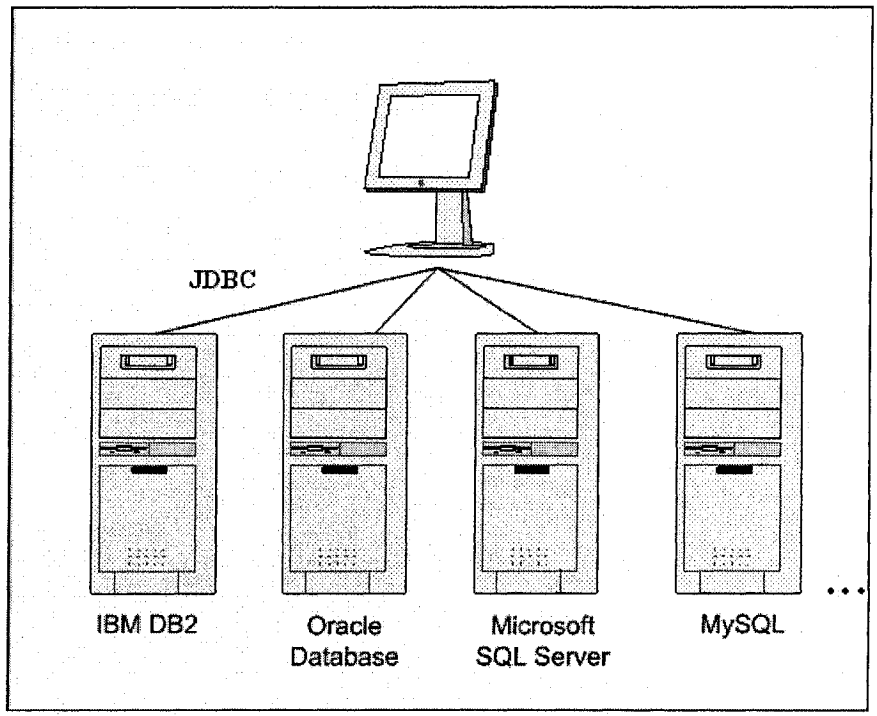

Figure 21: Client Server Architecture of the Inter Database Browser

Explore the Data in Tables and SQL Queries with a Drill Down Capability

The data exploration in a database is achieved using a tree structure where database connections appear at higher levels followed by tables and views (SQL queries). Browsing a database connection opens a new level where tables and views are displayed, each one in a separate branch. Browsing a table (a view) opens a new level where all records in the table (the view) are displayed, each one in a separate branch.

Browsing a record, opens a new level where attributes of the record followed by their values are displayed on deferent branches, with extra branches associated with foreign keys. These extra branches, allow to achieve a drill down capability by offering the possibility to directly navigating in the records referenced by the corresponding foreign keys.

Create Relationships between Different Databases

To make the drill down capability span among databases, the user can dynamically create relationships between tables, even if these tables are in different databases. When some fields in a database table are related (i.e., have the same meaning) to some fields in another database table, a relationship between these tables can be created and persisted. This relationship will play the same role as the one played by foreign keys in establishing connections between tables in the same database. A relationship is set by specifying the two related tables and their corresponding related fields.

Issue and Persist SQL Statements

The users can query a database by issuing SQL statements from the interface. Whenever an SQL statement is of some interest, it can be persisted for future usage. This feature is similar to the creation of views in a database, except that the created SQL statement is saved locally in the file system instead of the database. To persist an SQL statement, the user has to provide a name for it and a text description, if needed. Persisted SQL statements are listed in a tree structure in categories. These categories are dynamically created by the user. A statement can be executed at any moment, the results may be displayed in a table or chart and can printed or saved.

Prepare Data Sets to Create Data Mining Models Using Weka

To construct a data mining model using the tool Weka, the requited data need to be organized in a data set. The user may use the functionalities implemented in the inter database browser (see figure 23) to create data sets from connected databases or he/she can also provide data sets from external sources. A data set is simply a set of records similar to a table in a database.

Weka native data storage method is ARFF format (see figure 22). It is easy to convert a spreadsheet or a result of an SQL query to ARFF format. The bulk of an ARFF file consists of a list of the instances. The attribute values for each instance are separated by commas.

```
@relation weather

@attribute outlook {sunny, overcast, rainy}
@attribute temperature real
@attribute humidity real
@attribute windy {TRUE, FALSE}
@attribute run {yes, no}

@data
sunny,85,85,FALSE,no
sunny,80,90,TRUE,no
overcast,83,86,FALSE,yes
rainy,70,96,FALSE,yes
rainy,68,80,FALSE,yes
rainy,65,70,TRUE,no
overcast,64,65,TRUE,yes
sunny,72,95,FALSE,no
sunny,69,70,FALSE,yes
rainy,75,80,FALSE,yes
sunny,75,70,TRUE,yes
overcast,72,90,TRUE,yes
overcast,81,75,FALSE,yes
rainy,71,91,TRUE,no
```

Figure 22: Example of an ARFF File

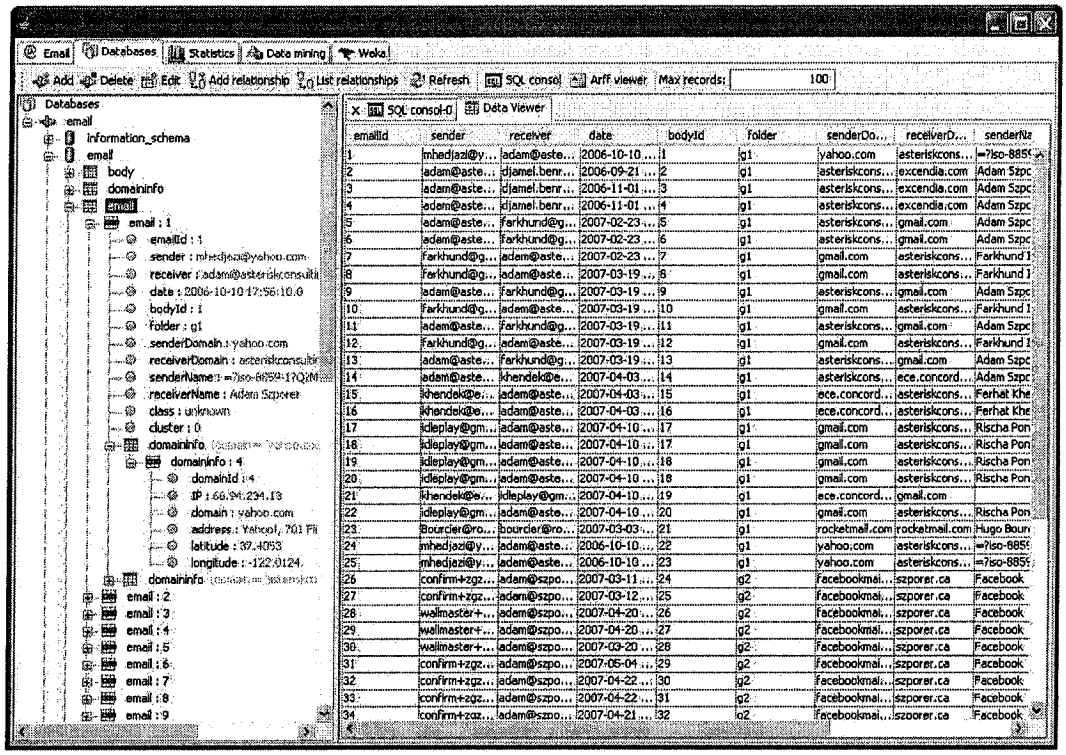

Figure 23: Inter Database Explorer

5.6.2 Statistics Explorer

The statistic explorer allows to associate SQL query with very elegant charts to gain better insight from the data. Charts are constructed using the ExpressChart API. This API provides a chart viewer with very sophisticated interactive functionalities ranging from simple rotation, resizing, zooming operations to switching between several possible views in two and tree dimensions.

Charts are displayed in a tree structure and groups of categories. The user can dynamically create new categories and new charts (see figure 24)), and display them by a simple click of a mouse.

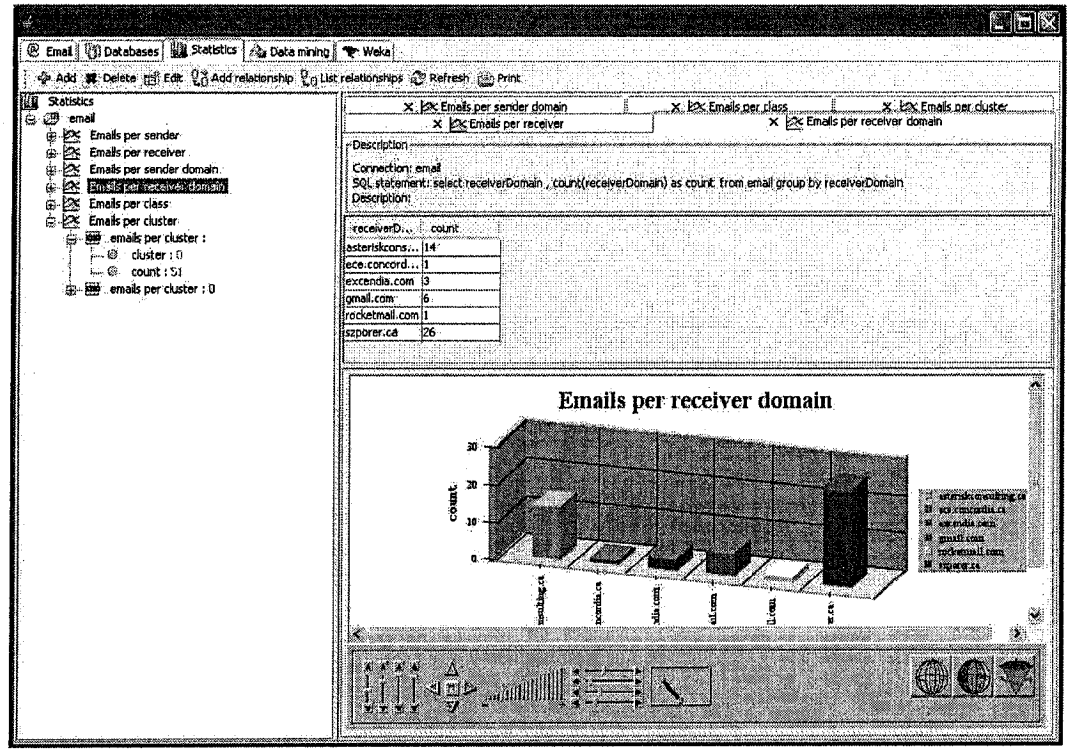

Figure 24: Statistics Explorer

5.6.3 Data Mining Explorer

Traditional business intelligence tools such as reports, interactive query and reporting only report on what has happened in the past. They report on historical activities and current status values. Online Analytical Processing (OLAP) provides rapid drill-down for fast, more detailed information, roll ups, forecasting and trend analysis but usually only for averages, sums, trends, and group-by aggregates. None of these approaches can provide the deeper insights and views to the future like data mining. Data mining uses machine-learning techniques, developed in the last decade and doesn't suffer from the same limitations. Data mining sifts deeper into data to discover information, patterns, factors, clusters, profiles, and predictions that remain "hidden" in the data. It allows to discover new insights, segments and associations, make more

accurate predictions, find the variables that most influence business activities, detect anomalies, and in general, extract more information from the data.

The data mining explorer (see figure 25) enables the user to directly interrogate data mining models constructed using the tool Weka. The models are organized in a tree structures in three main categories: Classifiers, Clusters and Association rules. Data mining models can be dynamically constructed with the tool Weka, then integrated in the data mining explorer in the appropriate category. Further categories can be added dynamically to the data mining explorer tree to customize the organization of its data mining models .

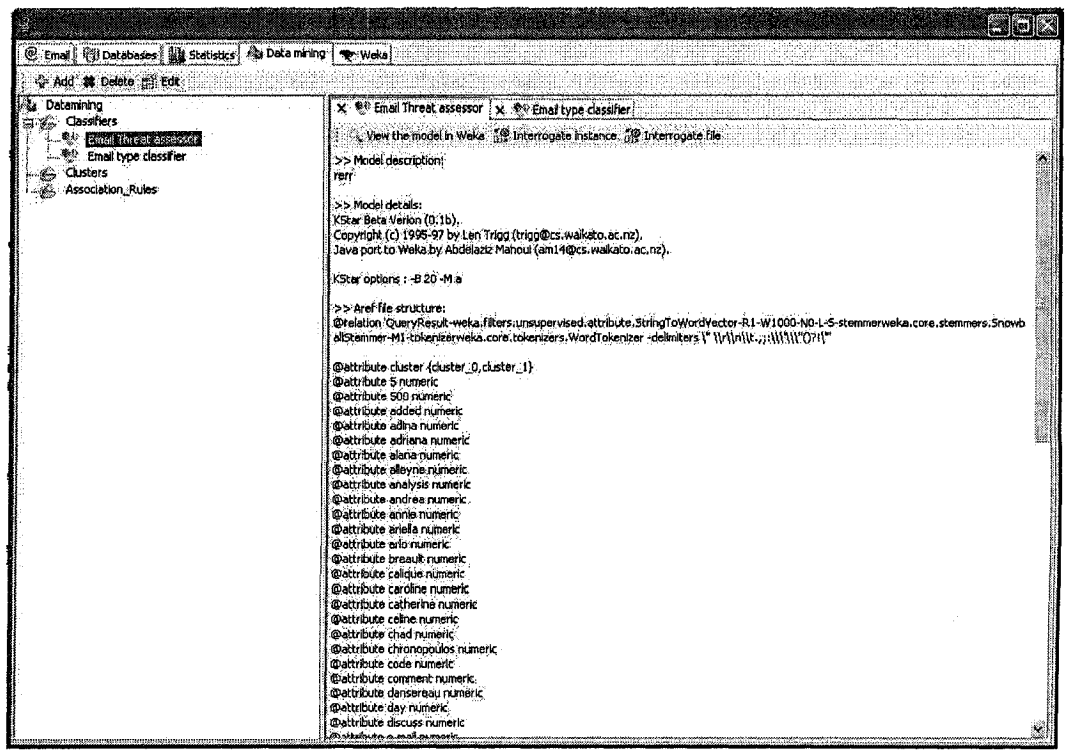

Figure 25: Data Mining Explorer

5.6.4 Weka Submodule

The Weka submodule is actually a complete integration of the Weka workbench. Weka is a collection of state-of-the-art machine learning algorithms and data preprocessing tools. It provides extensive support for the whole process of experimental data mining, including preparing the input data, evaluating learning schemes statistically, and visualizing the input data and the result of learning. With a wide variety of learning algorithms, it also includes a wide range of preprocessing tools. It is designed so that one can quickly try out existing methods on new data sets in flexible ways. This diverse and comprehensive toolkit is accessed through a common interface so that its users can compare different methods and identify those that are most appropriate for the problem at hand.

Weka was developed at the University of Waikato in New Zealand, and the name stands for *Waikato Environment for Knowledge Analysis.* The system is written in Java and distributed under the terms of the GNU General Public License and available from http://www.cs.waikato.ac.nz/ml/weka. It runs on almost any platform and has been tested under Linux, Windows, and Macintosh operating systems, and even on a personal digital assistant. It provides a uniform interface to many different learning algorithms, along with methods for pre and postprocessing and for evaluating the result of learning schemes on any given data set.

Weka provides implementations of learning algorithms that can be easily applied to a data set. It also includes a variety of tools for transforming data sets. For instance, it is possible to preprocess a data set, feed it into a learning scheme, and analyze the resulting classifier and its performance, all without writing any program code at all. The workbench includes methods for all the standard data mining problems: regression, classification, clustering, association rule mining, and attribute selection.

Getting to know the data is an integral part of the work, and many data visualization facilities and data preprocessing tools are provided. All algorithms take their input in the form of a single relational table in the ARFF format (a very simple text format), which can be read from a file or generated by a database query. One way of using Weka is to apply a learning method to a data set and analyze its output to learn more about the data. Another is to use learned models to generate predictions on new instances. A third is to apply several different learners and compare their performance in order to choose one for prediction. The learning methods are called *classifiers* and in the interactive Weka interface they are selected from a menu. Many classifiers have tunable parameters, which are accessible through a property sheet or *object editor*. A common evaluation module is used to measure the performance of all classifiers. Implementations of actual learning schemes are the most valuable resource that Weka provides. But tools for preprocessing the data, called *filters* come a close second. Like classifiers, filters are selected from a menu and tailored to the needed. Weka also includes implementations of algorithms for learning association rules, clustering data for which no class value is specified, and selecting relevant attributes in the data.

WEKA Usage

The easiest way to use Weka is through a graphical user interface called the *Explorer* (see figure 26). This gives access to all of its facilities using menu selection and form filling. For example, the user can quickly read in a data set from an ARFF file (or spreadsheet) and build a decision tree from it. But learning decision trees is just the beginning: there are many other algorithms to explore. The explorer interface helps to do just that. It guides the user by presenting choices as menus, by forcing him to work in an appropriate order by graying out options until they are applicable, and by

presenting options as forms to be filled out. Helpful *tool tips* pop up as the mouse passes over items on the screen to explain what they do. Sensible default values ensure that the user can obtain results with a minimum of effort, but he has to think about what he is doing to understand what the results mean. There are two other graphical user interfaces to Weka. The *Knowledge Flow* interface (see figure 27) allows to design configurations for streamed data processing. A fundamental disadvantage of the explorer is that it holds everything in the main memory when a data set is open, it immediately loads it all in. This means that it can only be applied to small to medium-sized problems. However, Weka contains some incremental algorithms that can be used to process very large data sets. The knowledge flow interface lets the user drag boxes representing learning algorithms and data sources around the screen and join them together into the needed configuration. It enables the user to specify a data stream by connecting components representing data sources, preprocessing tools, learning algorithms, evaluation methods, and visualization modules. If the filters and learning algorithms are capable of incremental learning, data will be loaded and processed incrementally.

Weka's third interface, the *Experimenter* (see figure 28), is designed to help answer a basic practical question when applying classification and regression techniques: which methods and parameter values work best for the given problem? There is usually no way to answer this question a priori, and one reason the workbench was developed was to provide an environment that enables users to compare a variety of learning techniques. This can be done interactively using the explorer. However, the experimenter allows to automate the process by making it easy to run classifiers and filters with different parameter settings on a corpus of data sets, collect performance statistics, and perform significance tests. Advanced users can employ the experimenter to distribute the computing load across multiple machines using Java

remote method invocation (RMI). In this way one can set up large-scale statistical experiments and leave them to run. Behind these interactive interfaces lies the basic functionality of Weka. This can be accessed in raw form by entering textual commands, which gives access to all features of the system. When Weka is started, it is possible to choose among four different user interfaces: the explorer, the knowledge flow, the experimenter, and the command-line interface (see Figure 29). An important resource when working with Weka is the online documentation, which has been automatically generated from the source code and concisely reflects its structure. It gives the only complete list of available algorithms because Weka is continually growing and being generated automatically from the source code, the online documentation is always up to date. Moreover, the online documentation becomes essential to access the library from a separate Java programs.

WEKA Screenshots

5.6.5 E-mail Explorer

The E-mail explorer (see figure 30) allows for a multi-stage analysis of a set of E-mails using social networks, text mining techniques, geographic rendering and statistical analysis, to gain in depth view into the underlying information.

The E-mail explorer works with a database (DB) back-end for a fast and convenient analysis of E-mail data. E-mails are first loaded from their raw files using the Java E-mail API, then structured and stored in a database according to the schema, as shown in Figure 31.

The E-mail table collects information about the sender, receiver and the date and time when the E-mail was sent. In addition, the table archives E-mail domains and other information related to the clustering and classification of E-mails using data mining tools. When an E-mail has many receivers, the table inserts a row for each

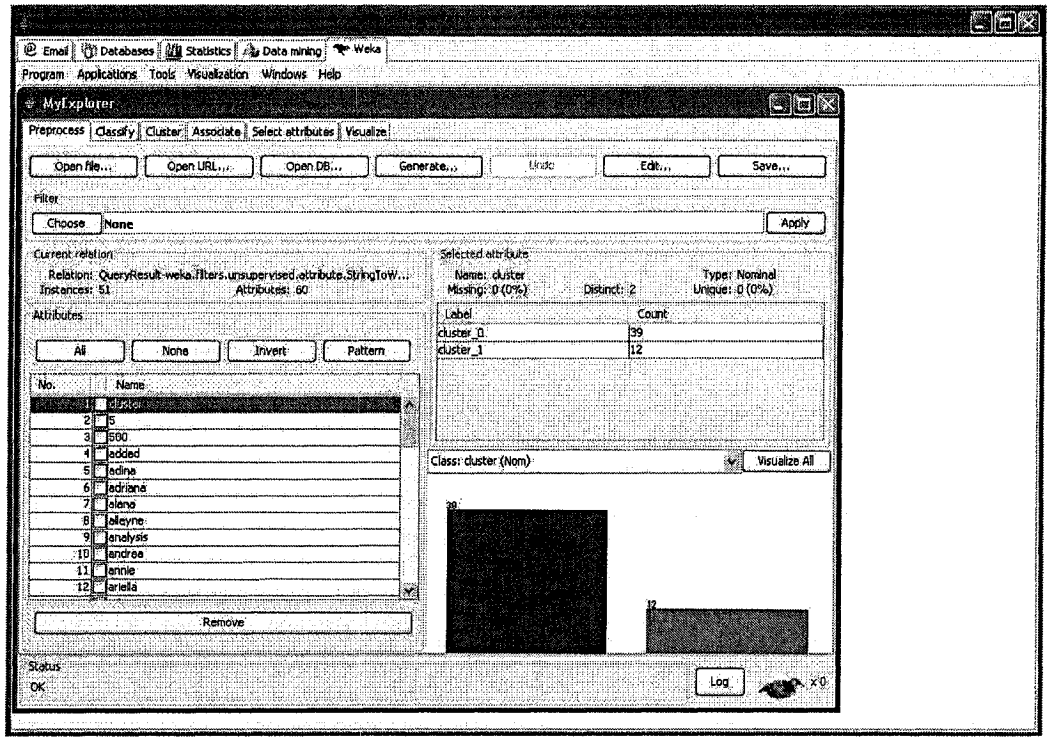

Figure 26: Weka Explorer Panel

receiver. This is because, conceptually, there is no difference between sending an E-mail to many people at once and sending it to each one individually.

Each E-mail is associated with a folder name attribute, for a virtual classification of E-mails in folders. The body table stores information of E-mail body composed of the subject, the E-mail message, and its raw format (by raw we mean the E-mail without formatting).

The domain table stores information about domain addresses and geographic coordinates. This information is used to get geographical localization of E-mail information on a map.

The user table stores information about E-mail users and their physical addresses.

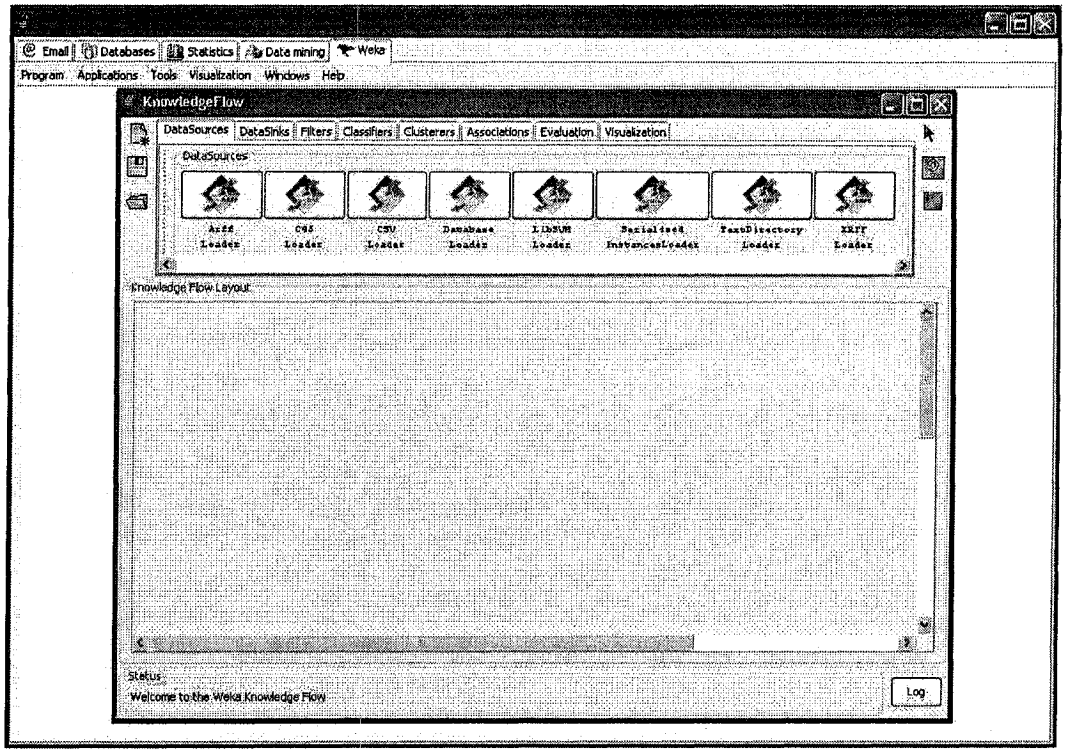

Figure 27: Weka Knowledge Flow Interface

5.6.6 E-mail Browsing and Exploration

The E-mail browser provides a structured view of E-mails in folders. The content of each folder may be viewed separately or together. A folder is viewed in a folder viewer where the list of its E-mails is displayed. Several folder viewers can be opened at the same time with the possibility to perform E-mail movement between them.

An E-mail folder viewer offers some classical functionalities related to as E-mail sorting, searching and also some advanced functionalities related to E-mail data mining and social networks visualization.

Advanced functionalities are presented in the following view sets:

1. Details editor

2. Map viewer

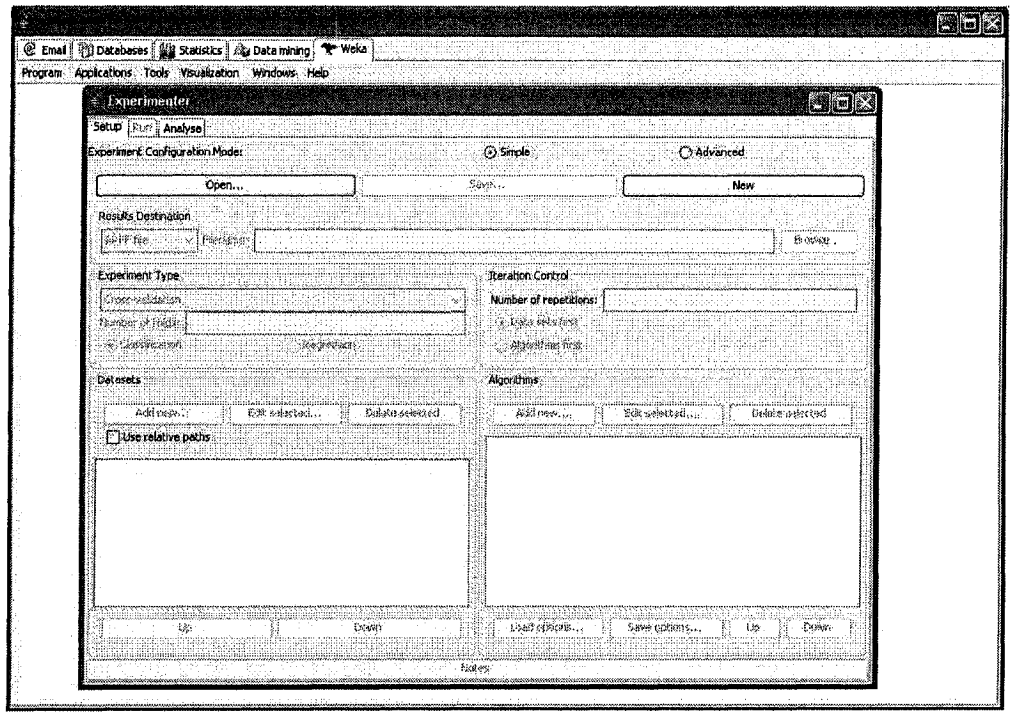

Figure 28: Weka Experimenter Interface

3. Statistic viewer

4. Social network viewer

5. Data mining viewer

Details Editor

The details editor (see figure 32) offers three different views of an E-mail content: text, html and raw format. The text view shows the text content of an E-mail, the HTML views shows its HTML view if it has an HTML content, and the raw view, presents the E-mail in its raw state, as it was received.

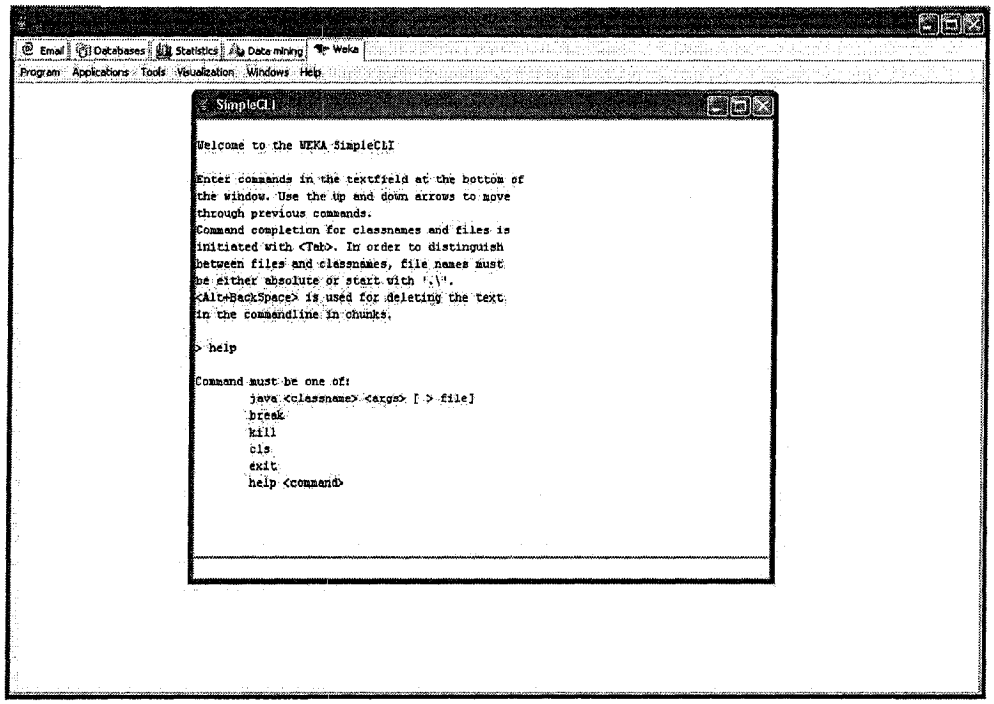

Figure 29: Command-Line Interface

Map Viewer

The map viewer (see figure 33) renders some E-mails information directly on real geographic maps using an internet GIS. For instance, when an E-mail is selected in the folder viewer, an arrow is drawn between the sender and receiver E-mail domains geographic locations. If the physical addresses of the sender and receiver are known, an arrow between these two locations is also drawn.

Other information such as the flows of E-mails between E-mail participants can be rendered directly on the map by appropriate labeling of the arrows that connect them.

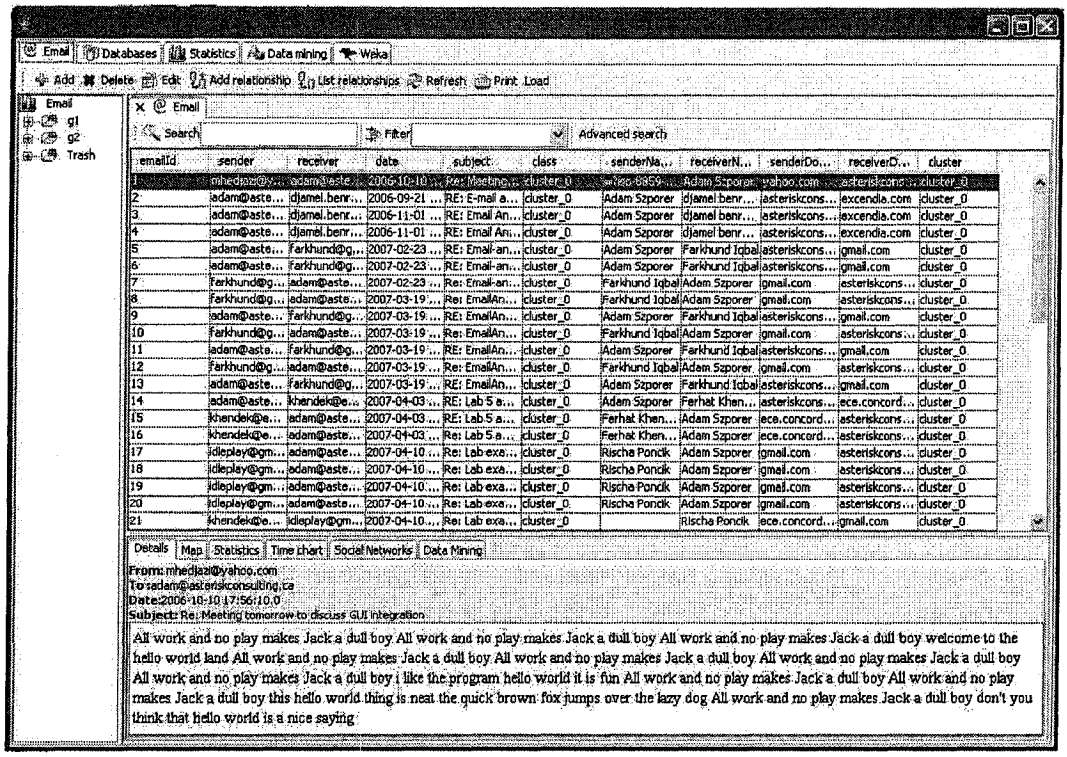

Figure 30: E-mail Explorer

Statistics Viewer

We compute several statistics on E-mail accounts and E-mail traffic, view them using appropriate charts for better and easy interpretation. The computed statistics include the distributions of E-mail flow per sender, recipient and domain. Other statistics relate to the distribution of E-mail per class and cluster in a selected periods of times. Statistics models are created in the statistic explorer then automatically inserted in the static viewer (see figure 34). A statistic model is created by specifying an SQL query to extract relevant data and associating it with the appropriate chart.

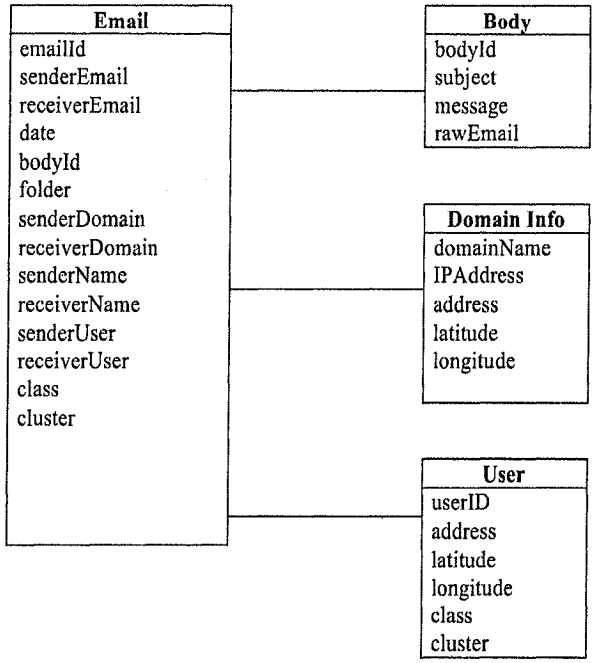

Figure 31: E-mail Database Schema

Social Network Viewer

In order to study E-mail flows between users and groups of users, we compute a set of cliques and display them in the form of networks (see figure 35), through a fully fledged graph editor. The user may explore those networks and transform them if needed to gain insight in the dynamics of E-mail traffic. These capabilities offer the possibility to identify groups of related E-mail accounts that participate with each other in common E-mail communications.

Different views of E-mail traffic can be displayed. The flow of E-mail/s between users and groups of users is reflected by the thickness of the links that connect them. This thickness increases with the intensity of the E-mail traffic. Colors are also used to easily identity the different classes and clusters of E-mails.

Figure 32: Details Editor

Data Mining Viewer

The *Data mining viewer* (see figure 36) enables to build machine-learning models from different sets of E-mails, using several data mining models. It also allows to evaluate these models on other sets of E-mails. The functionalities of the Data mining viewer are split into two parts, classification and clustering. Classification allows to build data decision models on sets of E-mails which are already classified, where as clustering seeks to identify hidden relationships between unclassified E-mails.

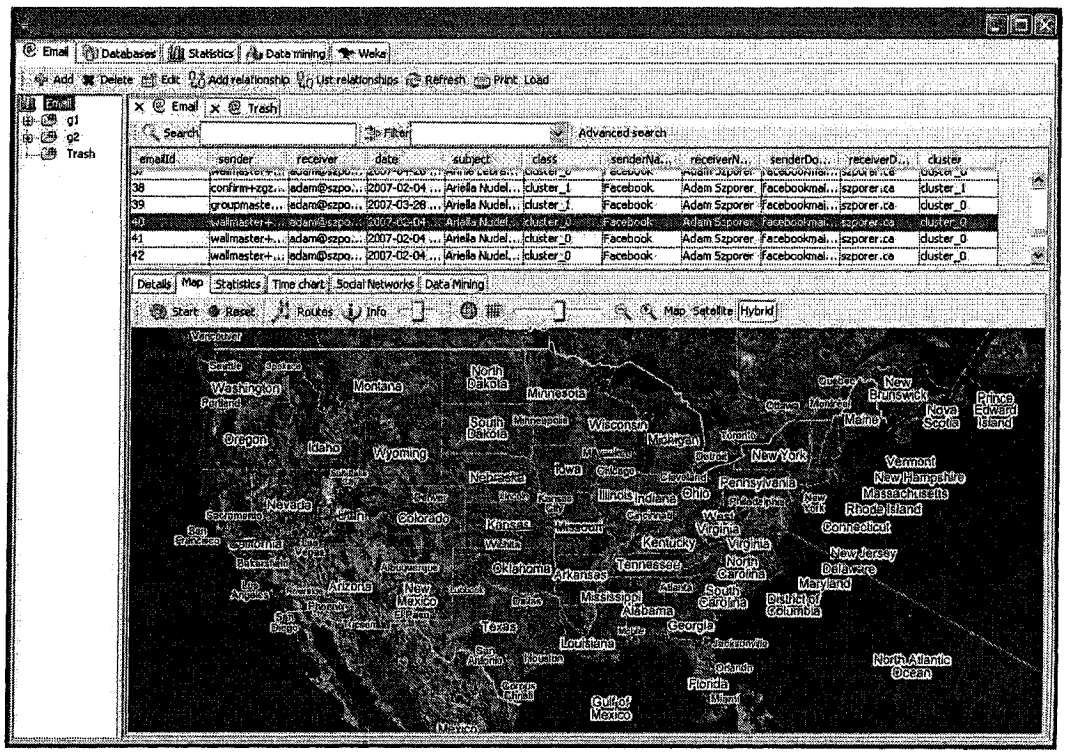

Figure 33: Map Viewer

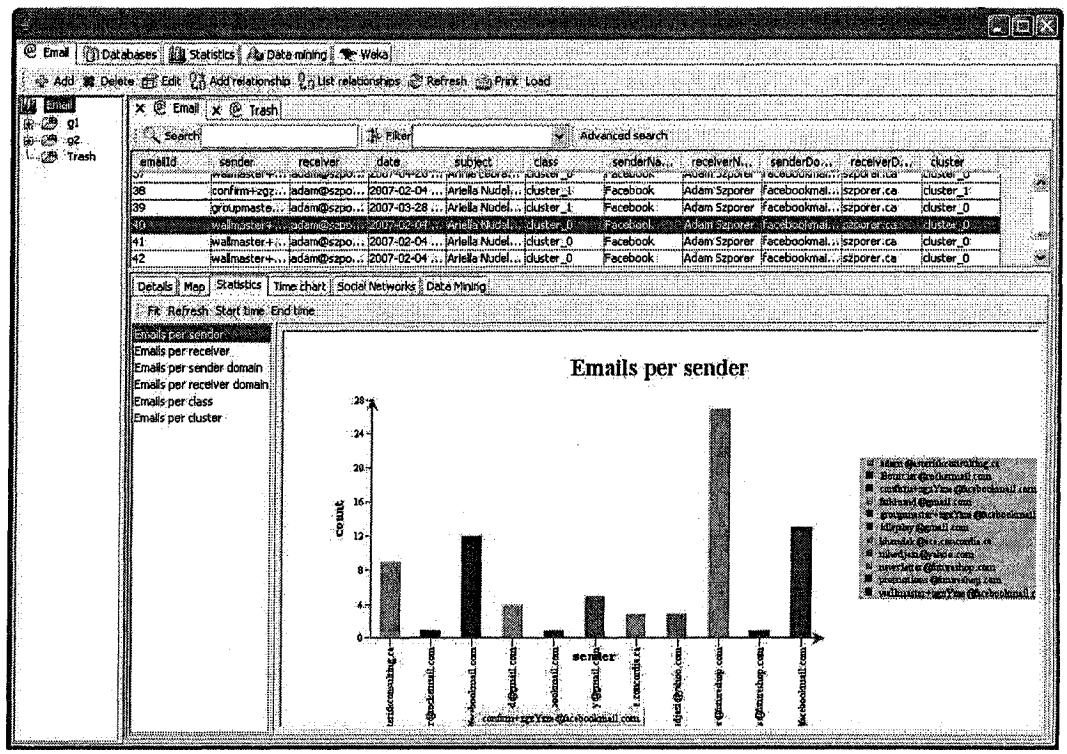

Figure 34: Statistics Viewer

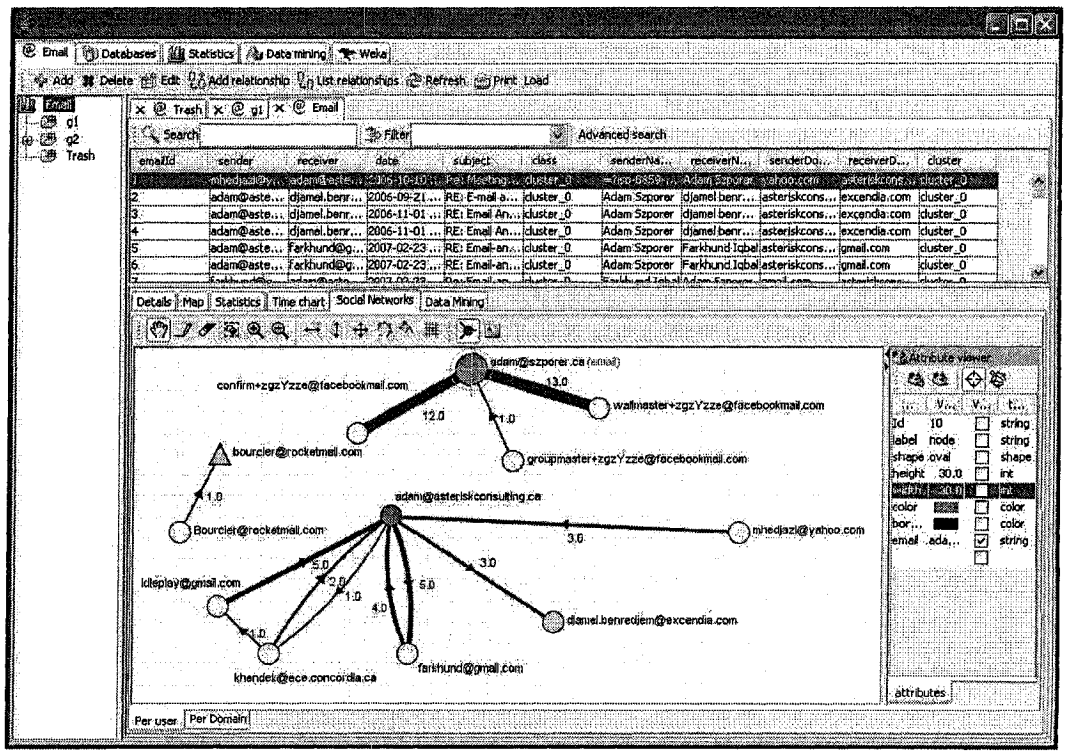

Figure 35: Social Networks Viewer

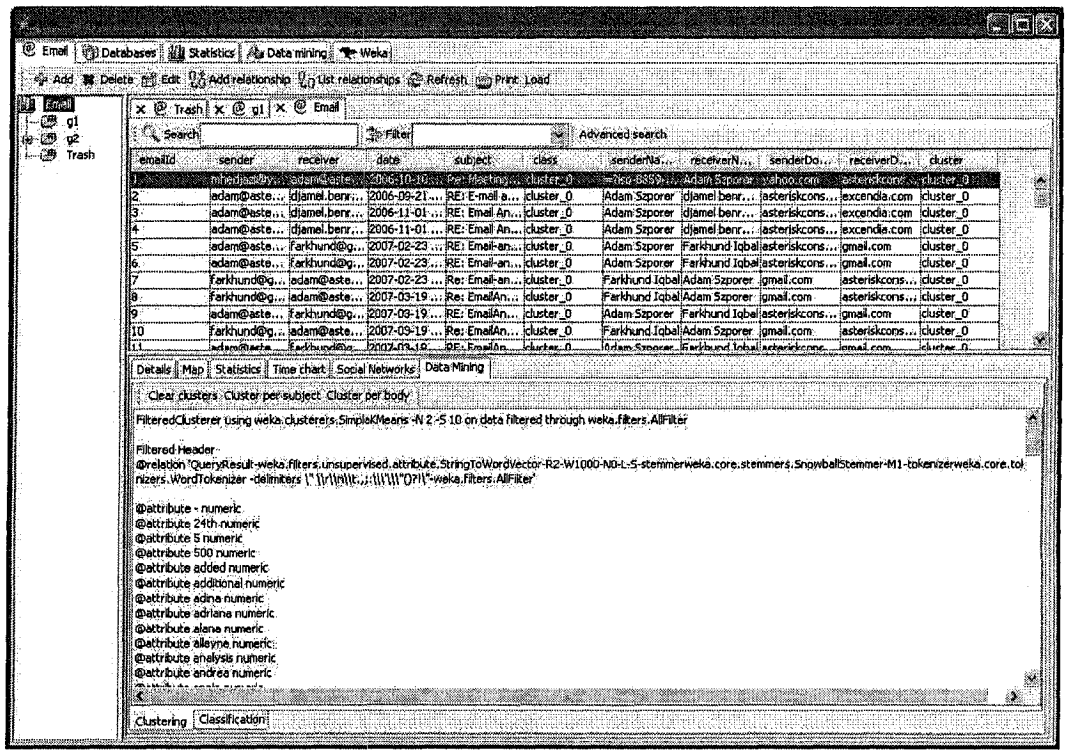

Figure 36: Data Mining Viewer

Chapter 6

Conclusion

Security is emerging as one of the most important challenges that are ever faced by computing research. Individuals, corporations and organizations are exposed to various attacks that lead to severe consequences ranging from life-threat, financial loss to information theft, denial of service, etc. With the growing pace of Internet technology, previously developed security mechanisms are becoming ineffective to hinder the unauthorized and disruptive use of cyber resources and thereby protecting the safety and privacy of cyber environments. In this context, cyber-forensics plays a major role by providing scientifically proven methods to detect and respond to cyber incidents by gathering, processing, and interpreting digital evidence in order to establish a conclusive description of cyber-crime activities. Even though cyber-forensics is at its infancy as a discipline, it is extremely pertinent as a security tool now-a-days more than ever. Accordingly, there is a desideratum for designing and implementing a framework that incorporates dedicated processes, methodologies, and techniques to enclose the whole cyber-forensics life cycle. Such a framework will help the investigator in verifying incidents, responding to deliberate attacks, collecting clues/evidence, and analyzing the evidence.

The contribution to this thesis is twofold. The first set of contributions is relative

to cyber-forensic processes. In this respects, this thesis presents a detailed study of the state of the art processes. It also focusses on a comparative study of the existing digital forensic processes. With the detailed understanding of the potential and pitfalls, a proposal is provided for a new process that may leverage the major features of existing processes, while fixing their identified deficiencies. In addition, the unified cyber-forensic process provides new features which can be summarized as follows :

- Detailed principles
- Investigative phases
- Quality assurance
- Control and iteration
- Description language
- Computer support

The second set of contributions is relative to cyber-forensic analysis. Actually, an emphasis has been placed on e-mail forensic analysis. Database and data mining techniques are used to provide practical insights and useful abstractions that may help the digital forensic investigations. The e-mail forensic analysis framework provides the following functionalities:

- Statistical analysis
- Geographical localization
- Social Networks
- Data mining analysis

- Authorship identification

As the future work, we first require the incorporation of this process into the software framework, being developed at the Computer Security Laboratory. Besides, we suggest the extension of the e-mail analysis capabilities by having a high-level query language together with the underlying execution engine and interface, which will allow the forensic investigator to browse and navigate inside the e-mail evidence and answer important questions that are related to the ongoing investigations. In addition, we propose the elaboration of a language for the description, planning and execution of cyber-forensic activities.

Bibliography

[1] Ó. Séamus. An Extended Model of Cybercrime Investigations. *International Journal of Digital Evidence*, 3(1):1–22, Summer 2004.

[2] B. Carrier. Basic Digital Foresnic Investigation Concepts. *Digital investigation and Digital Forensic Basics*, 2006.

[3] N. Beebe and J. Clark. A Hierarchical, Objectives-Based Framework for the Digital Investigations Process. *Digital Investigation*, 2(2):146–166, June 2005.

[4] C. Brian and S. Eugene. Getting Physical with the Digital Investigation Process. *International Journal of Digital Evidence*, 2(2):1–20, Fall 2003.

[5] J. Burrows. Word Patterns and Story Shapes: The Statistical Analysis of Narrative Style. *Literary and Linguistic Computing*, 2:61–67, 1987.

[6] C. Martindale and D. McKenzie. On the Utility of Content Analysis in Author Attribution: The Federalist. *Computer and the Humanities*, 29:259–270, 1995.

[7] E. Casey. *Digital Evidence and Computer Crime.* Academic Press, San Diego, California USA, 2 edition, 2004. Chapter 4.

[8] D. Holmes. The Evolution of Stylometry in Humanities. *Literary and Linguistic Computing*, 13(3):111–117, 1998.

[9] D. Holmes and R. Forsyth. The Federalist Revisited: New Directions in Authorship Attribution. *Literary and Linguistic Computing*, (10):111–127, 1995.

[10] D. Mealand. Correspondence Analysis of Luke. *Literary and Linguistic Computing*, 10:171–182, 1995.

[11] E. Stamatatos and N. Fakotakis and G. Kokkinakis. Computer-Based Authorship Attribution without Lexical Measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[12] F. Tweedie and R. Baayen. How Variable May a Constant be? Measures of Lexical Richness in Perspective. *Computers and the Humanities*, (32):323–352, 1998.

[13] F. Tweedie and S. Singh and D. Holmes. Neural Network Applications in Stylometry: The Federalist Papers. *Computers and the Humanities*, 30(1):1–10, 1996.

[14] G. Ledger and T. Merriam. Shakespeare, Fletcher, and the two Noble Kinsmen. *Literary and Linguistic Computing*, 9:235–248, 1994.

[15] G. Yule. On Sentence Length as a Statistical Characteristic of Style in Prose. *Biometrika*, (30):363–390, 1938.

[16] G. Yule. The Statistical Study of Literary Vocabulary. *Cambridge University Press*, 1944.

[17] R. Ieong. FORZA – Digital Forensics Investigation Framework that Incorporate Legal Issues. *Elsevier Digital Investigation*, 3:29–36, 2006.

[18] J. Burrows. An Ccean Where Each Kind...: Statistical Analysis and Some Major Determinants of Literary Style. *Computer and the Humanities*, (23):309–321, 1989.

[19] J. Diederich and J. Kindermann and E. Leopold and G. Paass. Authorship attribution with Support Vector Machines. *Applied Intelligence*, 19:109–123, 2000.

[20] J. Farringdon. Analyzing for Authorship: A Guide to the Cusum Technique. *University of Wales Press*, 1996.

[21] J. Hoorn and S. Frank and W. Kowalczyk and F. Ham. Neural Network Identification of Poets using Letter Sequences. *Literary and Linguistic Computing*, 14(3):311–338, 1999.

[22] J. Quinlan. Induction of Decision Trees. *Machine Learning*, 1(1):81–106, 1986.

[23] T. Joachims. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of European Conf. Machine Learning (ECML'98)*, pages 137–142, 1998.

[24] M. Bhattacharyya and S. Hershkop and E. Eskin and S. Stolfo. MET: An Experimental System for Malicious Email Tracking. In *Proceedings of the 2002 New Security Paradigms Workshop (NSPW-2002)*, 2002.

[25] M. Corney and O. de Vel and A. Anderson and G. Mohay. Gender-preferential text mining of e-mail discourse. In *Proceedings of the 18th annual Computer Security Applications Conference (ACSAC'02)*, Las Vegas, NV, USA, 2002.

[26] M. Koppel and S. Argamon and A. Shimoni. Automatically Categorizing Written Texts by Author Gender. *Literary and Linguistic Computing*, 4(17):401–412, 2002.

[27] M. Newman and S. Forrest and J. Balthrup. Email Networks and the Spread of Computer Viruses. *The American Physical Society*, 2002.

[28] M. Reith and C. Carr and G. Gunsch. An Examination of Digital Forensic Models. *International Journal of Digital Evidence*, 1(3):1–12, Fall 2002.

[29] K. Mandia, C. Prosise, and M. Pepe. *Digital Evidence and Computer Crime.* Brandon A. Nordin, Emeryville, California USA, 2 edition, 2004. Chapter 2.

[30] Mitchell97. *Machine Learning.* McGraw-Hill, 1997.

[31] F. Mosteller and D. Wallace. *Applied Bayesian and Classical Inference: The Case of the Federalist Papers (2nd ed.).* Springer-Verlag, New York, 1964.

[32] N. Cristianini and J. Shawe-Taylor. An introduction to Support Vector Machines. *Cambridge University Press*, 2000.

[33] O. de Vel. Mining E-mail Authorship. In *ACM International Conference on Knowledge Discovery and Data Mining (KDD)*, 2000.

[34] O. de Vel and A. Anderson and M. Corney and G. Mohay. Mining E-mail Content for Author Identification Forensics. *SIGMOD Record*, 30(4):55–64, 2001.

[35] G. Palmer. Report from the First Digital Forensic Research Workshop. Technical Report DTR-T001-01 Final, Digital Forensic Research Workshop (DFRWS), New York, November 2001. A Road Map for Digital Forensic Research.

[36] R. Agrawal and J. Gehrke and D. Gunopulos and P. Raghavan. Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications. In *Proceedings of the ACM SIGMOD'98 Conference*, pages 94–105, 1998.

[37] R. Baayen and H. Halteren and F. Tweedie. Outside the Cave of Shadows: Using Syntactic Annotation to Enhance Authorship Attribution. *Literary and Linguistic Computing*, 2:110–120, 1996.

[38] R. Forsyth and D. Holmes. Feature Finding for Text Classification. *Literary and Linguistic Computing*, 11(4):163–174, 1996.

[39] R. Lippmann. An Introduction to Computing with Neural Networks. *IEEE Acoustics Speech and Signal Processing Magazine*, 24:4–22, 1987.

[40] R. Rowlingson. A Ten Step Process for Forensic Readiness. *International Journal of Digital Evidence*, 2(3):1–28, Winter 2004.

[41] R. Zheng and Y. Qin and Z. Huang and H. Chen. Authorship analysis in cybercrime investigation. In Springer-Verlag, editor, *Proceedings of the 1st NSF/NIJ Symposium, ISI2003*, pages 59–73, 2003.

[42] S. Argamon and M. Saric and S. Stein. Style Mining of Electronic Messages for Multiple Authorship Discrimination: First Results. In *Proceedings of the 9th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.* ACM Press, 475-480 2003.

[43] S. Peter. A Comprehensive Approach to Digital Incident Investigation. *Elsevier Information Security Technical Report*, 2003.

[44] Schultz and Shumway. *Incident Response.* Sams, 1 edition, 2002.

[45] P. Stephenson. The DFRWS Framework Classes, 2003.

[46] T. Mendenhall. The characteristic curves of composition. *Science*, 11(11):237–249, 1887.

[47] Technical Working Group. *Electronic Crime Scene Investigation - a Guide for First Responders.* U.S. Department of Justice National Institute of Justice, http://www.ojp.usdoj.gov/nij, July 2001.

[48] W. Cohen. Learning Rules that Classify E-mail. In Springer-Verlag, editor, *Proceedings of AAAI Spring Symposium on Machine Learning in Information Access*, pages 18–25, 1996.