
Electronic Theses and Dissertations, 2004-2019

2013

The Hermeneutics Of The Hard Drive: Using Narratology, Natural Language Processing, And Knowledge Management To Improve The Effectiveness Of The Digital Forensic Process

Mark Pollitt

University of Central Florida, mark@digitalevidencepro.com



Part of the [Forensic Science and Technology Commons](#)

Find similar works at: <https://stars.library.ucf.edu/etd>

University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Pollitt, Mark, "The Hermeneutics Of The Hard Drive: Using Narratology, Natural Language Processing, And Knowledge Management To Improve The Effectiveness Of The Digital Forensic Process" (2013).

Electronic Theses and Dissertations, 2004-2019. 2994.

<https://stars.library.ucf.edu/etd/2994>



University of
Central
Florida

STARS
Showcase of Text, Archives, Research & Scholarship

THE HERMENEUTICS OF THE HARD DRIVE:
USING NARRATOLOGY, NATURAL LANGUAGE PROCESSING, AND KNOWLEDGE
MANAGEMENT TO IMPROVE THE EFFECTIVENESS OF THE DIGITAL FORENSIC
PROCESS

by

MARK POLLITT
B.S. Cornell University, 1973
M.S. Syracuse University, 2002

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Texts & Technology
in the Department of English
in the College of Arts and Humanities
at the University of Central Florida,
Orlando, Florida

Spring Term
2013

Major Professor: John D. Applen

ABSTRACT

In order to protect the safety of our citizens and to ensure a civil society, we ask our law enforcement, judiciary and intelligence agencies, under the rule of law, to seek probative information which can be acted upon for the common good. This information may be used in court to prosecute criminals or it can be used to conduct offensive or defensive operations to protect our national security. As the citizens of the world store more and more information in digital form, and as they live an ever-greater portion of their lives online, law enforcement, the judiciary and the Intelligence Community will continue to struggle with finding, extracting and understanding the data stored on computers. But this trend affords greater opportunity for law enforcement.

This dissertation describes how several disparate approaches: knowledge management, content analysis, narratology, and natural language processing, can be combined in an interdisciplinary way to positively impact the growing difficulty of developing useful, actionable intelligence from the ever-increasing corpus of digital evidence. After exploring how these techniques might apply to the digital forensic process, I will suggest two new theoretical constructs, the Hermeneutic Theory of Digital Forensics and the Narrative Theory of Digital Forensics, linking existing theories of forensic science, knowledge management, content analysis, narratology, and natural language processing together in order to identify and extract narratives from digital evidence. An experimental approach will be described and prototyped. The results of these experiments demonstrate the potential of natural language processing techniques to digital forensics.

This dissertation is dedicated to my loving wife, Jane. Her unconditional love has sustained me for over 35 years, and without whom this, and much else in my life, would not have been possible.

ACKNOWLEDGMENTS

This dissertation is the result of work conducted over more than two decades, during which I have been fortunate to have worked for, worked with, learned from, and shared ideas with literally thousands of people. All of them contributed something to this work. I am especially grateful to the men and women from the military, law enforcement and the intelligence community who provided a lifetime's tuition in identifying and solving problems.

The catalyst for this research was Dr. Robert Nielsen's course at the Information Resources College. Knowledge management transformed how I viewed not only digital forensics, law enforcement and intelligence, but information itself.

I would like to thank Drs. Elizabeth Liddy, Kevin Crowston, and Anne Diekma for sharing their knowledge of natural language processing. Drs. Nicole Beebe and Carole Chaski were very generous in sharing their digital forensic and linguistic research and experiences with me.

I am indebted to the many, often nameless, open source programmers, without whom the code for this dissertation would not have been possible.

This dissertation would not have been written were it not for the Texts and Technology program at the University of Central Florida. The program is uniquely dedicated to studying technology in truly interdisciplinary ways. That dedication was demonstrated over a six year period by each of my committee members. J.D. Applen, Melody Bowdon, and Paul Dombrowski. They have been excellent teachers, supporters and mentors in my quest. You were willing to take on a very non-traditional student in a very non-traditional subject. I am in your debt. I also want to pay tribute to my external committee member, Dr. Philip Craiger, who has stood shoulder-to-shoulder with me, teaching and researching digital forensics.

TABLE OF CONTENTS

LIST OF FIGURES	x
LIST OF TABLES	xii
LIST OF ACRONYMS AND ABBREVIATIONS	xiii
CHAPTER ONE: DIGITAL FORENSICS: THEORIES, CHALLENGES AND OPPORTUNITIES.....	1
Introduction.....	1
9-11 Changed Everything	4
What Do We Mean By <i>Text</i>	5
The Legal Notion of Evidence	6
Roles, Responsibilities, and Definitions	9
Theories of Traditional Forensic Science	12
Validity and Reliability.....	16
Ethics and Accountability.....	16
Communication.....	16
Accreditation.....	16
Theoretical Construct of Traditional Forensic Science	17
Statistics and Probability in Forensic Science	18

Digital Forensics and Its Theories	20
Theoretical Construct for Digital Forensics.....	29
The Practice of Digital Forensics.....	31
The Digital Forensic Process	31
Digital Evidence Challenges.....	37
Opportunities in Digital Forensic.....	43
CHAPTER TWO: FRAMING DIGITAL FORENSICS AS A TEXTUAL PROBLEM	46
A Knowledge Management Approach.....	49
The Law as a Narrative	54
Evidence as a Narrative	56
Extracting and Communicating Narratives.....	59
Structural Analysis and Semiotics	59
Narratology: Its Theory and Elements.....	62
Surrealism as a Tool.....	67
CHAPTER THREE: METHODOLOGIES FOR EXTRACTING NARRATIVES	72
Content Analysis.....	72
Extensible Markup Language (XML).....	76
Natural Language Processing	82
Levels of Natural Language Processing	85
Natural Language Processing in Digital Forensics.....	89

NLP Technologies	97
Applying Natural Language Processing	98
CHAPTER FOUR: DIGITAL FORENSIC THEORETICAL FRAMEWORK	105
Hermeneutic Theory of Digital Evidence	105
The Narrative Theory of Digital Forensics	109
Hypothesis One	110
Hypothesis Two	110
Hypothesis Three	111
Hypothesis Four	111
CHAPTER FIVE: APPLYING THE THEORY TO DIGITAL EVIDENCE	112
Experimental Design	112
The Enron Corpora	113
Description of Experiments	115
Email Header Extraction	115
Processing the Email Body	120
Description of the Software Developed	136
Analysis and Discussion of Results	139
Email Header Processing	139
Named Entity Extraction	140
Semantic Analysis	146

Summarization	148
Limitations	149
CHAPTER SIX: CONCLUSIONS AND FUTURE WORK	151
Conclusions.....	151
Hypothesis One.....	152
Hypothesis Two	154
Hypothesis Three	154
Hypothesis Four	155
Future Work.....	156
Hermeneutic Theory of Digital Forensics.....	156
The Narrative Theory of Digital Forensics.....	157
XML.....	158
Non-Forensic Applications for NLP Narrative Analysis.....	159
APPENDIX A—TEXT OF EXAMPLE EMAIL 443	161
APPENDIX B—FORENSIC XML CONTENT FILE EXAMPLE —EMAIL 443	163
APPENDIX C—DIGITAL FORENSIC NLP TOOLS PROGRAM	165
APPENDIX D—ELSEVIER COPYRIGHT RELEASE – ROUSSEV ARTICLE	171
APPEDNIX E—ELSEVIER COPYRIGHT RELEASE – WOOD ARTICLE.....	177
APPEDNIX F—SPRINGER COPYRIGHT RELEASE – VENTER ARTICLE	183
APPENDIX G—ACM COPYRIGHT RELEASE – FAN ARTICLE	186

REFERENCES 191

LIST OF FIGURES

Figure 1: Digital Forensic Process	32
Figure 2: Forensic Toolkit Overview Screen	35
Figure 3: Forensic Toolkit Search Screen.....	36
Figure 4: Areal density of HDD products vs. year of introduction. From Wood, R. "Future Hard Disk Drive Systems." Journal of Magnetism and Magnetic Materials Volume 321, Issue 6 2009 555 - 561.	37
Figure 5: Digital Forensics as Peeling the Onion	47
Figure 6: H. Porter Abbott's Construct of Narrative	65
Figure 7: Mieke Bal's Construct of Narrative	65
Figure 8: DEML Schema from Craiger, P. Digital Evidence Markup Language, 2009.....	77
Figure 9: DEML Data File from Craiger, P. Digital Evidence Markup Language, 2009	78
Figure 10: Garfinkel's XML Description of a File Object.....	79
Figure 11: Notional Email DTD	81
Figure 12: An Example of Text Mining. From Fan, et. al. Tapping the Power of Text Mining ..	91
Figure 13: CRISP-EM Level 2 From Venter, Waal & Willers.....	92
Figure 14: CRISP-EM Level 3 From Venter, Waal & Willers.....	92
Figure 15: WordNet Search Example	97
Figure 16: NLTK Tokenization and Tagging Example	100
Figure 17: Email Header and Body	117
Figure 18: Extracted Header in XML	118
Figure 19: Example of Parsed Sentence in XML	121
Figure 20 NLTK Parse Tree	123

Figure 21: Email 443 Nouns and Verbs.....	125
Figure 22: Email 443 Noun-Verb Sets.....	126
Figure 23: Email 443 Named Entities List	129
Figure 24: Email 443 Named Entities Set.....	129
Figure 25: Named Entity Extraction and Collection.....	130
Figure 26: Email 443 Named Entities and Verbs	131
Figure 27: Email 443 Extracted Nouns and Verbs	131
Figure 28: Noun and Verb Frequency Distribution	133
Figure 29: Email 443 Five Most Common Nouns and Verbs	133
Figure 30: Verbs Associated with the Common Nouns.....	134
Figure 31: Summary Snippets from Email 443	136
Figure 32: Digital Forensic NLP Program Flow.....	138
Figure 33: Named Entity Set from Lavorato's California Email Folder.....	141
Figure 34: Five and Ten Highest Frequency Words from Email Directories.....	148

LIST OF TABLES

Table 1:RCFL cumulative case load: FY 2003–2010. From Vassil Roussev, and Candice Quates. “Content Triage with Similarity Digests: The M57 Case Study.” Digital Investigation 9.Supplement S60–S68. Print.	38
Table 2: Subset of NLTK Tagset	99
Table 3: Comparison of Technical and Investigative Approaches	107
Table 4 Results of Named Entity Extraction from Ch. 17 and 2385 Emails	143
Table 5 Results of Ch. 17 Experiment, After Data Scrubbing.....	144
Table 6 Results of Named Entity Extraction from Ch. 11 and Emails	145
Table 7 Results after Data Scrubbing of Ch. 11 and Emails	146

LIST OF ACRONYMS AND ABBREVIATIONS

ASCII Text	American Standard Code for Information Exchange. This is a set of 128 codes used to represent text to a computer. It is considered the most basic way to represent text, since it does not have any formatting information other than spaces, carriage returns and line feeds.
Forensic Examiner	An individual whose primary responsibility is to conduct scientific examinations of evidence in legal matters.
FRE	Federal Rules of Evidence – rules established by the Federal judiciary dealing with the use of evidence in Federal Courts.
Investigator	An individual whose primary responsibility is to conduct civil, criminal, regulatory or administrative investigations in support of the legal system.
Metadata	Literally, information about information. In the digital forensic context, it typically consists of file system information such as the date and time of file creation (external metadata) or information embedded in the files, such as the file creator or version number (internal metadata.)
NIST	National Institute of Standards and Technology, as U.S. government agency charged with developing standards, including computing, computer security, and forensics.

OSI	Open Systems Interconnect. A model of how the different layers of a network work.
SWGDE	Scientific Working Group on Digital Evidence – a group of forensic science managers and practitioners who develop standards and guidelines for the examination of digital storage, communication, and processing devices and the data they contain.
SWGIT	Scientific Working Group on Imaging Technologies - a group of forensic science managers and practitioners who develop standards and guidelines for the examination and use of imaging technologies in forensic science. This includes the examination of audio, video, and visual images for evidentiary purposes.
Tuples	An ordered set of objects labeled by the number of elements in each set. The set (a, b) is a 2-tuple and (x, y, z) is a 3-tuple.
XML	Extensible Markup Language is a system of annotating and organizing data by utilizing “tags” that are both human and machine readable.

CHAPTER ONE: DIGITAL FORENSICS: THEORIES, CHALLENGES AND OPPORTUNITIES

Introduction

Computers and their associated digital technologies have transformed human society. They have altered the way we communicate, the way we learn, and our relationship with information. Digital technologies have mediated our interactions with each other and with our institutions (Ong). Marshall McLuhan's prediction of the medium becoming the message has, in the 21st century, come to pass. These changes have had great impact on the legal system. One of the most significant impacts has been how these digital technologies have become, in the legal sense, evidence. The means and methods by which digital information is acquired, examined, analyzed and presented, are called digital forensics. One definition of this field is "the application of science and engineering to the legal problem of digital evidence" (Slay, et al. 38).

If the history of ancient Egypt was written in hieroglyphics inscribed on the temples of the Nile and that of the Victorian Era infused into leather-bound volumes, our current era has, and is, being recorded in binary form and stored electronically. As in any age, there is a need to deal with personal and collective wrongs by means of a judicial system. Forensic science has evolved to assist the judicial system by finding and explaining "invisible," or as it is styled in legal terminology, latent evidence. With the rise of digital technology, there needed to be a means of providing forensics in this new environment. Thus, in the late 1980's, computer forensics began and has evolved into a broader field, involving not only computers, but all electronic devices that store or communicate information in binary form, a field called digital forensics (Casey 25-27; Pollitt, History 6-12).

This dissertation will propose that, adding a textual approach to the scientific and engineering approaches to digital evidence, will increase the value of digital forensics to the legal, law enforcement, intelligence, and information security communities. A theory, integrating the concepts of knowledge management, content analysis, narratology, and natural language processing, will be developed and applied to digital evidence. I will then demonstrate how this could be applied to a real world example of emails from the Enron Corpus.

This dissertation includes material from a number of disparate fields, some of which the reader may not be familiar. As a result, it is necessary to set a very wide and deep stage before getting into the details of the theory and its application. I ask the reader's indulgence for the length of the first two chapters, as it is necessary to provide a solid foundation for the remainder of the dissertation. The first chapter is divided into an Introduction and four sections. The Introduction will describe the nature of *text*, in a texts and technology context, and the legal concept of evidence. The second section will define traditional forensic science, as well as, its current theories and methodologies. The third section will focus on the theory, practice and process of digital forensics. The last two sections will explore the challenges and opportunities in the field of digital forensics.

The history of digital forensics is relatively brief. Initially called "computer forensics," the first law enforcement units created to look for evidence that was stored or transmitted by computers were formed in the late 1980's. Initially, most of the cases that were examined for digital evidence were relatively high tech crimes. In those days, computers were very expensive and required a lot of technical expertise to utilize. So, not surprisingly, the people who used computers in crime, as well as those who would investigate them, had to be technically oriented.

By the mid-1990's, computers became cheaper, easier to use, and were connected to the World Wide Web, via the Internet. This resulted in increased access to computers as tools of criminal behavior and concomitant increase in targets, human and digital, for criminals. As a result, there was a rapidly increasing need for law enforcement to collect and analyze digital information from criminal cases (Pollitt, History 6-12).

In the days leading up to September 11, 2001, the largest investigative use of digital forensics was the investigation of online child pornography. It was this focus which drove the state-of-the-art in digital forensics. Child pornographers, in the pre-digital days, had to take photographs, develop the pictures in clandestine photo labs, and then carefully exchange them with other criminals. Each of these steps was fraught with the potential to be caught by the law. Digital technology provided less risky methods at each step of the process, and significantly more anonymity. Digital photography and the Internet transformed child pornography. As a result, the online sexual exploitation of children became *the* high tech crime and fostered the creation of a large number of investigative programs. The ability to process the increasing volume of digital evidence in lagged, and continues to lag, the investigative needs (U.S. DoJ IG 5, 10-12).

In these types of cases, the evidence of most interest to investigators were the photographs that formed the *prima facie* case of child pornography. If the investigator could show a knowing and willful possession of prohibited images, the suspects would usually confess, and rather than go to trial, offer a guilty plea. But as the size of hard drives rapidly increased, the numbers of photographs that could be and were extracted from subject's computers increased dramatically. The visual review of every photograph found on a suspect's hard drive had gone from disagreeable tedium to impossible.

The immediate solution was to calculate a mathematically unique number, called a cryptographic hash, for each file and compare that value to a database of known, provable child pornography. If a sufficient number of known, provable, child pornographic images were located, the case was made without having to review thousands of pictures. While this provided a solution to finding some of the known pornographic images, it could not identify previously unknown files, or identify any unknown victims.

Because the evidentiary needs of the investigators were technically modest, straightforward data recovery from well-known file systems was usually sufficient. Digital forensics practitioners focused on the technical aspects of the systems that produced the data, the devices that stored the information, and their ability to view and extract this information. This might be described as a “computer science” view of the forensic problem. The focus is primarily on the operation of the computer, and not on the content.

9-11 Changed Everything

The morning of September 11, 2001, changed many things. Within a few hours, the FBI turned its primary focus from the investigation of crimes to the prevention of terrorism. Since information is the raw material for investigative and intelligence work, one of the first orders of business was to collect as much information, about a wide range of individuals, organizations and activities, as quickly as possible. For the first time in the FBI’s history, most of that information would be in digital form.

In the days just after September 11, 2001, I was the Chief of the FBI’s digital forensic program, and was scheduled to brief the Attorney General of the United States, John Ashcroft, the Director of the FBI, and his executive staff concerning the forensic examination of a particular computer. The initial examination had been conducted by a well-qualified digital

forensic examiner in the field. In preparation for my briefing, I had a copy of the computer's hard drive sent to me and I re-examined the computer to verify the examiner's results and to be able to report, with personal knowledge, what was found.

At the actual briefing, I articulated virtually everything found on the computer which might have had any possible significance to the case. When it came time for questions, the Attorney General looked at me and asked, in words to this effect, "What does it mean?" There was a long, and painfully awkward, silence, as I realized that I had no idea. My briefing had answered the wrong questions. For me, it was an epiphany – my field, digital forensics, was often answering the wrong questions, not because we didn't want to answer the "right" questions, but we were going about it in the wrong way. We had been confusing the ability to collect data with knowledge. We provided data, not meaning.

What the Attorney General wanted was not a recitation of the data found on the computer, but knowledge, or at least information, which would tell him if there was another plot hatching. Was this individual targeting some particular location? Were there other co-conspirators? He desired knowledge that would tell him what to do. The military and the Intelligence Community have a term for this missing link: "actionable intelligence." In more simplistic terms, he wanted a "story" of who was going to do what, when, why and how. To him, this "story" would have meaning. I began to think about the relationship between digital evidence, knowledge and story. I have been thinking about this issue for over a decade since that briefing.

What Do We Mean By *Text*

There is one more issue that needs to be addressed upfront. That is the use of the term *text* in the context of this dissertation. The study of "texts and technologies" is "an interdisciplinary approach to the intersections of texts and technology from cultural, historical,

and communicative perspectives” (T & T Brochure). As such, it views the notion of *text* in a very broad way; not only with the traditional format of written and printed media, but expands the view to include “emergent forms of video, audio, and multimedia texts” (T & T Brochure). This perspective is particularly appropriate to digital forensics, as the meaning developed from digitally stored data go beyond merely extracting documents and other, character-based, information. The items examined as digital evidence, such as, hardware, storage devices, data bases, and email, are in many respects, the ultimate form of emergent multimedia. Digital evidence may yield data in the form of graphics, audio, video and even computer programming code. In this dissertation, I will try to make it clear when I am specifically talking about character-based or alphanumeric data. In other cases, the reader is encouraged to interpret the term *text* in the broadest sense.

The Legal Notion of Evidence

As this entire dissertation revolves around the use of texts in legal proceedings, this is a good time to define the legal notion of evidence. Evidence is used in court to lay out a sequence of events which form the physical acts (*actus reus*) and any required mental state (*mens reus*) of the alleged crime or tort. Crimes are those acts which are defined, as a matter of law, as “behavior that the law makes punishable as a public offense,” whereas torts are “a civil wrong which can be addressed by awarding damages” (Cornell).

People commonly use the word *evidence* in a variety of ways. However, in a legal context, it has some very specific meanings. In court, something is only *evidence* if the judge allows, or “admits” in legal jargon, it into the court proceedings. So, the first requirement of evidence is that it must be admissible. The *Federal Rules of Evidence* define admissible evidence as evidence that can be used in court, and “has any tendency to make a fact more or less probable

than it would be without the evidence and the fact is of consequence in determining the [legal] action” (“Federal Rules” Rule 401).

We have described *evidence* as something admitted by the court and explained that these things can be in a variety of formats. The courts recognize sub-classes of evidence, some of which will be described below. Unfortunately, even within the law enforcement and legal community, it is common to refer to anything which supports a theory or position as *evidence*. Similarly, in the forensic science community, the technical term for an object that is to be examined is *submission* or *exhibit*. Unfortunately, these are also often colloquially referred to as “evidence.” In this dissertation, I will attempt to make clear when I am specifically referring to the notion of admissible evidence, its various sub-classes, a submission, or the more general notion of evidence.

For hundreds of years, western judicial systems have relied upon evidence in either oral or physical form. Oral evidence, the testimony of a witness in a legal proceeding, is perhaps the oldest form of evidence. It relies upon the “trier of fact,” meaning a judge or jury, to evaluate the testimony with respect to the particular legal proceeding. It is the job of the judge or jury to listen to the content of the witness’s testimony as well as measure their credibility. Lawyers will ask the witnesses questions to try and elucidate their particular position, but it is the witness that provides the information. Determining how much “weight” to give a particular testimony is another of the trier of fact’s duties.

The use of physical objects as evidence is common in western judicial trials. (Note: the author has little experience in judicial proceedings outside of Europe and the Americas, and thus has limited this discussion to western legal systems.) Most readers are familiar, from the televised proceedings of court cases, with the admission of a gun or a bloody glove into evidence

at trial. Computers, cell phones, and other electronic devices can also be used as purely physical evidence. That is, where they are examined as physical objects, independent from their electronic capabilities. As physical evidence, their presence at a crime scene may demonstrate some element of the alleged crime. For example, if the alleged crime is counterfeiting, then the presence of a high-end scanner, computer and printer would demonstrate an ability to commit this crime. These physical objects may also contain fingerprints or DNA which may serve to establish the presence of a particular party.

But in most cases, the principal use of computers and other electronic devices will revolve around their capabilities to store and communicate information. The information stored on these devices is considered “documentary evidence,” which is a legally particular form of physical evidence. There is a well-developed case law concerning documentary evidence

In a sense, a document stored on a computer is like a letter or memo, only in electronic form. But, it differs in one significant way. A letter is a physical object that can be seen by investigators, judges and jurors. The law considers a physical item, which can be readily seen, as “patent” evidence. Since we cannot look at a computer, or even its hard drive, and read the documents which it contains with our bare eyes, the law considers this “latent” evidence in the same sense that fingerprints or DNA are. While digital evidence may be latent, it is nonetheless a form of physical or real evidence. Similar to other latent forms of evidence, the law recognizes that there needs to be some technical process which makes these “invisible” forms of evidence visible. In addition to making the evidence visible, the law requires that these forms of evidence be demonstrably “authentic” and that the significance and interpretation of these forms of evidence be the result of a reliable process. This is one of the principle roles of forensic science:

to support the legal system in identifying, collecting, preserving, examining, and reporting the results of the examination of latent evidence to the court.

As part of that reporting process, charts, diagrams, simulations, or other instructional aids are utilized. When used as such, the legal system calls these “demonstrative evidence” (Wex). While this kind of evidence must be admitted by the court, it does not, necessarily, have to be factual in and of itself. However, it is common to take information from physical or documentary evidence and diagram or summarize it in a chart or diagram (Deehl 1-2). For example, in a “check kiting” case, where the accused is alleged to knowingly write checks with insufficient funds, the fraud investigator might create a chart listing all of the checks the dates they were submitted, the account balances and the bank’s written notices to the defendant. This chart would be demonstrative evidence, which also contained elements of documentary evidence, in the form of the information from checks and the bank’s notifications to the customer. In order for the summary chart to be admissible, the checks and notifications must have been previously admitted (Deehl, 2).

Physical and demonstrative evidence are admissible (FRE 101). But in order to be admitted, a “foundation” must be laid. Normally, this is done by having a witness describe the item, where it came from, and what it purports to be. The court then evaluates the evidence for “authenticity,” meaning that it is what it purports to be, and “relevance,” meaning that it tends to show a fact is more or less likely (FRE 401, 901).

Roles, Responsibilities, and Definitions

Throughout this dissertation, I will utilize some terms to represent the individuals who perform various roles in the legal process. It is important to understand that these roles are not

immutable, and in many cases, a single individual will perform several of the functions. I will briefly describe these roles.

An investigator is any individual charged with conducting an investigation, collecting evidence, and/or making recommendations to others concerning a civil or criminal case. These individuals may be civilians or sworn law enforcement personnel, information security professionals, or auditors. The matter investigated, the “case,” may be criminal in nature, civil, regulatory, or administrative in nature. The point is that the person performing this function focuses on the alleged behavior, collecting information and objects (evidence), and tries to synthesize the results for the use of others. In this dissertation, I will focus on investigators who perform these responsibilities with respect to digital evidence and I will use the term, investigator to include law enforcement, private sector investigators and cybersecurity personnel conducting investigations of criminal, civil, administrative, or regulatory investigations.

While the investigator focuses on the case, the forensic scientist focuses on the evidence. Forensic science is often defined as “the application of scientific, technical, or other specialized knowledge to assist courts in resolving questions of fact in civil and criminal trials” (Forensic Sciences Foundation). There are a number of different practitioners within the forensic science domain. Unlike what is seen on many television shows or movies, “Crime Scene Investigators” do not usually carry guns, conduct interviews or make arrests. In fact, most are not sworn (badge and gun carrying) members of law enforcement agencies. Their job is to go to crime scenes and collect evidence – nothing more. The evidence they collect is then submitted to a laboratory and/or Medical Examiner’s Office to be examined by a forensic scientist trained in the specifics of a type of evidence. The job titles for these forensic scientists vary widely, and occasionally forensic scientists will conduct crime scene investigations, but it is not the norm.

For purposes of this dissertation, I will use the terms forensic scientist, examiner and forensic practitioner synonymously, as someone who conducts forensic examinations of evidence. And while these people may collect digital evidence, they are, by training, experience and function, distinct from those who merely collect a variety of evidence at a crime scene.

There is one additional role, that of the analyst. The role of the analyst is to take the data and information, as documented by the investigator and examiner, and to add additional meaning to it by organizing and contextualizing it. Analysis, and therefore analysts serve a critical role in the investigative and intelligence process. One cinematic example of a professional analyst is Tom Clancy's character, Jack Ryan, especially in the early novels, such as *The Hunt for Red October*. While the story line has him becoming involved in the actual operations, his role in the initial part of the story is to analyze Soviet naval activities. His background is as a naval historian, which he uses to contextualize the data on the mystery submarine and its Captain (Clancy). The "Jack Ryan" novels gain narrative tension from Ryan being forced to abandon his comfortable, intellectual life, to become tactically involved in physical conflict. While this is good entertainment, it is not often done in the "real world," nor would it be generally effective. In my experience, the psychological makeup, work style, skills, and intellectual nature are very different between analysts and investigators.

The use of specifically trained analysts is common in the Intelligence Community and in major criminal investigations. However, in most cases, the analysis is done by the investigator and/or the forensic examiner. This situation is far from ideal, as the goal-oriented behavior of investigators is to prosecute as many cases as possible whereas the appropriate goal of the forensic examiner is to only make accurate and reliable scientific conclusions, without consideration of the outcomes.

Theories of Traditional Forensic Science

In order to understand the theoretical basis of digital forensics, we must first explore its parent: traditional forensics. By “traditional forensic science” I refer to the scientific disciplines that constituted forensic science up until the late 1980’s. These include specialties such as forensic pathology, forensic chemistry, toxicology, firearms/toolmark examination, questioned document examination, and serology. Once we understand the principles utilized in these traditional forensic disciplines, we can then explore their similarities and differences to digital forensics.

While there are a number of definitions for forensic science, most are generally similar, and for this discussion, I will use Saferstein’s definition: “Forensic science is the application of science to the criminal and civil law that are enforced by police agencies in a criminal justice system”(4).

Saferstein’s book is entitled *Criminalistics*, and so, while the reference to police agencies and the criminal justice system are appropriate to that context, the first part of the definition correctly asserts that forensic science is utilized in both civil and criminal law. Indeed, if we merely look at the Table of Contents for Wecht and Rago’s *Forensic Science and the Law*, we see numerous chapters devoted to civil law issues. For those not familiar with the distinction, criminal law is the prohibition of actions, by the government, which are punishable by fine or incarceration. Civil actions are actions of private or governmental parties which seek financial recompense or injunctive relief for legal wrongs, known as torts committed by other parties (Wecht and Rago 139-40,165). Murder is a crime, as it is prohibited by statute, prosecuted by the government and may result in a prison sentence. The same act, killing another human being, deprives the victim’s family of that person’s support and income and may, therefore, also be a

civil tort. Perhaps the best known example is the O.J. Simpson case. Simpson was tried and acquitted of murder, but found liable for wrongful death and battery in the killing of Nicole Brown Simpson and Ronald Goldman (PBS “Oppression and Malice: The O.J. Simpson Civil Trial”).

Inman and Rudin take a more nuanced view of forensic science. They identify three elements of forensic science: discussion or debate in the context of a court, the use of the scientific method, and forensic science as an “applied science” (4-7, 15). But they also recognize a larger context for this field. They integrate Edward O. Wilson’s view that science is “the organized systematic enterprise that gathers knowledge about the world and condenses the knowledge into testable laws and principles” into their construct of forensic science. They also recognize that there are elements of art within forensic science (9). While forensic science strives to maintain scientific objectivity, it must be realized that even comparison, between a piece of evidence collected at the crime scene and a known reference sample, is not free of human subjectivity (12).

Inman and Rudin also make an interesting distinction between forensic sciences that are focused on specific types of evidence, such as fingerprints, firearms and documents, which they call criminalistics, and the applied sciences, such as forensic pathology and forensic toxicology, which apply an external discipline (e.g., pathology), to a forensic purpose. It is an academic versus craft dichotomy, that while alternately leaning from one side or the other, has evolved both groups towards a central discourse that recognizes that there is both art and science in all forensic disciplines (10-17, 40-43).

Perhaps Inman and Rudin’s greatest contribution is their attempt to define a “unifying paradigm of forensic science.” In doing so, they invoke Kuhn’s explication of how the study of

paradigms defines a scientific discipline (75). In Inman and Rudin's paradigm, they include two principles, and five processes, although two of the processes are an extension of one to the other.

Two principles, divisibility of matter and transference, characterize the nature of physical evidence. Physical objects can be mechanically or chemically separated and by this division, produce evidence. A very simple example of this would be a criminal breaking a window; there are particles of the glass found still attached to the window frame, pieces of glass at the crime scene and perhaps shards of glass caught in the criminal's clothing. It is the same material, only divided. The latter part of the example demonstrates the second principle, transference. This is the ability of material, once divided, to adhere to something that comes into contact with the divided material (76).

The processes, described by Inman and Rudin, are: identification, classification/individualization, association, and reconstruction (76-80). Because, as we will shortly see, individualization is an extension of classification, we will refer to these as Inman and Rudin's four forensic questions.

Identification is simply the scientific ability to define the nature of an object. Legally, it may be sufficient to chemically identify a controlled substance, e.g. cocaine. The possession of cocaine is prohibited; therefore, once benzoyl-methyl-ecgonine (the chemical name for cocaine) is identified in the sample, no further qualitative examination may be necessary (78).

Classification is the ability to define objects as coming from a common origin. A bullet recovered from a homicide victim, because of its dimensions, weight, shape and markings may be identified as coming from a Smith & Wesson 9mm pistol. Because there are millions of such pistols, this is considered "class evidence." The gun, which fired the bullet, belongs to the "class" of Smith & Wesson 9mm pistols. If we locate the actual gun used to fire this bullet, we can

determine, through the microscopic markings on the bore of the barrel, that, the recovered gun is the source of the bullet recovered from the victim, to the exclusion of all other pistols, and we have “individualized” the evidence. Because of the relationship between these two processes, we refer to them as a single forensic question. All individualizations are not binary – there may be a range of certainty. In their book, Inman and Rudin give us a very instructive approach to dealing with the logical problem of levels of certainty given the inability to prove a hypothesis (113-154).

The last two principles described by Inman and Rudin are association and reconstruction. The former is the ability to “infer contact” between two pieces of evidence, while the latter is the ability to order “events in the relative space and time based on the physical evidence” (178). A fingerprint may identify the perpetrator, but having been found at the scene of the crime, associates the perpetrator with the crime scene. The entry and exit wounds on the victim’s body, compared with the location of the spent bullets may allow forensic scientists to reconstruct both the order of the shots and the position of the body at each impact. These two principles differ from the first two in a very significant way: they put the physical evidence in an investigative context (165-178).

In addition to the scientific and forensic principles described, most forensic science texts define structured approaches, either implicitly or explicitly, that form what might be described as a forensic process. Saferstein includes, in many of his chapters, sections on “collection and preservation” and “analysis.” Inman and Rudin provide, in their Appendix F, the following processes: evidence collection, prevention of contamination, examination, interpretation and conclusions (356-57).

Because forensic sciences are “applied” sciences, a common set of practices have evolved from the discipline’s discourse. In this section, I will try to enumerate some of these.

Validity and Reliability

One of the tenets of forensic science is the requirement to ensure that the practitioner's work is valid and that the process is reliable (Inman & Rudin 221; NRC 8-13; Tilstone 62). This is in part a legal requirement derived from both common law and recent court decisions, most notably *Daubert v. Merrill Dow Pharmaceuticals* (Blackmun).

Ethics and Accountability

Similarly, ethics and the related notion of accountability are considered central to the practice of forensic science. Most professional forensic science organizations have a Code of Conduct or similar canon of ethical behavior (Inman & Rudin 299-321; Append. B-D). In addition, since forensic practitioners often work for lawyers, they are bound by some of the Rules of Professional Conduct established for the legal profession (Wecht 663-74).

Communication

The ultimate results of forensic examinations are usually transmitted to the interested parties in the form of oral or written communications. The written communications are usually reports submitted to the government, attorneys and courts. In addition to ordinary, informal oral communications, the ultimate test of the forensic process is in courtroom testimony (Inman 293-98). The forensic practitioner usually testifies as an expert witness and, as such, is allowed to “teach” the judge and/or jury about the underlying scientific foundations of their testimony and how they applied their knowledge skills and abilities to the instance at hand (McKasson and Richards).

Accreditation

Tilstone, et al., in their encyclopedia, describe the related concepts of accreditation and certification. While the latter is the process of assuring an individual has the appropriate

education, training and experience to perform at a professional level, the former is the external evaluation of a laboratory's "policies, practices and procedures for compliance against a credible set of standards" (70-71). Anja Einsele, in her chapter, Forensic Laboratory Accreditation, highlights the need to imbed quality into the individual and "corporate" culture, as well as, the notion of continuous improvement (Mozayani and Noziglia 1-12). The National Research Council report concluded that forensic laboratory accreditation is an important means of ensuring reliable scientific outcomes (215).

Theoretical Construct of Traditional Forensic Science

From the preceding, we may distill the following:

1. Forensic science utilizes science to make deductions and inferences on things (evidence) of interest to the legal system.
2. Forensic science utilizes validated processes that preserve the integrity of the evidence and seek to answer questions of identification, classification/individualization, association and reconstruction within the context of the legal issue and about the evidence. In most forensic examinations, evidence is submitted with specific legal or investigative questions either explicitly stated or implied.
3. Forensic science seeks to act in an ethical manner and in as neutral a fashion as possible. Quality is actively managed and subject to both accreditation and judicial review.
4. Forensic practitioners communicate the results, conclusions and opinions obtained from their examination in the form of written reports and oral testimony.

Statistics and Probability in Forensic Science

The use of statistics, particularly Bayesian reasoning and logic, has “revolutionized” forensic science over the past 25 years. This impact has been central to the use of DNA analysis and has been extended to include fingerprints and other “trace evidence” disciplines, such as hairs, fibers and glass (Curran 141). The use of statistical interpretation of DNA profiles has become so accepted that there is little debate and few legal challenges to its use. For example, President Clinton did not challenge the results of the tests conducted on Monica Lewinsky’s dress (Weedn 428). This focus on mathematical results has, in many ways framed the public’s perception of forensic science. As the National Research Council stated in their 2009 report:

Although the forensic use of nuclear DNA is barely 20 years old, DNA typing is now universally recognized as the standard against which many other forensic individualization techniques are judged. DNA enjoys this preeminent position because of its reliability and the fact that, absent fraud or an error in labeling or handling, the probabilities of a false positive are quantifiable and often miniscule. (NRC 130)

But while the utility and accuracy of Bayesian statistics in the DNA context is undisputed, its use is not applicable to all forensic sciences, and has so far, demonstrated little application to digital forensics. As Anderson, Schum, and Twining state in their book, *Analysis of Evidence*, “such numerical judgments are quite difficult to make and justify because the events of concern either happened or did not happen on exactly one occasion.” They go on to describe five issues with regards to the probabilistic nature of evidence:

1. Evidence is always incomplete.
2. Evidence is commonly inconclusive.

3. Evidence is often ambiguous.
4. Evidence is often dissonant.
5. Evidence comes from sources that have varying levels of credibility. (246)

They review four statistical models, including Bayes, which address the probative value and/or weight of the evidence. Among these, the most germane to the examination of digital evidence is likely Cohen's Baconian approach. Core to this approach are the dual notions of hypothesis testing and that testing either supports or refutes a given hypothesis. As Anderson, et al, describe it: "Cohen's system of Baconian probability is the only system that takes specific account of how completely the evidence we have covers matters recognized to be relevant". It measures the weight of the evidence by comparing what questions the evidence has answered and against the questions that are unanswered (259).

Cohen describes the perceived differences between "mathematical" and "inductive" probability as "expert, exact, numerical way" versus "a popular, loose, qualitative way" (L. Cohen 39). He argues that inductive probability is not less valuable, but is inherently different, involving "a comparative or ordinal gradation of probability rather than a quantitative and measurable one" (40-1).

It is important to understand that statistics and the notion of probability have a broader, and more functional, role in both jurisprudence and forensic science. We can use probability to help us find probative information without the ability to demonstrate a finite, causal relationship. This will become clearer in a later section, where we describe the examination and analytical processes.

Digital Forensics and Its Theories

As stated earlier, digital forensics can be defined as “the application of science and engineering to the legal problem of digital evidence” (Slay, et al. 38) In turn, digital evidence is defined as “information of probative value stored or transmitted in” binary form (SWGDE). Carrier describes it as a “process where we develop and test hypotheses that answer questions about digital events” (4). In its simplest form, this means looking at computers, network devices and data storage devices, such as hard drives, for information significant to an investigation or a court case. Consider the following examples:

A young intern goes missing. In an attempt to find out what might have happened, police search her computer, where they find recent web searches for Klinge Mansion, located in nearby Rock Creek Park. Unfortunately, that is where they would ultimately find her body (“Chandra Levy”).

A 26-year-old man communicates with a 14-year-old girl half way across the country. Using a video chat; he pressures her into emailing him nude photos of herself (Plumlee).

Computer evidence was at the heart of the contested ownership of half of Marc Zuckerberg’s share of Facebook. The evidence in this dispute was located on personal computers, Facebook corporate servers, and Harvard’s email servers (Van Voris).

In each of these cases, the digital evidence is crucial, not only prove the case, but to tell the narrative or story of the case.

It is necessary to also understand the nature of evidence in a judicial setting. As we saw in the previous section, evidence is used in court to lay out a sequence of events which form the physical acts (*actus reus*) and any required mental state (*mens reus*) of the alleged crime or tort.

Crimes and torts are a sequence of events. They unfold in a largely sequential way and are often presented in court as a narrative. Effectively, criminal investigations and prosecution are the identification and communication of facts in a legally specific narrative. Similarly, forensic scientists seek to describe the “story” told by the evidence in the form of a written report and oral testimony, both often in a narrative format. In a very real sense, the justice system is founded on the notion of narratives, as will be explored in a later section.

In the 21st Century, we utilize our computers for everything from audio/visual/textual communications and data storage, to entertainment. Because they are such multi-purpose machines, everything on a criminal’s computer is unlikely to be associated with the crime being investigated. In most cases the subject of our investigation will have a wide range of information stored on their computer that is not pertinent to our investigation. Therefore, digital forensics needs to find (discover), in digital form, that which is pertinent to the particular inquiry and organize it in such a way that it reconstructs the narrative particular to the crime being investigated.

Depending on the crime or tort, the probative information may be the content of computer files or information created as a byproduct of the operation of the computer or networks to which the computer was attached. This latter information is “information about information” and is called *metadata*. Examples of metadata are things like: the file name, file creation/modification dates, and the application’s user. In effect, we have two classes of digital evidence: content and metadata. Metadata can be further subdivided into “internal” metadata, meaning metadata stored in the data file, and “external” metadata, which is associated with the file system. For example, Microsoft Office stores information within Word files, such as the identity of the user and how long the file has been edited. This is internal metadata. The file’s

name, creation date and location, produced by the file system and recorded in the file's directory listing, would be classified as external metadata.

Digital forensics is a process used to answer legal and investigative questions. Kruse and Heiser describe it thusly:

Computer forensics involves the preservation, identification, extraction, documentation, and interpretation of computer data. It is often more of an art than a science, but as in any discipline, computer forensic specialists follow clear, well-defined methodologies and procedures.... (2)

They further describe their basic methodology as "the three As:"

1. Acquire the evidence without altering or damaging the original.
2. Authenticate that your recovered evidence is the same as the originally seized data.
3. Analyze the data without modifying it. (3)

Carrier and others articulate slightly different processes, but, as a discipline, the use of a systematic approach to conducting forensic examinations/investigations is well accepted.

Digital evidence is somewhat unique in that its relationship to a legal or investigative case is multiplicitous. As Parker describes in his books, computers can serve a number of different roles in connection with crime (Crime; Fighting). He defines these roles: *object*, *subject*, *tool* and *symbol* (Casey 32). From my investigative background, I have found it more useful to utilize a different nomenclature. Computers can function as *weapons* for hacking or the theft of intellectual property. When used for counterfeiting, committing scams and soliciting sex crimes, they are *instrumentalities*. When used to keep records, such as in white collar crime or drug trafficking, they are *tools*. And last but not least, computers can be *victims* of malware, such as viruses or of hostile intrusions. Computers can even be combinations of these, as in the situation

where, as a result of being infectedd with malicious software, the computer becomes a “bot” used to attack websites. As such, it is first a victim and after activation as a “bot,” it becomes an instrumentality. If the hacker were to store malware programming materials on the infected computer, it would also be classed as a tool. These roles greatly increase the digital forensic practitioner’s problem of defining the scientific questions to be asked, and the tools and methods used to conduct the inquiry.

The traditional forensic practitioner offers a limited set of examinations that answer a relatively small set of questions. For example, a firearms examiner, given a fired bullet or casing, can only provide information concerning the physical characteristics of the bullet and/or casing which may answer what cartridge (e.g., 9mm or .38 Special) or firearm (e.g., Smith & Wesson revolver or Colt pistol) was used (Hueske 238-9). The digital forensics examiner’s task is far more complex. She must spend a great deal of time custom designing a process and the criteria which will recover, identify and extract only the most pertinent information for each individual case.

Digital forensics shares most of the framework and discourse of the traditional forensic sciences, such as fingerprints, forensic toxicology and forensic pathology. In these traditional forms of forensic science, the usual process, unlike what is seen on CSI and other entertainment venues, involves two groups of people: those who collect evidence at the crime scene (the “real” Crime Scene Investigators) and the forensic scientist, sometimes called an examiner. The former group goes to the scene of the crime to locate, document and collect physical objects (evidence), albeit sometimes the evidence is microscopic. While much of the work is done visually, often specialized cameras, lights, vacuums and lasers are used to collect the evidence. Crime Scene Investigators are trained to know what to collect and how. When done well, there is a feedback

loop between the case investigators, the forensic scientists and the CSIs (Saferstein 13-14). When done correctly, the evidence collected is properly preserved and identified, ready for examination by the forensic scientist.

In the laboratory, the traditional forensic scientist utilizes a combination of knowledge and tools to attempt to answer one or more of Inman and Rudin's four forensic questions: identification, classification/individualization, association and reconstruction (75-80). While knowledge and skill are required to answer novel questions, examine novel specimens or interpret results, most examinations are routine and many are largely automated.

Computers, and by extension the digital evidence they contain, for a given investigation, may have various roles, including: weapon, instrumentality, tool, record, or victim. The evidence itself is polysemous. The data obtained from computers and their storage, in its raw, binary form, has little investigative value until it is placed in some context. It may be an external context, such as content that describes persons, places, things or activities or an internal context, such as its location within the storage device or the software used to create the data. These two distinctive characteristics, the case context and the functional context, are further complicated by the ubiquitous networking of computers. A computer can store or access data anywhere on the network, therefore finding and preserving the digital evidence may be problematic. Compared to the relatively straight-forward problem of collecting evidence at a physical crime scene, the collection and preservation of digital evidence is much more difficult. Traditional CSI's know what to look for and have a fixed, finite crime scene. This might be viewed as a two-dimensional problem. In contrast, digital forensic personnel have a multi-dimensional problem of location, role of the computer, relationship of the computer/storage to the crime, the abstraction of the data and the investigative/knowledge management context of the evidence. Digital forensic

practitioners have software and hardware tools to conduct examinations, however, these tools, as I will demonstrate later, do not answer the “right” questions. Questions like Ashcroft's: "What does this mean?"

If you are investigating a computer hacking case, the evidence computer may be a *victim* or a *weapon*, and we would likely want to look for pertinent information in the context of how the computer operates. We would scan the computer's operating system for rogue, or altered system files, analyze log files, and try to reconstruct a timeline of events. If we were investigating a child pornography case, one of the key sets of evidence would be photographs, and we would be looking for data files stored on the hard drive. Since the computer is acting as a data repository, we would scour the computer's hard drive for active and deleted image files. The investigation of a suspected terrorist is a bit trickier, as we need to find out how he is using the computer (as a weapon, a record, a tool or some combination of these), what information he is storing, with whom he is communicating, and ultimately, what his motivations and intentions were. We would look at the user's network and Internet history, emails, chats, data files and suspicious application files, like viruses and hacking tools. Most of the questions that investigators or prosecutors need answered in this last kind of investigation are not Inman and Rudin's sort of questions, but the traditional investigative framework of: who, what, when, where, why and how. In short, we need to look for different things, in different places, for different purposes (Beebe, Terabyte; Research).

The spectacular increase in volume of digital evidence is, in and of itself, a major problem. In a recent capital homicide case, one of the court appointed experts was quoted in the press: "The old idea of linear review — of a defense attorney laying his eye on every page — is not a feasible idea in the digital age" (Barry).

Knowing where to look is only part of the problem. Knowing that you have found something of investigative or legal value is often difficult or impossible. Utilizing the traditional hypothesis-based scientific approach, you will only select those files that either support or refute your hypothesis. This presumes that you know the actors, actions and sequence of events. If the hypothesis is too narrow, then important evidence will not be selected, analyzed or utilized. If the hypothesis is too broad, the volume of evidence quickly becomes overwhelming and consequently of little value.

An alternative is to take the “artistic” approach, where you evaluate each item to see if it might reasonably be of value. Given that computers have storage capacities larger than most libraries, this is problematic (Lesk). While Inman and Rudin’s forensic questions are very useful in answering questions about physical evidence (including digital evidence), they are, in many ways, too granular to contextualize much of the data in digital evidence. The investigator and prosecutor have a different set of questions. They tend to use the: *who, what, when, where, why* and *how* paradigm described above. And while those questions are useful for investigators and prosecutors, it is difficult to directly address these using currently available digital forensic tools or techniques. As a result, we must find a way to bridge the gap between the two paradigms.

Most digital forensic textbooks focus on the tools to use and what various artifacts might tell us. Further, most textbooks provide only generalized approaches to the actual examination/analysis of digital evidence (Carrier; Carvey; Casey; Kruse and Heiser). While the digital forensic practitioner needs this information, it does not help her to decide what information to collect, analyze or report, and this is another area where narrative can be of service. The examiner is not only collecting narratives, but is creating, by selecting probative evidence, drafting a report, and providing oral testimony, a meta-narrative of the case.

While Farmer and Venema's book, *Forensic Discovery*, is likewise focused on tools and artifacts, they make several very profound observations. They analogize the deleted data on a hard drive as "fossilized." (12) They then extrapolate this to geology and archeology. They analogize the creation of data, by the operation of the computer, to the physical forces of nature, such as plate tectonics and volcanos. This they call this "digital geology." What defines this aspect is that the artifacts are caused by the inherent operation of the computer and are not products of user action. In contrast, they use the term "digital archeology" to describe the artifacts of human intervention (12-13). These metaphorical approaches are helpful in differentiating those examinations in which the activity is focused on a computer's activities from those involving user data. The use of the term archeology also implies the authorial nature of the data. The users are "writing" their history. This is not particularly helpful in identifying and selecting probative files or constructing knowledge from the data. It does, however, suggest an analogous approach, which I will describe shortly.

The other insight from their book is the notion that data are the product of a "Hierarchy of abstractions" and that computer information is comprised of "layers and illusions." As they explain, even the notion of files and directories are metaphors, rather than objective reality. Files are collections of ones and zeros, not manila folders, just as directories are merely specialized files containing pointers to other files, not a poster in the lobby. They go so far as to say: "As we peel away the layer after layer of illusions, information becomes more and more accurate because it has undergone less and less processing" (Farmer and Venema 8-9). I would argue that, from their archeological viewpoint, this may allow for a more accurate assessment of the mechanics of the computer system's activities, but it does not provide much, if any, assistance

with the knowledge management aspects of digital forensics and, may in fact, be counter-productive.

Farmer and Venema's notion of "discovery" is an archeology of system artifacts. By careful extraction of data and analysis, the examiner can reconstruct the computer's activity. This approach has great merit in the situations in which the computer's role in an investigation is as a *victim*, a *weapon*, and sometimes as an *instrumentality*. Unfortunately, these represent the minority of digital forensic cases. Far more common are the frauds, the child pornography, the intellectual property theft and the forgeries that comprise most of the personal and economic crimes. For these latter cases, what is required from the forensic examination is what can be gleaned from the contents of files and their fragments. Investigators and lawyers want the emails, the memos, the photographs, the spreadsheets, and the social/personal connections which speak to the actions and intentions that comprise the case.

I propose extending their scientific metaphor to the ethnography, history, literature and sociology of the computer. We are searching for the activities of users and the textual (in the broadest sense) record of their activities. Both the user's activities and the content of the files are mediated by cultural, social and technical factors unique to the user. Rather than analyze the content of the computer's hard drive as the record of the machine, I propose to look at the data as a large, semi-organized repository of data that is part archeological dig, part anthropological study and an un-edited anthology. Our mission is to find those data (files and fragments) which answer the relevant questions appropriate to the particular case, characters and crime.

The multiplicitous roles served by computers greatly increase the digital forensic practitioner's problem of defining the scientific questions to be asked and the tools and methods used to conduct the inquiry. Unlike the traditional forensic practitioner, who offers a limited set

of examinations that answer a relatively small set of questions, the digital forensic practitioner must spend a great deal of time custom designing the process and criteria to be used for each individual case. While similar to traditional forensics, digital forensics is far more complex.

But what is truly revolutionary about digital forensics is that the product of the examination/investigation is not an inanimate molecule or an objective instrumental value, but information or knowledge. And while data can be valuable, in most investigative and legal contexts, it is the higher levels of information, knowledge and wisdom, which are the desired products of the forensic process. The terms *data*, *information*, *knowledge* and *wisdom* have specific meaning in the field of knowledge management and will be explained in a later section. In this dissertation, they are used in that context.

The second theoretical construct of traditional forensics is therefore also pertinent to digital forensics, with two additions. First, practitioners often face a high degree of difficulty in defining the forensic questions to be answered. Secondly, the identification and extraction of the pertinent data is of limited value. It, in turn, must be further contextualized to form information, knowledge or wisdom.

Fortunately, both the third and fourth constructs of traditional forensic science are identical in their application, so we need make no changes with respect to those constructs.

Theoretical Construct for Digital Forensics

From the preceding, we may derive the following theoretical construct for digital forensics:

1. Digital Forensics utilizes science to make deductions and inferences on evidence of interest to the legal system that is stored or transmitted in binary form.

2. Digital Forensics utilizes validated tools that preserve the integrity of the evidence. Preservation of the original evidence, without alteration is critical to the reliability and authenticity of the results.
3. Digital forensics seeks to answer questions of identification, classification/individualization, association and reconstruction, as well as questions of: *who, what, when, where, why* and *how* within the context of a given piece of evidence and/or case.
4. Digital Forensic examinations may yield information that relates to the operation of the digital system, the content of stored or transmitted data, or both.
5. In most digital forensic examinations, evidence is submitted with limited specific legal or investigative questions either explicitly stated or implied. Forensic examiners must develop logical, effective and defensible strategies for location, selection and presentation of information in cooperation with the interested parties. Digital forensics is often a knowledge management process.
6. Forensic science seeks to act in an ethical manner and in as neutral a fashion as possible. Quality is actively managed and subject to both accreditation and judicial review.
7. Forensic practitioners communicate the results, conclusions and opinions obtained from their examination in the form of written reports and oral testimony.

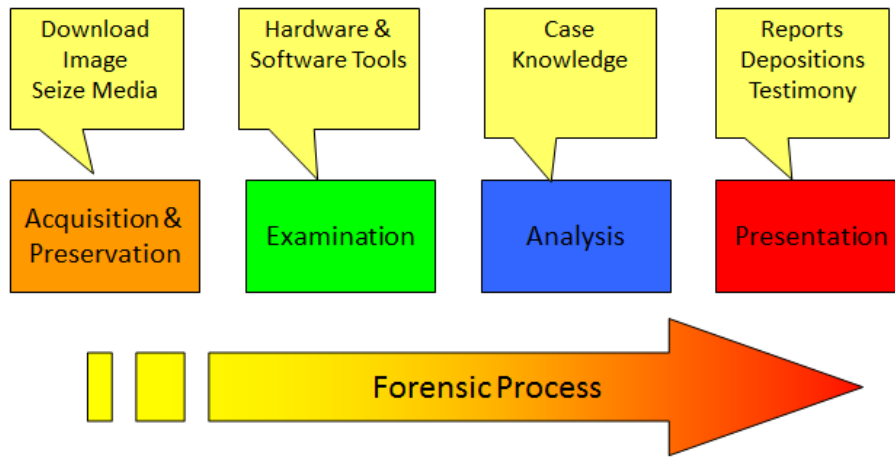
The Practice of Digital Forensics

There have been many models developed to describe the digital forensic process (Pollitt, Ad hoc; Reith). Virtually all are process models and many have iterative phases or sub-processes. What they all have in common is they attempt to conceptualize the relationship between the data created by computers, and the data created by humans and attempt to extract the pertinent materials. Somewhat implicitly, many of the models attempt to maximize the automation of the process, while minimizing, or at least facilitating, the examiner's human interaction. By doing so, these models may be trying to accomplish two goals. From a practical standpoint, automating the examination of data is perceived as more efficient, if not more effective, than the use of human judgment. A secondary reason may be to support the notion of objectivity. There is a common perception, even among scientists, that "scientific" testing is "better," meaning more objective, when performed by machines than humans. While the latter assertion is logically correct, it does not measure the pertinence, or utility, of the result. Automated processes may give you accurate answers, but not necessarily useful or correct answers. In this dissertation, I suggest that the integrity of the evidence can, and should, be demonstrably preserved, but that finding meaning in human communication requires a human intellect. The goal of this research is to find ways to augment that intellect with linguistic tools. Once the original evidence is preserved and the content reliably extracted, the analytical process can proceed. The interpretation of this reliable evidence, as well as, the conclusions and opinions obtained from this analysis, can be tested against the original evidence or through other evidence.

The Digital Forensic Process

For the purposes of this dissertation, I adapted a commonly-used four step model of digital forensics that I have been using since at least the mid-1990's and that has been codified in

the National Institute of Standards and Technology's Special Publication 800-86 (Kent 3-1). I have annotated the model with some explanatory notes as shown in Figure 1.



Copyright 2008 Mark M. Pollitt

Figure 1: Digital Forensic Process

Many models, including the one described above, articulate a set of steps, some of which use the terms “examine” and “analyze. In the model adopted for this dissertation, there are four phases: acquisition/preservation, examination, analysis and presentation. I will briefly describe each of the phases in the following sections.

Acquisition/Preservation

The first phase, acquisition/preservation, is where the original hard drive, storage media or electronic data are captured, exactly reproduced, and stored by a process that ensures that the original data has not been altered in any way. In legal terms, this allows the evidence to be admitted into court as “authentic.” As this phase is largely technical and serves primarily to ensure the reliability of the evidence, it does not have any further relevance to this dissertation.

Examination

By *examination*, I mean the observation, documentation and interpretation of the technical artifacts of all of the evidence, such as the file systems, metadata, log files and physical structure of the data. This might be characterized as a “computer science” view of the evidence. In this phase, all of the data is extracted, documented and evaluated. If appropriate, objective conclusions and/or opinions about the characteristics of this data may be made by the forensic examiner. For example, after examining the database that the computer’s operating system keeps of devices attached to the particular computer, the examiner could conclude that there is no entry for a specific USB drive and form an opinion, based on this observation and additional review of other entries in the database, that it is unlikely that such a drive was ever attached to this computer.

The most common first step of the examination phase is to extract all of the data. All of the “active files,” that is, the files that are available to the user, are extracted first. Then the “deleted files” are recovered. Deleted files are those whose data are intact, but are not available to the user. Most operating systems merely mark files as deleted, and do not over-write the actual data, at least initially¹. Once both of these sets of files are recovered, then the examiner will utilize software to look at all of the remaining space on the storage device. This space is called un-allocated file space, as the file system does not keep track of the data stored in this area – it believes that this space is available for storing new data. Most modern forensic software will parse these areas looking for things it recognizes as files in a process called “carving.” When it finds what it believes is a file, it will make it available to the forensic software. Any remaining data in unallocated space will be stored as “objects.”

¹ It is possible that portions of a “deleted file” may be recoverable, while other parts are not. The portion of the file which is recoverable can be classified as a “recovered file,” albeit only partially. The un-recoverable portion of the file will be classified as un-allocated space.

Technically, the first three items are correctly called “files” as they are data created by an application and stored by the computer’s file system. The last item, the unstructured data located in unallocated space, are called *objects* and not *files*, as they have lost their connection to the file system and identifiable structure as *files*. *Objects* are defined as data that can be manipulated by software, and since most forensic software operates on both *objects* and *files* interchangeably, we will, for the purpose of this dissertation, include them in the use of the term *files*.

Analysis

The next phase of the examination is a review of the data, using the “lens” of the investigative case or legal issue. This is the “analysis” phase. The review is conducted with the goal of putting the data, both technical and content, into the investigative/legal context; how does this data relate to the case? The result of this phase of the examination is information which directly relates to the case at hand. For example, the examiner finds an email with incriminating information. While the examination phase will conclude that this email exists and is a reliable product of the computer’s operation, in the analysis phase, we identify the significance of the data to the case. The import of this email may not be merely in the content of the file, but the metadata conclusively shows an electronic connection to the sender or recipient. This latter conclusion leverages the examination finding of the existence of the email with the content and metadata.

This distinction is an important one, as the reliability of all the evidence is established in the acquisition and examination phases. If the first two phases, acquisition and examination have preserved the integrity of the evidence, then the subsequent analysis is, at least with respect to the content and metadata, reliable. Given the increasing immense quantities of data in digital forensic cases, there is a need to have a separate process to select relevant information. The

analytic process, cannot be wholly objective, as human judgment is utilized, either directly, by viewing the data, or indirectly, by setting up search parameters. But, if the examination phase was done completely and correctly, then whatever is selected during the analysis will, nonetheless be admissible in court.

Many of the digital forensic tools (software), that have been developed, demonstrate this approach. Figure 2 shows the summary screen for a popular digital forensic software application. It classifies all of the files by application type and can display the file's metadata. A third window allows the examiner to view the contents of the file. The software has another tab, seen in Figure 3, which will allow string searches of all, or a sub-set, of the files. While the interface is relatively elegant, certainly compared to a command line tool, it is clearly not a very efficient approach to finding a narrative.

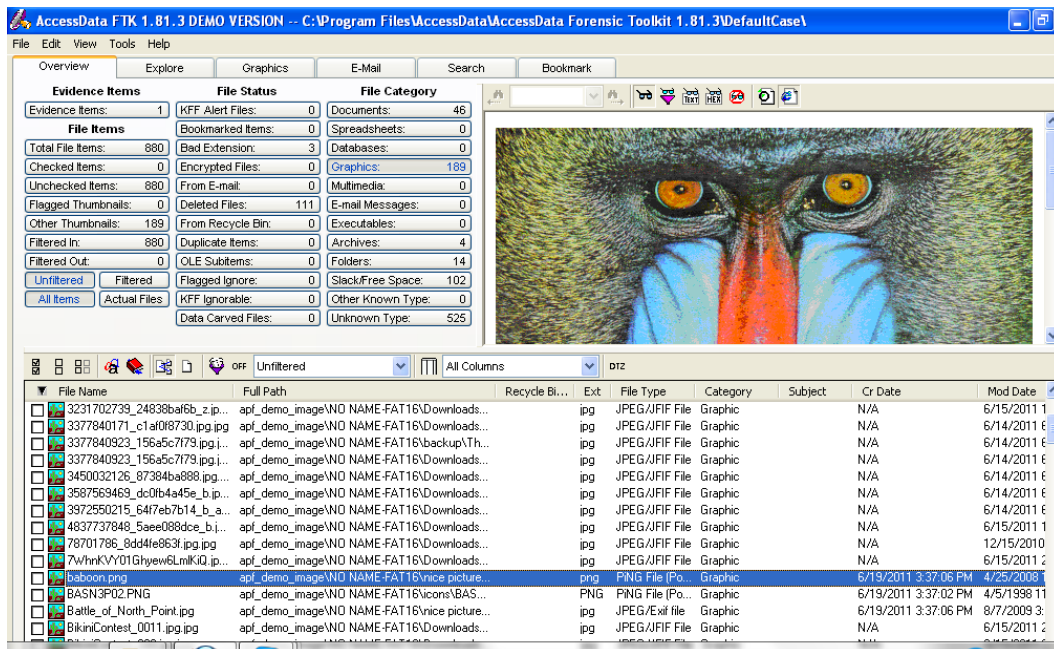


Figure 2: Forensic Toolkit Overview Screen

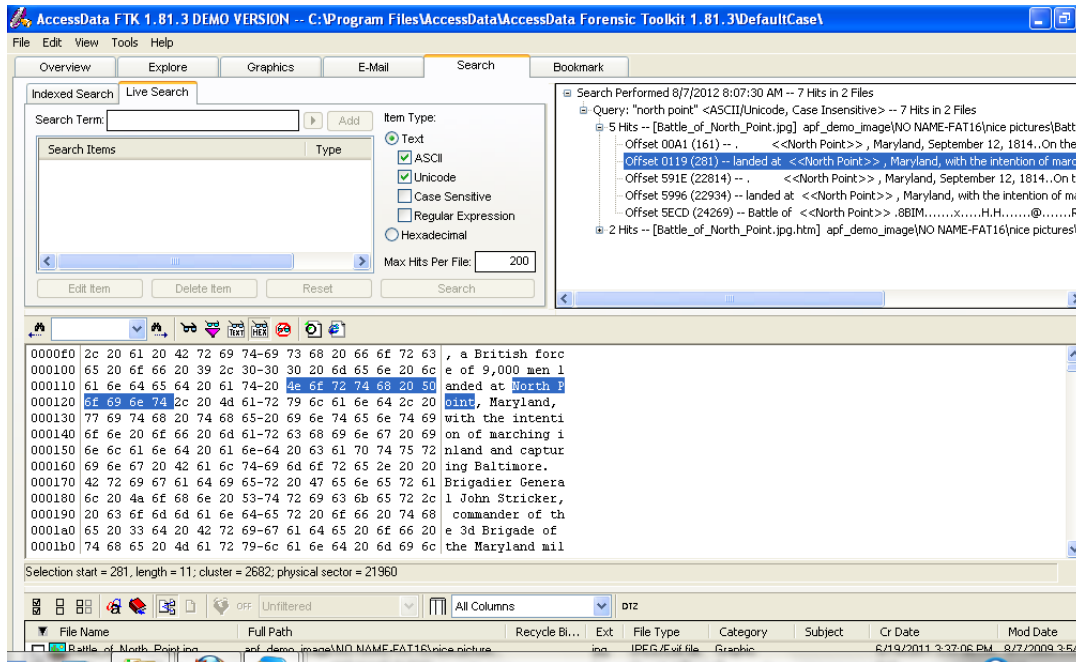


Figure 3: Forensic Toolkit Search Screen

In this dissertation, I will focus primarily on the content of the files with the goal of identifying probative narratives within the text contained in the file contents.

Presentation

The last phase of the forensic process communicates the results of the previous phases to the “clients” of the forensic examination. These may include corporate security officials, law enforcement officials, lawyers and ultimately jurors. Commonly, a written forensic report is submitted which is sometimes followed up with briefings, sworn depositions and/or courtroom testimony. In addition to transmitting the results of the examination phase, the examiner will sometimes make scientific conclusions and/or form professional opinions. And while the communications generated during this phase are crucial to the use of digital forensics and strongly mediate the results, for the purposes of this dissertation, we will not comment on this phase further.

Digital Evidence Challenges

In many ways, digital forensics sits at the intersection of the individual, society, information, and technology, and as such, it has huge and increasing challenges. Some of the challenges can be divided into those which are intrinsic to the process of digital forensics and those which are extrinsic. There are technical challenges, as well as challenges to the equities involved in both the resolution of the legal proceeding, and also, issues involving privacy. These problems are not new; they have existed since the very first legal proceeding involving technology. With the ever-increasing volume and dispersion of technologies, the challenges are becoming broader, more problematic, and urgent.

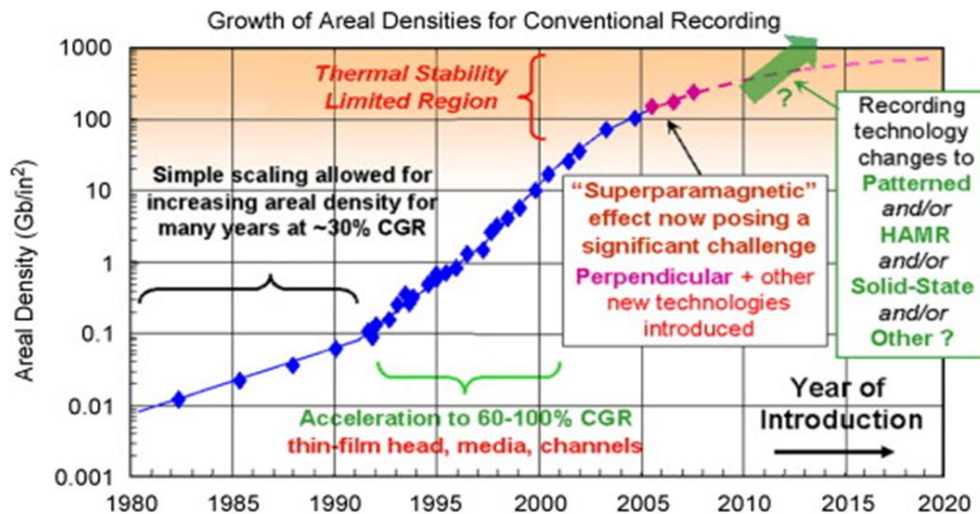


Figure 4: Areal density of HDD products vs. year of introduction. From Wood, R. "Future Hard Disk Drive Systems." *Journal of Magnetism and Magnetic Materials* Volume 321, Issue 6 2009 555 - 561.

If you were to ask almost any digital forensic practitioner what their single largest problem was, it is very likely that they would state it is the constantly increasing volume of data. This can easily be seen in the sales of digital storage devices, the growth in size of typical storage units and the volume of casework performed by digital forensic practitioners. In a 2009 article, Wood documented that the ability of hard drives to store information, in a physical space,

referred to as “areal density,” had grown at a compound rate of 44% per year for over 50 years. This is graphically shown in Figure 4. Nanotechnology has the potential to continue this accelerating miniaturization (Menon & Gupta 1117).

Table 1:RCFL cumulative case load: FY 2003–2010. From Vassil Roussev, and Candice Quates. “Content Triage with Similarity Digests: The M57 Case Study.” Digital Investigation 9.Supplement S60–S68. Print.

Fiscal Year	Processed Data (TB)	Number of Cases	Avg Case (GB)
2003	83	987	84
2004	229	1304	175
2005	457	2977	154
2006	916	3633	252
2007	1288	4634	278
2008	1756	4524	388
2009	2334	6016	388
2010	3086	6564	470

As technology becomes more integrated into daily life, the demand for digital forensic examination has grown very quickly. In 2000, the FBI started a program to develop regional digital forensic laboratories (RCFLs) in partnership with other federal agencies, as well as state and local law enforcement. In 2005, there were 10 such labs employing approximately 200 digital forensic examiners, who processed 1.4 Petabytes (1,400 terabytes) of evidence data (Reagan). Roussev provides an excellent chart which depicts the growth in RCFL casework (Table 1). It is significant that not only are the numbers of cases, and the total volume of data increasing every year, but the volume of data per case is likewise increasing dramatically. In Fiscal Year 2011, the 16 RCFLs conducted 7,629 examinations and processed 6,243 terabytes of data. This report describes a single terabyte of data as roughly equivalent to 1,000 encyclopedias or a stack of paper (FBI, RCFL Annual Report FY2011, 4). In six years the volume has increased more than four-fold. Using the FBI’s data from FY11, each case averaged over 800 megabytes

(0.8 terabytes). Compare this to Roussev's 2005 data, where each case averaged a mere 154 megabytes (0.154 terrabytes).

Clearly, there have been, and continue to be, rapidly increasing volumes of data for the digital forensic examiner to process. The volumes are such that, as a former colleague of mine, Marcus Thomas, used to say: "We're not looking for a needle in a haystack; we're looking for a needle in a needle stack." It would be bad enough if the only problem were one of volume, but additional trends are making the problem even more difficult. One of these is the increasing complexity of the data, caused by principally two factors: structural complexity and complexity of use.

The structural complexity of the computing environment and the expansion of data types and locations, make the job of the digital investigator and forensic examiner substantially more difficult. From the mid-1990's onward, computing has become progressively more networked. It began with local area networks, dial-up Internet connections, and text-based desktop computers. In less than twenty years we have smartphones and tablets with both still and video cameras, social networking capability and constant network connectivity. In 2012, smart phone manufacturers expected to ship 567 million devices and by 2016, it is anticipated that over 1 billion phones will ship per year (DisplaySearch). It is not only the sheer volume, but the complexity of how they are used. As Kaplan explains: "The wide variety of document types, tremendous volume of dissimilar media, operating systems, programs, and compaction and encryption algorithms all present daunting tasks for the examiner to efficiently organize, process, and filter" (57).

Volume and complexity do not fully express the dramatic changes in the task of the digital forensic examiner. The combination of rich technologies, in the multimedia sense, with

the constant, universal connectivity has reified McLuhan's two aphorisms: "we become what we behold" and "We shape our tools and afterwards our tools shape us" (McLuhan & Lapham 19). The effect on the digital forensic examiner is the need to select and interpret the data on the hard drive in multiple, different ways, and to collate and synthesize these disparate data to form some coherent narrative of probative information to investigators and the courts. Traditional documents, such as letters and spreadsheets created in software such as Microsoft Office, must be combined with things like Internet history files, emails, instant messages, and artifacts of social networking software, to form a holistic view of the hard drive's owner. Cloud-based applications seem poised to overtake hard drive based applications (Foley). Applications, such as Google Docs, Windows Live, Dropbox, and Evernote™ have become commonplace; examiners must be aware of these sources of data and either deal with them directly or transfer the task of obtaining that data to an investigator or attorney.

Volume, complexity and the sheer velocity of evolution are, and always have been, the greatest technological challenges to digital forensics. There is no reason to believe that these elements will resolve in the foreseeable future. There is yet another challenge and it is the ability to select, organize and present the relevant information. It is, in a sense, a "reading comprehension" problem for this vast collection of texts.

In the early days of digital forensics, I like most of my peers, would literally print out every single character on the storage device, onto continuous, "green bar" computer paper. The examiner would then read, or at least scan, every page of the printout. In the 1992 World Trade Center bombing, we not only had to print out and read all the data, but we had to initial and date each and every page. Later, we would produce CDROMs, and later, DVDs of much of the data, and provide that to the lawyers and investigators. Now, it is common to provide investigators and

lawyers with software which allows them to conduct their own searches of the raw, or sometimes refined, data, over a network connection to a massive storage array (Craig Law Enforcement).

But regardless of how the data was provided, the often predictable outcome was that often the material was never looked at. Sometimes this was because investigators lacked even a modicum of technical ability, but more often it was a cultural bias. Investigators and lawyers are used to having people tell them what is important. In one sense, they preferred to interrogate the forensic examiner rather than look through the material themselves. But having been both an investigator and an examiner, I recognize something else at work. Investigators, as they are collecting information from interviews, are eliciting a narrative, which they in turn, extract information from and create an investigative narrative. For example the interview might sound like this:

Investigator: Mrs. Jones, tell me what you saw when the man came in to rob the bank?

Mrs. Jones: I noticed a tall man, wearing jeans, a red and blue plaid shirt and a ski mask, walk past the line of customers, directly to the blond teller. He took out a big gun, which he pulled on and it made a loud noise as the metal “thingy” was released...

The Investigator’s report might read something like:

The subject proceeded directly from the North entrance of the bank and proceeded directly to teller station #4, occupied by Ms. Bridget Smith. The suspect announced the robbery by producing a semi-automatic pistol, drawing and releasing the slide... The subject was described as...

When the examiner writes her report, it is often written in technical terms, but organized in a manner that facilitates the examiner's testimony. It is very common for investigators to not even read the report, but to call the examiner and ask a very narrative question like: "what are you going to be able to testify to?" The answer is very often communicated in the form of a narrative. As we will explore further in a later section, this notion of narrative is critical to the entire legal process.

One alternative to the process described above is for the examiner to take an active investigative role and review the evidence himself, trying to identify pertinent information. Taken to extreme, the investigator and examiner can be the same person. This is done in quite a few law enforcement organizations – the investigators do their own forensics. While at first glance, this would seem to have some benefit, there are several substantial drawbacks. The cost of training and equipping a digital forensic examiner is very high. As the technology continues to evolve, training and tools must be updated. From a resource management perspective, it does not make sense to only utilize these investments "part-time." From a human resources perspective, there are relatively few investigators who also have the technical background to perform complex examinations. Those that do have both of these skillsets are highly marketable, and therefore unlikely to remain in government service. But perhaps the strongest argument against this approach is the abandonment of any verisimilitude of scientific objectivity. This lack of independence may undermine the unique privilege accorded to scientific expert witnesses in court.

In summary, the three major technical challenges to digital forensics are the increasing volume of data; the increasing complexity of both the data and the use of technology by individuals; and the problem of identifying, extracting, and reporting probative information.

Beyond the technical issues, the search for “meaning” in digital evidence is an old and continuing problem. Almost from the start of written communications, questions have arisen about how to interpret what is written.

Early structuralists, such as Saussure, recognized that text itself is not simple, rather, it is a complex interaction among signs, signifiers, and the signified. Others, such as Barthes and Derrida, explored the notion that a text may have a multiplicity of meaning (Leitch 1466-70, 1822-76). Iser explored the notion that there are two roles in the interpretation of text: the reader and the author, and there is no face-to-face interaction to correct intent or interpretation. Iser describes how the “blanks” and “gaps,” as well as the structure of the text “as a tacit invitation to find the missing link” (Leitch 1677) We will explore these theories later, but it is important to note that while this problem is well known, little study or research has been undertaken. This dissertation seeks to add to that discourse.

Opportunities in Digital Forensic

For over two decades, law enforcement, the legal community and the intelligence community have struggled to capture, understand and utilize information in digital form. The law enforcement and intelligence communities have struggled with notion of digital forensics, as either a forensic science or an investigative art. It has proved to be resistant to either classification, and making the recognition of it being both the only viable option. But the mixing notions of science and art is an uncomfortable in many circumstances, but especially in the law enforcement and intelligence communities.

In the traditional forensic sciences, there is a core, underlying science, or group of sciences, which define a particular forensic science. For example, in forensic toxicology, the underlying sciences are chemistry and physiology. For the forensic toxicologist to understand the

mechanisms of poisons on the human body, they must understand the chemistry of the substances, both organic and inorganic, as well as the physiology of how the body deals with these substances. In the examination of firearms, it is necessary to understand the materials involved through the study of chemistry and engineering, as well as learning the physics of how items interact with their surroundings.

If digital forensics is a science, what branch of science is it? The traditional view has been, that since all things digital are fundamentally that combination of mathematics and engineering that is commonly called computer science, and therefore the scientific component of digital forensics must be computer science. The fact that we use computers, software and mathematical algorithms to process our evidence surely defines this branch of forensics as a form of computer science, or perhaps not. It is true that computer science has created the means of production, storage and examination of digital evidence. Some of the software tools developed over the last two decades have proved useful in identifying data of probative value. But once we recover all of the data from the storage device, we are faced with what are inherently “investigative questions,” the traditional *who, what, when, where, why,* and *how* questions. Given the tremendous volume of digital evidence there is an even greater question: “what is important?” And while some of these questions may be answered utilizing the technical information and metadata from the digital evidence, most of these questions are fundamentally about the content of the digital storage. Being able to identify and select the probative content is the next great challenge in digital forensics.

Merely selecting probative information is a worthy goal in and of itself. Given the volume of the evidence, it is likely that the probative information will, itself, be voluminous. How to select the most probative information, analyze, organize, and present that information is

a second major challenge. It is interesting to note that the 2006 study on “The Future of Intelligence Analysis” synthesized, from a series of workshops, a list of background, skills and traits for successful analysts. Of the ten skills listed, nine of them were cognitive or communication skills, and only one was information technology skill. In addition to being an effective researcher, writer, communicator, and briefer, the desired skills included problem solving and interpersonal skills (Lahneman Append. A). This suggests that a humanities approach may provide the foundation for the next level of digital forensic analysis.

The study of how digital forensics might identify, extract and analyze probative information from large corpora of digital data has applications beyond the law enforcement, legal and intelligence communities. While great strides have been made in the areas of artificial intelligence and data mining, the ability to more efficiently extract information from unstructured data, in a way that more closely mimics the way human beings understand, remember and communicate knowledge, may be a useful adjunct to the computational methods currently in use. Conversely, a structured and computationally enhanced methodology for examining the structure and meaning of narratives might provide new ways for humanities scholars to explore both traditional and “new media” texts. It might suggest pedagogy for introducing the technologically native student of today with a deeper appreciation of the aesthetics of narratives, as well as the social and emotional context of communications.

CHAPTER TWO: FRAMING DIGITAL FORENSICS AS A TEXTUAL PROBLEM

In the previous chapter, we examined the theories of traditional and digital forensics, as well as their methodologies. The tools and techniques, currently in use, do an excellent job of acquiring and examining digital evidence within the computer science context. They are far less useful in extracting probative content from large bodies of data. We have seen that a computer science approach, where we analyze items of potential evidence, by reference to the items technical characteristics and their literal values, does not provide a sufficiently powerful methodology to achieve the desired goal of articulating the probative value contained in the item.

Finding all of the deleted material merely adds additional volume to examine. Locating instances of words, or simple phrases, may sometimes be useful, but often it is not. There is a disconnect between the levels of abstraction: looking at ones and zeros usually does not tell us who did what, to whom. In some ways, it is a bit like using a microscope to understand the geology of a volcano. The microscopic examination of material from a volcano does provide a great deal of useful information about the volcano, but several aerial photographs may provide much more actionable information, if the goal is to protect the surrounding community.

Computers, and by extension the networks to which they connect, are complex. In order to get the core functionalities of input, output, storage and computation, the designers of computers must begin with physical devices, such as transistors and connectors. These are sometimes referred to as the “glue logic” of the computer; how the various elements are wired together determines what they can do. These are normally assembled onto “motherboards,” to which additional devices, such as hard drives and monitors can be attached. At the other end of the spectrum, there are the software applications, such as Microsoft Word®, which the user actually utilizes. In between, is software called the operating system, which acts as the interface

between the hardware level and the applications. There are various operating systems, such as Microsoft Windows®, Linux, and OS X®. Each will run on a specific set of hardware, but all share the same core functionalities, one of which is the need to permanently store data. In order to do so, they utilize standards agreed upon by computer manufacturers and operating system developers. These standards, call “file systems,” bear names such as FAT 16, NTFS, HPFS, and ext3. In a similar fashion, there are a number of layers which constitute a computer network. What makes digital forensics so technically challenging is the complex and multiplicitious interactions between hardware, operating systems, applications, networks and users. As a result, digital forensic practitioners often use a layered abstraction to examine computers.

Eoghan Casey, in the first edition of his book, *Digital Evidence and Computer Crime*, pioneered the use of the OSI model of computer networks, a conceptual metaphor that describes the layered functions that allow networks to function, that describes what evidence would be found at each layer (Casey). Brian Carrier, in his book, *File System Forensic Analysis*, also used a layered approach to reconstructing the digital evidence recorded on storage devices (Carrier). I have used similar approaches, which I called “peeling the onion” (see Figure 5), or the “hierarchy of access,” to describe the examination of digital evidence.

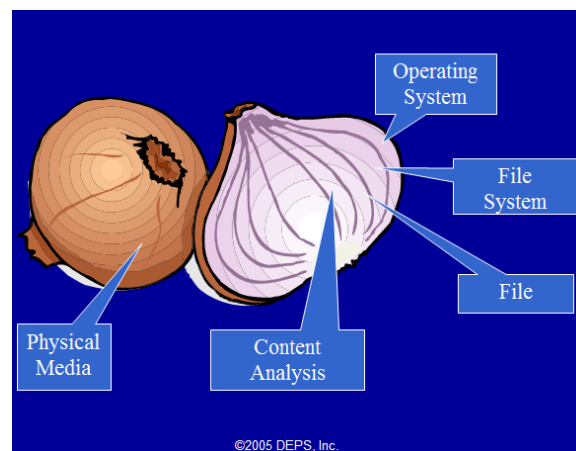


Figure 5: Digital Forensics as Peeling the Onion

What is the appropriate level of abstraction for digital forensics? Clearly, there is not a single answer to this, as the law and the facts will be different in each case. There are cases, such as the examination of malicious software (worms, viruses, and logic bombs), where a very technical examination is required. However, most legal matters deal in the physical world, with physical acts committed by human beings. It is on this level that, most of the time, the courts operate.

We have previously identified two of the most pressing problems with respect to digital forensics as: volume and meaning. Instead of focusing on the mechanics of the computer and their storage devices, I suggest focusing on the issues concerning with these two issues more directly.

For the volume issue, it would be useful to be able to reduce the volume by identifying what is important and ignoring that which is not. Knowledge management, as an approach, has evolved over the last two decades to deal with the problems of managing, evaluating, and distributing large volumes of data.

We have demonstrated the textual nature of all evidence, including digital evidence. We have recognized that story, or narrative, is an important element of investigations, forensic examinations, and litigation. I suggest we explicitly frame digital forensics, not as a computer science problem, but as a textual one. In so doing, we open up the analysis of digital evidence to many different analytical approaches. For this dissertation, I will focus on several textual concepts, or theories, that might be useful and I will describe how these theories might be applied to the digital forensic process.

A Knowledge Management Approach

In the latest study of the total volume of digital information, published by USC, “How Much Information Is There in the World?” there is a stunning analogy; the information stored in the world in 2007, was 291 exabytes, a number approximately 315 times the total number of grains of sand on the planet. Ninety-four percent of this information was stored in digital form. “From 1986 to 2007, the period of time examined in the study, worldwide computing capacity grew 58 percent per year, ten times faster than the United States' GDP” (Hilbert and Lopez 60-65). This exponential growth in data is demonstrated every day in the digital forensics field. The result is the classic data glut and information famine. Somehow, forensic examiners must find ways to “distill” the vast amount of electronic evidence into useful, “actionable” evidence. As Dustin Wax says in his *Lifhack* article:

To tame information overload, then, is not simply a matter of restricting ourselves to sources that advance our immediate goals in some way....Instead, we need to rethink our relationship with information and with work. Because information is, in the end, the building material that meaning is made of.

We now recognize the need to organize, prioritize, and extract value from ever greater corpora of data. Humans have been trying to get a handle on these issues for millennia. While the modern history of knowledge management dates to the 1990's, Davenport and Prusak's seminal work, *Working Knowledge*; argues that it began far earlier, perhaps as far back as the Greeks, who, as they transitioned from an oral to a written culture, attempted to organize their newly codified knowledge. Ong tells us that writing has both “created history” and “transformed human consciousness” (Ong 168,77-79).

In fact, it was the Greeks that gave us the modern term, *encyclopedia*. Romans, such as Cato and Marcus Terentius Varro, wrote collections of knowledge. Pliny the Elder wrote his *Natural Histories*, containing some “20,000 facts” (Carey 8-9). Later the Arabs and Chinese developed systematic structures to organize knowledge (Stockwell 17-27). In Medieval times, Isadore of Seville struggled to organize knowledge in his *Entymologies*, which he intended to be a compendium of, as Brehaut describes it: “that which ought to be known” (Brehaut 36-45). In *Omne Bonem*, James le Palmer is the first to textually classify information by means of alphabetization (Sandler 5, 16). During and after the Enlightenment, many encyclopedias and dictionaries were published. As Richard Yeo’s book, *Encyclopaedic Visions: Scientific Dictionaries and Enlightenment Culture*, points out, there were two major challenges faced by all of the historical encyclopedists: what to select for inclusion and how to organize the “facts” selected.

What then is knowledge management? In the many books and articles on knowledge management, there is a myriad of definitions. Perhaps the most succinct is that found in Hicks, Dattero, and Galup’s article “The Five-Tier Knowledge Management Hierarchy,” they offer the following definition, quoted from the Gartner Group:

Knowledge management promotes an integrated approach to identifying, capturing, retrieving, sharing, and evaluating an enterprise’s information assets. These information assets may include databases, documents, policies, and procedures, as well as the un-captured tacit expertise and experience stored in individual workers' heads. (19)

What this definition does not make clear is that knowledge management is a process by which the “value” of information is preserved or enhanced. Like a number of the theories described in this dissertation, the field of knowledge management has yet to adopt a singular set of definitions.

Davenport and Prusak state a hierarchy of value ranging from *data*, through *information*, and finally *knowledge*. They also recognize there are additional “higher order” notions, such as *wisdom* and *insight*, but they chose to include them within the definition of *knowledge*. They describe, as Hicks, Dattero, and Galup explicitly articulate, that there is a process, whereby less useful items, e.g. *data*, are capable of being transformed into more valuable entities, e.g., *information*, in Davenport and Prusak’s terminology. This transformation is the result of some value being added to the lesser object. That value may include a classification of the data or information, correlation to other data or information, or some form of contextualization. Both sets of authors describe both computer-based and socially based approaches to identify, select, and disseminate knowledge.

Within the field of digital forensics, what is truly unique is that the product of the examination/investigation, is not an inanimate molecule, or an objective instrumental value, but information and knowledge. And while data can be valuable, in most investigative and legal contexts, it is the higher levels of *information*, *knowledge*, *insight*, and perhaps even *wisdom* that are the desired products of the forensic process. Developing information may be probative, knowing how something occurred is more useful, understanding how a criminal deals with problems is better, and knowing the most efficient and just investigative approach is a worthy goal. In applying knowledge management to the digital forensic process, we seek, by means of a

process, to add value to our data and information. The goal is to produce investigative and legally probative *knowledge*.

For the purposes of this dissertation, we will utilize Davenport and Prusak's terminology. The core terms, as defined by these authors, are *data*, *information*, and *knowledge*. "*Data* is a set of discrete, objective facts about events" (2). *Information* is "*data* that makes a difference". It is data that is "transformed" in one of five ways: contextualization, categorization, calculation, correction, and condensation. And last, *knowledge*, is "a fluid mix of framed experience, values, contextual information, and expert insight that provides a framework for evaluating and incorporating new experiences and information. It originates and is applied in the minds of knowers" (3-5). I offer the following to exemplify these concepts.

"Tuesday," "AA11," "Boston," "September 11th," "hijack," and "Atta" are *data*; each word, taken individually, has little "meaning" in the investigative sense. The terms: "Tuesday," "AA11," "Boston," and September 11th," when taken collectively, identify a specific flight and are an example of *information*. If one were to connect the name "Atta," and the concept "hijack," with the preceding *information*, you would have what we have described earlier as "actionable intelligence," or in Davenport and Prusak's terminology, *knowledge*. You would know that Mohammad Atta would hijack American Airlines flight 11 on Tuesday, September 11th.

Pemberton describes two fundamental views of knowledge management. One view is a technological view, focused on computers and systems. The other paradigm assumes "that knowledge is primarily a human, rather than a technological resource." He further articulates two basic kinds of knowledge that acquired by acquaintance and obtained by description. The former is by "direct apprehension of sense data" and the latter "is mediated in the sense that it can be

developed, stored, delivered, and then acquired through the speech, or books, or written messages of others.” Knowledge by description can be challenged, tested, repeated, transmitted over time, and verified by others” (Pemberton). In digital forensics we utilize both approaches; however, we tend to be biased toward the former, as it “feels” more scientific, despite the testable notion of the “knowledge by description” model.

How does knowledge management work in practice? Forensic examiners are given an item of physical evidence: it could be a hard drive, a CDROM, or a database file. The evidentiary item can be likened to a corpus of data. Data aren’t useful until it has been interpreted as relevant to the case. It is the job of the examiner and/or analyst to select information that has value to the investigator. They do so, in part, by performing the transformations described by Davenport and Prusak. They contextualize and categorize the data “internally,” by identifying the data’s location, data type, file system metadata (file names, dates, and time stamps from the operating system), and application metadata, such as licensed user, print time. They also contextualize and categorize the information “externally,” based upon their knowledge and understanding of the type of crime in general and the specifics of the case at hand. It may be useful to provide an example of this process.

On September 1, 2012, a victim received an extortion letter in the mail. The police have conducted an investigation that identified a subject, Joe Doaks. Probable cause was developed and a search warrant for Doak’s home and computer was obtained. A digital forensic examination of Doak’s computer was later conducted. A file with the exact verbiage in the extortion letter was located. The file, a Microsoft Word document, was located in the “Downloads” directory of the computer. This file’s modified, accessed, and created dates are all September 2, 2012. A further examination of the file reveals that the internal metadata, which

was created on August 15, 2012, by the Microsoft Word program, indicates that the file's author was John Smith, of the Acme Company. This file, at first glance, appears to be a "smoking gun." However, by analyzing the internal and external metadata, it is clear that the file was likely created by Smith and downloaded to Doak's computer. These small details completely change both the narrative, and the significance of the file, to the investigation. Additional investigation and examination will be needed to clarify the relevance of this file, but this scenario demonstrates how the technical issues of internal and external metadata mediate the narrative.

For the most part, calculation, correction, and condensation play less direct roles, at least for the present, in the knowledge management aspect of the digital forensic process. While the internal contextualization is very amenable to software solutions, the external contextualization is a very difficult computational problem, as we will see, in the natural language processing section.

The Law as a Narrative

Criminal statutes define the behaviors that constitute crimes. They are the legislature's way of informing the government and the citizens of what is unlawful. They are written in prose, albeit technical, legal jargon, that is essentially narrative. For example, the following is an excerpt from the Federal statute prohibiting the use of weapons of mass destruction:

18 USC § 2332a - Use of weapons of mass destruction

(a) Offense Against a National of the United States or Within the United States.— A person who, without lawful authority, uses, threatens, or attempts or conspires to use, a weapon of mass destruction...

(3) against any property that is owned, leased, or used by the United States or by any department or agency of the United States, whether the property is

within or outside of the United States;...shall be imprisoned for any term of years or for life, and if death results, shall be punished by death or imprisoned for any term of years or for life. (Cornell)

In criminal cases, investigators collect information from interviews, documentary and physical evidence. These activities are usually compiled into one or more reports. If the crime is to be prosecuted, the accused will be formally charged. This is done using one of a variety of documents, such as a *criminal complaint*, an *information*, or an *indictment*. The former is a formal, written statement, proffered by the government, alleging that a crime has been committed. It is used for misdemeanor charges, or when the defendant waives his right to a Grand Jury Indictment. The latter, also known as a "true bill," outlines the actions which the Grand Jury believes constitute the violation of the law. The following is an excerpt from the indictment of Timothy McVeigh and Terry Nichols, for the bombing of the Oklahoma City Federal Building:

I N D I C T M E N T COUNT ONE

(Conspiracy to Use a Weapon of Mass Destruction)

The Grand Jury charges:

1. Beginning on or about September 13, 1994 and continuing thereafter until on or about April 19, 1995, at Oklahoma City, Oklahoma, in the Western District of Oklahoma and elsewhere,

TIMOTHY JAMES McVEIGH and TERRY LYNN NICHOLS,

the defendants herein, did knowingly, intentionally, willfully and maliciously conspire, combine and agree together and with others unknown to the Grand Jury to use a weapon of mass destruction, namely an explosive bomb placed in a truck (a "truck bomb"), against persons within the United States and against property

that was owned and used by the United States and by a department and agency of the United States, namely, the Alfred P. Murrah Federal Building at 200 N.W. 5th Street, Oklahoma City, Oklahoma, resulting in death, grievous bodily injury and destruction of the building.

2. It was the object of the conspiracy to kill and injure innocent persons and to damage property of the United States. (United States v. Timothy James McVeigh)

As can be seen from reading the law and the corresponding indictment, both are essentially narratives: they tell a story.

Evidence as a Narrative

William O’Barr, in his textbook, *Linguistic Evidence, Language, Power, and Strategy in the Courtroom*, provides an excellent description of a trial:

A trial might be thought of as a situation in which many people, often as many as 10 or more, present various versions of what happened. Their versions overlap to some degree and together tell a story. As the trial unfolds and opposing sides present evidence, it becomes clear that all versions cannot be equally correct. It is the role of the jury [or judge]... to decide which witnesses to believe and whose testimony to hold above others in reconciling differences. (11)

As was discussed in the Legal Notion of Evidence section, there are fundamentally two kinds of oral testimony: lay witness or expert witness. Since the former is limited to that which the witness “has personal knowledge of the matter” (Cornell FRE 602), the normal question, by the attorney calling the witness, will usually be something similar to: “Tell us what you saw when the man entered the bank?” The witness is, in effect, asked to tell a story. This serves two

purposes; it helps the witness organize her testimony, and it helps the jury to follow the sequence of events.

Expert testimony, under FRE 702, is different in many ways, but from a narrative perspective, has many similarities to the “story” told by the lay witness. The expert witness’s job is to apply scientific or technical knowledge to the submitted evidence, utilizing reliable principles and methodologies. If appropriate, the expert may render an opinion or state a conclusion, based on the results of the examination (Cornell FRE 702). In effect, the expert has at least three testimonial tasks. First, the expert must “tell the story” of the expert’s interaction with the evidence, for example:

Expert Witness: I received a hard drive marked ‘ABC123’ from Detective Jones. I then made a forensic copy of this drive and conducted my examination...I searched for... I located... I copied all of the files which contained information concerning XYZ Corporation to a CDROM, which I provided to Detective Jones.

In order to lay a foundation for the jury to understand the results of the examination, the forensic examiner often will have to explain how the science or technology works. For pedagogical reasons, this is often done before the actual results are offered:

Expert Witness: Detective Jones requested that I specifically look for any emails between... and Joe Doaks. I would like to take a few minutes to explain something about how email works, so that we can all understand what I found. In order to create an email you must have... The email program then sends the message to...The email server “forwards” the email...

The expert witness then needs to explain the significance of that information and how it “fits” into the context of the “case narrative,” which is, itself a narrative or story:

Expert Witness: On the computer I examined, I located a program for sending and receiving e-mails. The program, known as Thunderbird, was setup with a user account of Joe Doaks. I located an email from Sam Smith, whose email address was... The email stated “this is what I sent to the SOB,” and had a link to ...By looking at the email headers, I determined that this email was sent on September 2, 2012 at approximately...”

Usually, the expert witness will organize the information in a logical progression, either by chronology or topic. In effect, the expert witness is creating a meta-narrative, combining the narrative of the examination of the evidence, the narrative developed from the evidence, and the case narrative. By doing so, the examiner is organizing the information for the jury, placing it in into the narrative of the case, and laying out the logic of the opinions or conclusions. As used in this context, a conclusion is a scientific statement of fact, based upon the results of reliable principles or methods, while an opinion is a suggestion of likelihood, based upon the evidence, the results of the examination, and the experience of the examiner (Jones).

The preference for using narrative in a trial is embedded in the training of lawyers and “appears to be based on the implicit assumption that narrative answers are better received than fragmented ones” (O’Barr 77).

In the Introduction, the theorists of structural analysis of text were briefly mentioned. A core foundation for these theories has been the notion that text, in whatever form or genre, is a form of communication. As such, it has a sender and, one or more, receivers. In some forms, or

genres, this is explicit. In a business letter, the document is explicitly a communication between the sender and the addressee. In a novel, it is a bit more abstract, as we have the writer “narrating” a story for the reader. But something like a database, comprised of many records of transactions, does not look, at first glance, to be a form of communication. However, if you look at a database from the owner of the data’s perspective it is a bit like, each time he adds a record, saying “...note to self: remember this data.” Conversely, when the database is queried, he is “saying:” “Computer, tell me the answer to the following question.” Each time we store or access a computer’s data, or interact with software, we are having a conversation/communication with the human intelligence that created, stored, organized, or owns that data and/or program. It is this notion that marks computers and their data as *text*, in the texts and technology sense, and the organized transmission of that data, or the chronology of a computer’s activities, as potentially narrative.

Extracting and Communicating Narratives

In order to identify, extract, and evaluate narratives from the forensic corpus, we need some form of framework and theoretical underpinning. I have chosen to utilize some well-known linguistic theories with which to provide a foundation for our analytical techniques.

Structural Analysis and Semiotics

Ferdinand de Saussure, in his “Course on General Linguistics,” articulates the notion that all language is arbitrary, in the sense that we agree upon a systematic representation of communication through signs that, in and of themselves, have no intrinsic meaning. Since oral communication predated written communication, these signs were initially sounds, which, in most cases, had no connection to that which was being communicated. The *signs* are composed of two parts, the *signified* and the *signifier*. The former is the concept the speaker/writer is

wishing to communicate and the latter is the sound or word used, by mutual agreement, to represent the concept. The signified may be an orange-colored, dimpled, spherical fruit and the sign is the spoken or written word *orange* (Leitch 963-966). As we explore the application of linguistics to the field of digital forensics, there are two key concepts from Saussure's work which will help us form a theoretical construct. First, that language is an arbitrary system and as such, it can be studied systematically. It is not that non-arbitrary systems cannot be studied systematically, merely that because it is arbitrary, the choices made in constructing the system can be studied. Secondly, all language, written or spoken, is a form of communication where concepts are translated to an aural, visual, or tactile manifestation.

There were other theoreticians who extended Saussure's work. It is not necessary, for this dissertation, to expand on all of them, but there are several, whose work is useful to our inquiries.

Claude Levi-Strauss, the French anthropologist, recognized that language is a social construct, and therefore has meaning beyond the mere signified. The signified does not merely represent a physical object; it has a cultural meaning as well. A "lemon" may be a yellow, citrus fruit, but it can also represent the concepts of bitter, bright, or shoddy workmanship. He theorized that culture, and all that it embodies, is a form of symbolic communication (Leitch 1415-18). This suggests that the content of digital evidence must be interpreted in a social context, and these digital communications constitute a cultural structure.

A number of theorists, such as Lacan and Mulvey, have extended the notion of texts beyond the spoken and written word. Roland Barthes, in his *Image Music Text*, begins his book with the statement: "The press photograph is a message" (15). Barthes goes on to describe three levels of signification for visual images: informational, symbolic, and significance (52-54).

Barthes' "Introduction to the Structural Analysis of Narratives," begins with the following:

The narratives of the word are numberless. Narrative is first and foremost a prodigious variety of genres, themselves distributed amongst different substances... Able to be carried by articulated language, spoken or written, fixed or moving images, gestures, and the ordered mixture of all these..." (Barthes Image, Music, Text 79)

Digital evidence encompasses all of the categories enumerated by Barthes, and so I posit that digital evidence can be studied as a genre of narrative. But, it is a genre that, in turn, may be comprised of many subordinate genres encompassing all of the categories described by Barthes, and more. He goes on to suggest that narratives can be studied scientifically, as a deductive process. In so doing, he requires us to "first devise a hypothetical model of description (what American linguists call a 'theory') and then gradually to work down from this model towards the different narrative species..." (Barthes 81). This suggests that it may be possible to scientifically analyze digital evidence, from a narrative perspective, to identify genres of digital evidence, and to develop *meaning*, in the Saussurian sense of identifying shared concepts.

Todorov, in his "Structural Analysis of Narrative," described, structural analysis as: "kind of propaedeutic for a future." He takes a grammatical as opposed to a semantic view of narratives, which he likens to sentence structure. He suggests that the elements of narrative are: the *subject*, identified as a noun; the *predicate*, which he states "is always a verb"; and the *adjective*, which he describes as infusing a "quality" without changing the situation. After describing a grammatical analysis of structure, he suggests this grammatical approach can be used to further the study of narrative syntax, theme, and rhetoric (Leitch 2098-2104).

Narratology: Its Theory and Elements

H. Porter Abbott, in *The Cambridge Introduction to Narrative*, quotes Jameson as saying: “the all-informing process of narrative” is “the central function or instance of the human mind,” and quoted Lyotard describing narrative as “the quintessential form of customary knowledge” (1). From the earliest writings of the Greeks to the latest digital genre, the narrative has been a powerful way to convey, not only information, but the context of this information that transforms it into knowledge. To see just how powerful this context is, consider the following examples:

1. “yesterday,” “Tuesday,” “Sally,” and “restaurant”
2. “Sally wants all of us to meet on Tuesday, at the restaurant where we were yesterday.”

In the first example, we have data. One might, consciously or unconsciously, concatenate these, on the assumption, that since they are grouped together, they must have some relationship to each other. You would, in effect, be trying to construct a narrative. But, by placing all of the data into a sentence, complete with connective and modifying words, we not only organize the data, but give it an external construct, thus creating knowledge. We understand the “meaning” of the sentence. Interestingly, if we start with the complete sentence, remove the connecting words, leaving only the nouns and verbs, we still (usually) have enough to understand the meaning of the sentence.

In the case of digital forensics, doing string searches for words provides us with only data. Even if we were able to “know” what words were the truly pertinent, in terabytes of data, we would still not have a context to create meaning from the search results. Searching only for individual words ignores their grammatical context. If words do not sufficiently convey information, perhaps sentences are a more appropriate source? By utilizing sentences, we can

identify the use of words in their grammatical (parts of speech) and semantic contexts. Clearly the second example above tells a story, but is it: a) the story that we need to know? and b) is it part of a larger story that is the more important? From a digital forensic perspective, it suggests two questions: What makes something *a* narrative? What makes something *the* narrative? Or perhaps there are multiple, layered, and/or embedded narratives?

Abbott defines this notion of narratives, within other narratives, as “framing narratives,” each of which serve to mediate the “embedded narrative.” He invokes the notion of “frame theory,” developed by Goffman, which he articulates as “models of understanding” between the text and the audience (28-30). This helps to explain why extracting meaning from text is so difficult. We not only need to identify narratives, but we must figure out how they are layered and/or connected, and we need to understand the model of understanding.

Mieke Bal defines narratology as “the ensemble of theories of narratives, narrative texts, images, spectacles, events; cultural artifacts that ‘tell a story’” (Narratology 3). Abbott tells us that “narrative is the principal way in which our species organizes its understanding of time” (3). Bal points out, in her paper, “The Point of Narratology,” how the study of the narrative can be integral to science, as in the study of anthropology, or clarifying, as in studies that examine the rhetoric of science. It is through these two authors we will explore how narratology could be used to assist in our analysis of digital evidence.

At the core, narratology tries to dissect a narrative to see how it works. This is very useful for a number of purposes, such as teaching writing or literary criticism. For the purposes of digital forensics, we will use it for very limited purposes: helping us identify narratives in our evidence, perhaps to extract the narratives embedded in the computer files, and possibly even to help us create a meta-narrative from the collected embedded narratives, external metadata,

internal metadata, and operating system artifacts. Like many emergent disciplines, narratology struggles with some of its fundamental definitions. Abbott and Bal have very similar views about what narratology is about, but develop somewhat parallel theoretical frameworks with almost opposite terminology: a point Abbott recognizes in his book (18).

For Abbott, there are two possible kinds of narrative, one “compact and definable” and the other “loose and generally recognizable.” In the former, he refers to small, relatively simple narrative structures that, as he describes them, are the building blocks of larger structures. The latter are broader constructs, such as genres, which have an overall “narrative coherence” (14). The notion of narrative genres is useful to assist in parsing the content of digital evidence. We can utilize tools which help us “read” the data based on its form and structure. We can recognize that a particular text is email, because we see the *from*, *to*, *subject*, and dates from the data.

Abbott’s construct can best be shown by the following quote: “So far we have established three distinctions: narrative is the representation of events, consisting of story and narrative discourse; story is an event or sequence of events (the action); and narrative discourse is those events as represented”. He defines a narrative in two parts: story and discourse. The story is further broken down into events and entities (19).

For comparison, Bal describes, in *Narratology: Introduction to the Theory of Narrative*, her construct of the elements which make up a narrative as follows:

A narrative text is a text in which an agent or subject conveys to an addressee (‘tells’ the reader) a story in a particular medium... A story is the content of that text, and produces a particular manifestation, inflection and ‘colouring’ of a fabula... A fabula is a series of logically and chronologically related events that are caused or experience by actors. (5)

The two authors' differing views can be graphically represented by Figures 6 and 7.

H. Porter Abbott's Construct

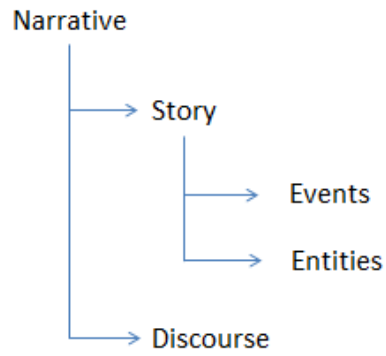


Figure 6: H. Porter Abbott's Construct of Narrative

Mieke Bal's Construct

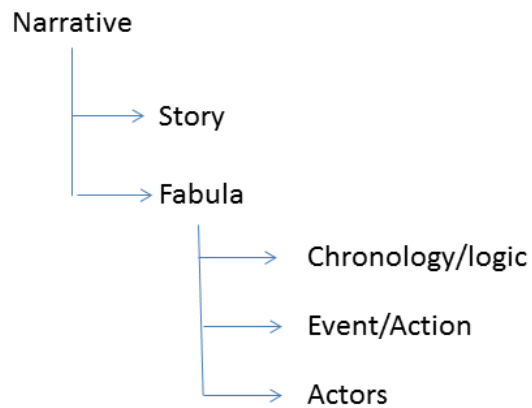


Figure 7: Mieke Bal's Construct of Narrative

As we can see from the two quotations, they use the term “story” in very different ways. Both authors differentiate between the content of the narrative and the way in which the narrative is told. The infamous phrase, “It was a dark and stormy night,” are the words used to describe, to the reader, the current weather at the scene. In other words, the content is the weather and “dark and stormy” are how the author communicates that content. The same weather could be written as: “The lightning pierced the black sky, through the torrential rain.” The weather is the same, as is whatever follows in the narrative, but the particular “telling” of the story is different. In Abbot’s construct, the actions, and events are labeled the “story,” and the words used to communicate the story are labeled “discourse.” In contrast, Bal defines the events and action as the “fabula” and the particular instance that communicates those events, as the “story.” For Bal, “dark and stormy night” is the “story,” while Abbott would label this “discourse.” Abbott discusses these differences and ascribes some of the differences to the linguistic origins of some of the terms, such as *szujet* and *fabula* (Abbott 18).

For our purposes, I will adopt Bal’s terminology, in part because she makes clear the three elements of the fabula: chronology/logic, events, and actors. I recognize that Abbott’s use of the term story is closer to the general usage of the term, but for clarity and consistency, I will use Bal’s terminology.

Bal’s construct can be interpreted as saying that there can be many ways to “tell” the story, but the elements of the fabula are the “factual” basis of the story. It is the elements of the fabula: chronology/logic, events, and actors that I will concentrate on in this dissertation. If that is so, and we can identify these elements, then we can identify the narrative, and in the process find the investigative or forensic story. The goal of our forensic

analysis is to identify these elements in our data/evidence, and by extension, identify the investigative and forensic narratives therein.

Why focus on the fabula? The “story,” in Bal’s terminology is far more complex and nuanced. Because of this complexity, and that, according to Bal, each telling of the fabula produces a different “story,” we are likely to get a more generalized, and therefore more accurate/authentic result from our analysis by focusing on the core elements of the fabula, rather than what she calls the “story.”

In digital forensics, if we can map the content of our evidence files to Bal’s notion of the story, and the goals for the particular forensic analysis, then we may be able to identify investigative or intelligence information of value. But in order to reach these objectives, we have to identify what files and information are likely to contain such information and should, therefore, be analyzed in depth. It is simply not practical to review or analyze every single file in terabyte-sized datasets. This is where the elements of the fabula come to our aid. If we can identify the events and actors, we can begin the process of developing a selection process. When coupled with both the internal and external chronology, we can, perhaps, “construct” a fabula.

Surrealism as a Tool

The level of cognition that would be required to extract all of the probative narratives, and none of the irrelevant ones, is unlikely, given our current understanding of digital forensics and the textual complexity of narratives. We cannot expect miracles. There is another theory that may assist us: surrealism. While the common understanding of this method, or theory, is artistic, its origins are in psychoanalysis, or more accurately, the psychoanalysis which existed in the very early part of the 20th century. Andre Breton, one of the originators of surrealism, served in a hospital during the First World War, where he observed and studied the treatment of soldiers

with mental illness. The treatment regime, “dynamic psychiatry,” called for the patients to “empty their minds of any conscious internal stimulation,” thus allowing for their “inner voice” to “speak, the results of which were written down “automatically.” Freud would later develop this into his free association technique (Gibson 56).

In 1924, Andre Breton wrote his “Manifesto of Surrealism” and he defined it thus:

SURREALISM, n. Psychic automatism in its pure state, by which one proposes to express, verbally— by means of the written word, or in any other manner — the actual functioning of thought. Dictated by the thought, in the absence of any control exercised by reason, exempt from any aesthetic or moral concern. (Breton 13)

Breton, and his fellow surrealists, sought to apply the notion of abandoning cognitive logic in order to free the artist’s imagination, and allow him to “automatically” create art without limiting one’s self to the constraints of consciousness. Ray describes how Benjamin, in his Arcades Project, describes the notion of “literary montage,” where “details counted.” Benjamin’s objective, “to convince... without conceptualization,” was described by Adorno as being “at the crossroads of magic and positivism.” Ray observes that Breton recognized “that science’s mistake was not its goal, but its methods, particularly its faith in procedures constrained by traditional logic” (Ray 41-7).

Surrealism was not limited to a literary context. In addition to writing, it was adopted by painters, photographers, and cinematographers. Sergei Eisenstein, in his book *The Film Sense*, reconciled the seemingly disparate notions of montage and narrative, by recognizing that the act of juxtaposing objects “resembles not so much a simple sum of one shot plus another shot — as it does a *creation*” (7). In so doing, the cinematographer creates a narrative. This narrative need not be presented sequentially. It may be told in a disjointed fashion, but the “spectator” will

construct a narrative. Part of this process, according to Eisenstein, is through the process of “condensation,” wherein the viewer disregards “the chain of intervening links” (14). The surrealist notion of juxtaposing disparate objects was not limited to synchronicity. But the particular notion, that we can create a narrative by taking non-sequential data points (shots), may well be useful in trying to identify a narrative from a corpus of digital evidence. If we can develop a narrative from a relatively few data points, then we can “understand the meaning” of our data. This would seem, at first glance, a far reach.

Among the surrealist “inventions” were surrealist games. Ray outlines several of these: one of which is called “exquisite corpse.” In this game, a syntactical construct is utilized. This construct is: What is a [noun₁]? A [noun₂], [adjective₁], and [adjective₂]. Each of four players is assigned to contribute one of the bracketed words. These words are then placed into the construct. An example, taken from Ray is: “What is the black bird? A Gregorian Chant, lucid and economical” (50). Recalling Kuhn, he explains: “Since the philosophy of science has shown that all knowledge systems rest on a few basic metaphors, and that a new paradigm always proposes a metaphoric shift, this game might have more profound consequences than at first appears” (49).

One of Ray’s students, Christopher Dove, designed a cinematic variation of another game called “irrational enlargement.” In this game, six scenes from a film are selected; three of which are “narratively important,” while the other three are not. The viewer is shown these clips, in random order, and is asked some substantive and conjectural questions about the film. The results of this experiment were surprising, in that, without any *a priori* knowledge of the film, the viewer was able to correctly identify major aspects of the story. Ray “suggests that film works even more economically than we had imagined, conveying an enormous amount of information in only three shots” (73).

It occurred to me that there might be some potential to utilize this in a digital forensic setting. I subsequently wrote a class paper, which was subsequently published, outlining the use of surrealism in a forensic context. To illustrate how a surrealist approach could be applied to emails, I noted that during a period of 20 days, I had 238 emails in my university email account. The senders and recipients were superiors, peers, and students. If one were to read all of the emails they “would recognize that I am teaching two courses, enrolled as a student in two others, and I am on several program committees and editorial boards.” A reader of all the emails might find a few personal emails as well. In total, it would be a fairly complete view of my life at that time. I then suggested that if someone read only 80% of the emails, they would not lose 20% of the content. In fact, since many emails are repetitive and/or contain copies of previous communications, relatively little content would be lost. With careful reading, some of the “missing data” can be constructed or surmised, from that which remains. At the other extreme, if one were to read only 1% of the emails, which would work out to roughly 2 emails, the reader would have to be lucky to have anything significant. The reader would certainly not have a complete picture of my email “life.” We have, in effect, a non-linear, sliding scale of montage. (Pollitt Surreal Narrative 12-13).

In reviewing my paper on using surrealism as a digital forensic tool, Prof. Barry Mauer provided an excellent insight. He suggested that while the scientific examination of evidence “needs to follow strict (positivist) strictures, the investigation and analysis need follow no such strictures.” He reminded us of the 911 Commission’s conclusion that the failure to recognize and prevent the tragic events of September 11, 2001 was a “failure of the imagination” (Mauer; National Commission 1)

I will propose, in a later section, to tap this approach: to identify a narrative from a montage of elements extracted from digital evidence.

CHAPTER THREE: METHODOLOGIES FOR EXTRACTING NARRATIVES

In the previous chapter I have examined some of the textual theories which might support an improved digital forensic capability. In this chapter, I will look at some of the technologies which might be profitably employed to improve the analysis. I will focus on the use of content analysis, extensible markup language (XML), and natural language processing (NLP).

Content Analysis

Most of the examination and analysis discussed so far concerns the structure of the data as it relates to the operation of a computer. Clearly, as was pointed out in the Introduction, we are trying to make sense of the information on a target computer or storage media as text, in the broadest sense. To understand text, we have to examine its content. For the most part, it is the content that is the signifier. For this aspect of our research, we will look at a field of study known as “content analysis.” While the topic of content analysis could be an chapter of its own, we will only briefly look at a couple of key concepts that are useful to our digital forensic inquiries.

Kimberly Neuendorf, in her book *The Content Analysis Guidebook*, defines content analysis “as the systematic, objective, quantitative analysis of message characteristics” (1). This definition is important to us for two reasons. First, with forensic science, we are looking for something that is systematic and objective. And while true objectivity, concerning socially constructed knowledge, is impossible, Neuendorf (11), substitutes the notion of intersubjectivity — “do we agree it is true?” rather than, “is it true?” The last part of the definition, dealing with the notion of message characteristics, might, at first blush, look problematic. But, if we consider that a “message” is a product of communication, and the files on a computer are all communications of some sort, then it makes more sense. There are really three ways in which all of the content on the computer is a communication. Most obviously, there are emails, text

messages and letters, all of which are intentional communications between the computer user and the recipient. While the user may not have intended it to be so, all of the user-created data on the computer, by way of the forensic examination, is being communicated, from the user, to the computer, and thence to the examiner. One could view the original evidence as a communication between the user and the computer, and the analysis as a communication between the computer and the forensic examiner/analyst. And finally, some data located on the hard drive are records of activity by the computer's operating system, including its communications sub-systems. Thus, they are messages between the computer, itself and other computers.

Klaus Krippendorff, in his book, *Content Analysis: An Introduction to Its Methodology*, provides another, slightly different definition: "content analysis is an empirically grounded method, exploratory in process, and predicative or inferential in intent" (xvii). He could have been describing digital forensics when he describes how content analysts seek to understand meaning and impact of communications and thus answer questions "for which natural scientists have no answers, and for which their methods are generally insensitive" (xviii). Similarly, he understands the blossoming volume of data, stating: "The large volumes of electronically available data call for qualitatively different research techniques, for computer aids. Such aids convert large bodies of electronic text into representations if not answers ..." (xxi). This dissertation seeks to generate representations which will assist the examiner and investigator to find narrative answers.

Ellen Hijmans recognizes in her article "The Logic of Qualitative Media Content Analysis: A Typology" that most published studies identified with content analysis as a methodology are quantitative and are often studies of mass communications. What literature that she was able to find concerning qualitative use of content analysis has been in the social sciences

(93-4). After reviewing 57 empirical studies, she identified five “logical” approaches to content analysis, based on Krippendorff’s model of framework and logic: rhetorical, narrative, discourse, structuralist-semiotic and interpretative analyses.

Based on Hijman’s descriptions, rhetorical analysis, because of its focus on word choices and structure, may be of use to indicate intent, while the structuralist-semiotic approach seeks latent meanings that are usually below the level of empirical evidence. Interpretative analysis is a methodology used for developing theories, especially in the social sciences, and as a result, is of limited value in a forensic examination. Similarly, she describes discourse analysis as focused on “intentions and conventions,” and thus, not a particularly good “fit” for a screening methodology. It is the narrative analysis that is of most interest to the digital forensic examiner. And while she emphasizes the “character” aspect of the narrative, it is the focus on actors and actions that make this important for forensic examination (4-12).

Roussev first proposed, in 2006, stream-based disk forensics, a technique which processes all of the data on the storage media without re-constructing the file system hierarchy or context. Garfinkel, in his 2010 article, “Digital Forensics Research: the Next 10 Years,” he discusses how “it may be possible to recover a significant amount of useful information from the drive without building the hierarchy” and points to his own 2007 article “Carving Contiguous and Fragmented Files with Fast Object Validation” (S70). In that article, he also suggests that files could be re-constructed from data streams by utilizing “semantic validation.” He describes how he “solved part of the 2006 Challenge using a manually tuned corpus recognizer that based its decisions on vocabulary unique to each text in question. Although this is an interesting approach, automating it is currently beyond our abilities” (S8). In his 2010 article, he also states that “Today there are only five widely used forensic data abstractions (S69). Four of those he

lists, are products of the technology, while one, “extracted named entities,” for which he gives examples such as “names, phone numbers, email addresses, credit card numbers, etc.” are content.

The notion of utilizing thematic clustering for searching digital evidence was suggested by Beebe and Clark in a 2007 paper. In this approach, they sought to reduce the “information overload” by utilizing a similar approach to web search engines; they would prioritize the results of a given search. Recognizing that the extensive data indexing utilized in most Internet search tools, for example Google Search, would be computationally infeasible in a digital forensic setting, they looked at a number of data mining techniques. They recognized that most were unsuitable for the purpose of digital forensic searching. One approach that seemed to have merit was the technique of text clustering, which they describe as an “algorithm [that] automatically derives the thematic categories from the data.” They would leverage this approach by utilizing van Rijsbergen’s cluster hypothesis, which they describe as “computationally similar documents tend to be relevant to the same query.” They list a number of studies which indicate that “clustered query results improve information retrieval effectiveness over traditional ranked lists.” They proposed to further extend this approach with the use of a neural network approach of Kohonen’s Self-Organizing Map (SOMs) (Beebe & Clark; Text String Searching S50-52).

Garfinkel’s and Beebe’s approaches are pioneering in the examination of content in the digital forensic setting. Their approaches suggest that by combining natural language tools and the textual content of digital evidence, knowledge can be extracted from textual data, while not bypassing the information stage, but rather contextualizing it by means of the content, rather than the metadata. The authors do not suggest that internal or external metadata is not valuable, rather they explore how the process could be more efficient by utilizing content.

An important goal in digital forensics is scientifically-driven knowledge management. We recognize that our “raw materials” are the texts encoded on the hard drive. To increase the value of those texts we need to examine their content. Content analysis becomes the “root” of our theory of forensic analysis. Given the strengths and weaknesses of the different content analysis approaches, it would appear that a narrative analysis may provide us with the most efficacious approach.

Extensible Markup Language (XML)

Drawing on Johnson-Eilola, Selber, and Selfe; Applen and McDaniel, tell us in their book, *The Rhetorical Nature of XML*, technical writers have become “symbolic analysts,” who are helping to transition us from “an industrial to an information economy,” by means of the knowledge management processes. They make a strong argument that these technical writers/symbolic analysts are knowledge managers (9). They describe the role of these symbolic analysts as:

1. Identifying what constitutes relevant and meaningful information.
2. Breaking this information down into specific elements.
3. Providing names for these elements.
4. Contextualizing these elements of information to best meet the rhetorical needs of their audiences. (8)

The roles they describe are, essentially, the same roles as those of the digital forensic analyst. In this context, I use the term *analyst* in the context of the analysis phase of the digital forensic process. Applen and McDaniels characterize extensible markup language (XML) as a “robust tool” for the analyst’s conduct of knowledge management (8).

The use of XML in digital forensics is not new, but has not been extensively employed. Beginning in 2005, Craiger began extending the Global Justice XML Data Model (GJXML) to describe the physical evidence, such as hard drives and other storage devices, as well as the file system and external file metadata. The class properties were developed, in part, by reference to existing digital forensic tools. Examples from the XML schema and a snippet from an XML data file are seen in Figures 8 and 9. The final version of DEML was published on the National Information Interchange Model website in 2009 (Craiger, DEML).

```
<xs:element name="File">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="filename"/>
      <xs:element ref="startSector"/>
      <xs:element ref="cDate"/>
      <xs:element ref="cTime"/>
      <xs:element ref="mDate"/>
      <xs:element ref="mTime"/>
      <xs:element ref="aDate"/>
      <xs:element ref="aTime"/>
      <xs:element ref="logicalLocation"/>
      <xs:element ref="applicationAssociation"/>
      <xs:element ref="deleted"/>
      <xs:element ref="attribute"/>
      <xs:element ref="examinerComments"/>
      <xs:element ref="bookmarkType"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
```

Figure 8: DEML Schema from Craiger, P. Digital Evidence Markup Language, 2009

```

<File>
  <filename>string</filename>
  <startSector>-9223372036854775808</startSector>
  <cDate>1999-01-21</cDate>
  <cTime>13:20:00-05:00</cTime>
  <mDate>1999-01-21</mDate>
  <mTime>13:20:00-05:00</mTime>
  <aDate>1999-01-21</aDate>
  <aTime>13:20:00-05:00</aTime>
  <logicalLocation>string</logicalLocation>
  <applicationAssociation>string</applicationAssociation>
  <deleted>>true</deleted>
  <attribute>string</attribute>
  <examinerComments>string</examinerComments>
  <bookmarkType>string</bookmarkType>
  <preview>string</preview>
  <pageBreak>string</pageBreak>

```

Figure 9: DEML Data File from Craiger, P. Digital Evidence Markup Language, 2009

In 2009, Simson Garfinkel published a paper, “Automating Disk Forensic Processing with SleuthKit, XML and Python,” where he describes the use of a standardized set of XML tags to represent forensic data “into a single XML structure that represents all of the file system and document metadata resident within a disk image.” He constructs an XML structure that accounts for information about the forensic image (copy) file, the tools utilized to create and examine the image, the structure of the original device, and information about each individual file. He suggests encapsulating each file’s XML data within a *<fileobject>* tag (Garfinkel, Automating Disk Forensic, 73-4). He provides the example shown in Figure 10.

```

<fileobject>
  <id>10</id>
  <filesize>16384</filesize>
  <partition>1</partition>
  <ALLOC>1</ALLOC>
  <USED>1</USED>
  <mtime>1227139058</mtime>
  <atime>1227081600</atime>
  <ctime>1227139058</ctime>
  <filename>file3</filename>
  <libmagic>ASCII text</libmagic>
  <byte_runs>
    <run fs_offset=' 31744' img_offset=' 64000' file_offset=' 0'
      len=' 2048' />
    <run fs_offset=' 35840' img_offset=' 68096' file_offset=' 2048'
      len=' 14336' />
  </byte_runs>
  <md5>3b4af47f8b542fb5d1bdaeec34563d89</md5>
  <sha1>b614d783fee50f053bd99da89401b89b013487a4</sha1>
</fileobject>

```

Figure 10: Garfinkel's XML Description of a File Object

Subsequently, Garfinkel published a Digital Forensic XML Document Type Description (DFXML DTD). His focus is on representing “the provenance of data subject to forensic investigation,” however, he only includes physical characteristics of a forensic image and the forensic tools utilized in the examination (Garfinkel, DFXML).

The Forensics Wiki (www.forensicswiki.org) article on the digital forensic use of XML lists a number of software tools which either create or utilize XML. A review of these entries shows that the primary focus is on documenting the structure of the evidence. This is no indication that it is being used to document the content of the data (Category: Digital Forensic XML).

Electronic discovery, sometimes referred to as “eDiscovery,” is the process of acquiring, analyzing and presenting evidence by means of the discovery rules in the Federal Rules of Civil Procedure. In traditional legal practice, there was an exchange of paper documents made available to the opposing parties. Since 2006, attorneys have been required to negotiate between the parties and the court what electronic evidence is needed for the case (Roberts). Recently, an

industry group has gotten together under the banner of the Electronic Discovery Reference Model. One of their projects entails creating XML schema for the exchange of evidence (ERDM).

The use of XML to document the structure of digital evidence is useful to standardize the nomenclature and representation of the data. It is also useful in processing the forensic data by acting as a kind of *lingua franca* for the field. However, little has been done with respect to the structure and representation of the content of files in a forensic setting. In a later section, I will suggest a Document Type Description (DTD) focused specifically on a narrative approach to forensic analysis.

Returning to our over-arching concepts of knowledge management, narrative and content analysis, the literature of digital forensics does not appear to demonstrate any substantive use of XML for these purposes. Clearly, the discipline would benefit from the use of XML as a content analysis tool. How then might that work? It would be nice if we could utilize *<who>*, *<what>*, *<when>*, *<where>*, and *<why>* XML tags for investigative questions and *<chronology/logic>*, *<event/action>*, and *<actors>* tags to display narratives. Obviously, this is a gross over-simplification. But that is, in effect, what the goal of digital forensics ought to be: the identification of probative knowledge in the appropriate context.

How then could we utilize XML for analytical and knowledge management purposes? The ability to designate arbitrary, in the mathematical sense, tags allows us to categorize the content of a given file or data in multiple ways. We can create tags which provide links to the physical structure of the data, such as the physical location on a hard drive; the logical characteristics of the file, such as the file name, directory, and file times; and for the contents themselves. By creating a set of tags for the physical and logical characteristics, we document

and preserve the legally required authenticity. Not only can we trace any tagged content back to its source, we link the external metadata to the content. We can create a set of tags which will allow us to identify semantic and lexical elements, such as parts of speech, as well as tags which will capture internal metadata, such as email dates and times. By using XML we can link the forensic, analytic, and investigative needs in a single structure.

In this dissertation I am proposing to extend both Craiger and Garfinkels' work by utilizing content tags. These tags will document both the internal metadata and the full content of the data. By utilizing NLP tools, I propose to extract semantic and lexical features which will assist in identifying narratives. As an example, I will utilize the email files from the Enron Corpus. This corpus will be discussed in detail in the experimental section of this dissertation, but for now it is sufficient to know that we will use a set of ASCII (plain) text files which represent individual emails between employees of the Enron Corporation during the period of time when the company's senior management were conducting a massive corporate fraud (W. Cohen).

```
<?xml version="1.0" encoding="ISO-8859-1">>
<!DOCTYPE Email [
<!ELEMENT Email (email_header, email_body)>
<!ELEMENT email_header (email_recp+, email_copy*, email_sender, email_subj, msg_id, email_date+)>
<!ELEMENT email_sender (CDATA)>
<!ELEMENT email_recp (CDATA)>
<!ELEMENT email_copy (CDATA)>
<!ELEMENT email_email_subj (CDATA)>
<!ELEMENT msg_id (CDATA)>
<!ELEMENT email_date (CDATA)>
<!ELEMENT email_body (sentence+)>
<!ELEMENT sentence (date_info, named_entity+, noun+, verb+)>
<!ELEMENT date_info (CDATA)>
<!ELEMENT named_entity (CDATA)>
<!ELEMENT noun (CDATA)>
<!ELEMENT verb (CDATA)>
]>
```

Figure 11: Notional Email DTD

In Figure 11, we see one example of how the internal contents of an email file might be tagged in XML. The emails are semi-structured data and consist of two major sections: a header and the body. The former includes the email sender, the recipients, the subject line, the message identification number and the date sent or received. The body consists of plain text, as organized by the sender. As such, it may consist of well-organized sentences and paragraphs or be substantially less grammatical. For reasons that will be explained in the experimental design section of this dissertation, we will use NLP to extract sentences (including non-grammatical phrases), as well as the nouns and verbs in each sentence. “Named entities,” which are typically proper nouns that are contained in a look-up file, will be tagged, as will any dates or time contained in each sentence.

Emails are a genre of documents. XML provides a way to build templates for varying genres of text. Properly designed, differing structures, organization and types of content could be encoded in structured and semi-structured files. Non-structured text could still be parsed and “sentences” could be created by combining noun and verbs, as well as named entities extracted.

Natural Language Processing

During the last half of the twentieth century, with the tremendous advances of science and technology, it was perhaps natural that scientists would seek to utilize computer technology to model, replicate, and understand human communication. This effort resulted in the development of theoretical approaches and technological explorations of the synthesis of these two areas. The field would come to be called natural language processing, or NLP.

Natural language processing, as described by Liddy, is an interdisciplinary study involving three primary disciplines: linguistics, computer science, and psychology (Liddy 1).

These disciplines represent both a skill set and an approach to problem solving. She defines the field in the following way:

“Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications.” (1)

One of the key elements of this definition is the “naturally occurring text.” The use of the word *texts* is used in the same broad sense as in the Texts and Technology program and includes both written and spoken language. The naturally occurring part of the definition differentiates text which is generated by people for communications purposes, from that produced either for research purposes or is produced by entirely computational means. I selected her definition because it incorporates two concepts that will help us decide what natural language processing can and cannot, contribute to our forensic analysis. The first is the notion of naturally occurring text. Analyzing known, structured (artificial) text, such as a database, is a straightforward computational process. Dealing with the variety of formats, genres, and content that comprise forensic data is a substantially more difficult problem. The second concept from the definition is that of multiple levels of linguistic analysis.

The definition makes clear the complexity and ambition of this field. The goal of NLP is not merely understanding, but replicating human communication. At its very simplest, communication requires a sender, receiver, and the code that defines the content of the message. But even the simplest communication is not that simple. Literary and linguistic theorists, psychologists, and psychiatrists, as well as computer and forensic scientists, have been trying to define and model the processes of human communication

for hundreds of years. While great strides have been made in the last few decades, natural language processing is still a work in progress.

Both Liddy (3) and Wilkes (2) agree that NLP had its origins approximately 50 years ago when computer scientists and linguists attempted to apply computer technology to handle real or near-real-time translation of foreign languages. This approach was somewhat successful in that literal translations could be accomplished with some degree of accuracy. However, the limits of literal translation quickly became apparent and it was clear that grammar and semantics were far more complex than computational techniques could address in the late 1940's through the 1960's. It was clear that text needed to be "represented at all their levels of meaning" and the machine translation was not the singular answer (Liddy, 2).

What followed was an examination, by computer scientists and linguistist, of the elements that composed language and the processes by which it could be examined and reproduced. Researchers also framed their work on some practical applications that dealt with specific dimensions of communication. Liddy suggests the following taxonomy of NLP applications:

1. Information Retrieval
2. Information Extraction (IE)
3. Question-Answering
4. Summarization
5. Machine Translation
6. Dialogue Systems

Levels of Natural Language Processing

Liddy also suggests a model for the levels of NLP. While these levels separate ways in which meaning is communicated, it is important to understand that they can, and often do, operate concurrently. These levels are described in the following paragraphs (Liddy 5-8).

Phonology

Phonology is the level of communication that deals with the sound of speech. Included are the notions of: the phonetic (sounds within words), the phonemic (variations of pronunciation), and the prosodic (fluctuation of emphasis or intonation). Liddy notes that these are significant from a vocal input or output point of view. Because this level focuses on speech, there does not seem to be much applicability to our present inquiries, except, perhaps, it might be of some value in interpreting misspelled words. However, that is beyond the scope of this dissertation.

Morphology

Morphemes are “the smallest units of meaning” (Liddy 6). As children learning to read, we are taught to separate the prefixes, suffixes and roots of words. Later, we learn to take complex words and break them down in order to ascertain their meaning. One NLP approach to dealing with morphology is to utilize software to determine the root word, from its morphological variants, in a process called “stemming” (Bird, Klein & Loper 107).

Lexical

The interpretation of individual words defines the lexical level of interpretation. At first this may seem the simplest, but Liddy reminds us that there are at least three different dimensions that can affect the interpretation of a particular word. Words can have different meanings depending on their part of speech. They can have different meaning in different

contexts. Lastly, their meaning may be modified by the surrounding words (semantic arguments). Lexicons are attempts to define these variables for use in a system of symbolic logic (7). As we will see later, there are software tools that “tag” each word with a part of speech.

Syntactic

The grammar of a sentence tells us a great deal about meaning. After all, it defines the notion of a subject and verb (action) of a sentence. At first glance it would seem that this would be the most powerful of all of the levels for determination of meaning. And although it is powerful, it also presents a great deal of ambiguity, as we use a wide variety of structures and organize these phrases through a system of dependencies. This continues to be a significant problem in NLP.

Semantic

It was Saussure who originated the concepts of the sign, the signified, the signifier, and the referent (Saussure). In NLP, the problem is to “disambiguate” these to establish true meaning. Liddy describes it as a similar problem to that of syntax, with the goal of permitting “only one sense of a polysemous word” (7-8). NLP uses multiple approaches to evaluate all of the potential senses of a given word, including: stemming, lemmatization, part-of-speech tagging, as well as, supervised and unsupervised word sense disambiguation (Jurafsky & Martin 47, 68,133-44,637-641).

Discourse

Discourse is the level that first looks at objects larger than the sentence. At this level, NLP attempts to understand the function of individual sentences within the greater text and anaphora resolution wherein referent words, such as pronouns, are replaced by their logical signifier.

Pragmatic

The pragmatic level is where we try to understand the larger or external context of the text. As Liddy describes it, it is how we “read into” the texts without tangible links or coding technologies (8). In knowledge management terms, this is the application of tacit knowledge.

As the first level, phonology, relates to speech, and as we are limiting our discussion to digital texts, we will discard this level for purposes of this dissertation (6).

The next two levels, morphology and lexical, relate to words, their meaning and structure. These two levels provide the foundation for our understanding of a text; for without these, there can be no grammar, syntax or discourse. Liddy describes these as “semantic primitives.” Scientists have developed software, called parsers, that deconstruct text into “tokens,” compare it to a database, and tag it with the properties of each token (6-7). The current state of the art for parsers, operating at these levels, is very effective (Diekema).

The syntactic level refines meaning by looking at the syntax of sentences to discover order and dependency. As she describes it, complete syntactic parsing is still a challenge, but for many applications, the current state of the art is sufficient (Liddy 7).

Liddy describes the semantic level as the part of the process where the disambiguation of multiple senses occurs and where humans analyze whole sentences for meaning. As she points out, while many people would think that this is the primary level for the development of meaning, it relies heavily on the preceding levels. Most tools for semantic parsing require either domain knowledge or sufficient training corpora (7-8). Given that most digital forensic cases consist of an eclectic mix of data; identifying appropriate training corpora is problematic.

The use of domain knowledge may be viable in well understood cases, containing large amounts of structured information, for example Medicare Fraud, where much of the evidence is

billing information and the criminality is systematic. For example, a common medical fraud technique is to “un-bundle” laboratory tests. When billed individually, laboratory tests cost much more than when they are charged as a group. Medical laboratories will sometimes utilize software to “un-bundle” the tests, resulting in fraudulent charges. A review of billing records by an experienced medical fraud investigator would instantly recognize this practice. However, in most cases, the human analyst may be the most effective and efficient “disambiguator,” able to draw upon the same human capacity that allows us to correctly identify the narrative in surrealist games (Ray). While there are problems in working at the semantic level, it is the “home” of the sentence. Sentences provide the “base pairs” of nouns and verbs, which are the DNA of our stories.

The discourse level interprets the larger structures, such as paragraphs, to develop meaning from the relationship between sentences. It serves to provide things like anaphora resolution, by inserting subject nouns for pronouns, e.g., *He took a walk*, is parsed as: *John took a walk*, and sense disambiguation, e.g., when we say *course*, we mean a compass heading to follow, rather than an academic program. This level can also be used to identify the structural components of the document genre, such as, introduction, main body, and conclusion. In order to do this, there has to be a mechanism to identify the genre and its component structures. Again, this is much easier with structured text than “natural text” (8).

The last level is the pragmatic, whose goal is to “explain how extra meaning is read into texts without actually being encoded in them.” In addition, they may need to be interpreted utilizing “world” or external knowledge (8). While this is a key element of the analytical process, it is difficult, given the state of the art, to implement this in forensic software.

Natural Language Processing in Digital Forensics

Natural language processing has seen some limited use in digital forensics. Roussev utilized some NLP techniques in attempting to analyze the internal metadata of files. His study relied on statistical and machine learning approaches, and his results indicated that any of the methods needed “manual tuning” to be effective (Roussev 7).

Beebe and Clark, in their paper “Digital Forensic Text String Searching: Improving Information Retrieval Effectiveness by Thematically Clustering Search Results,” explore the use of natural language processing to map the concepts contained in digital evidence. This is a particularly important paper. After a significant literature review, Beebe and Clark come to several conclusions. Their observation is that the current modality of looking for strings at a physical level is neither effective, nor scalable. The latter term, scalable, refers to the ability of a given process to be applied to greater volumes of data. In the case of string searches, given a sufficient volume of text, any given string becomes more common, lowering its discriminatory value. The computational “overhead,” coupled with “information overload” from the search results, make the current practice unusable on large data sets. The authors suggest exploring other information retrieval techniques, specifically text mining. They examine Google’s use of five page-ranking variables. Since several of these variables are particular to HTML, or web documents, and the initial parameter for these searches is a user-supplied query, they reject this technology; they then look at the text mining framework described by Fan, et al. (Beebe and Clark S49-51).

In their 2006 paper, “Tapping the Power of Text Mining,” Fei, et. al., described a knowledge management process utilizing text mining. They differentiated data mining, where the material to be processed is structured in defined packages, from text mining, which operates

on unstructured or semi-structured data. They use databases and XML documents as examples of structured data: they specifically mention full-text documents and emails as unstructured or semi-structured data. They describe a set of eight “technologies” that are useful in the process of text mining: information extraction, topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering (77-9).

Of the technologies described by Fei, et. al., several have particular relevance to digital forensics. Three of them — categorization, clustering, and concept linkage — deal with the grouping together of documents with similar themes or concepts. The authors describe, in the categorization section, that a common technique for processing words is to analyze them, independent of their grammatical structures, as a “bag of words.” The technique takes all of the words and views them without punctuation, paragraphs or even word order. This approach is often found in the NLP literature. It would seem somewhat counterproductive to ignore the semantic and narrative context when trying to evaluate the significance of a given word. This is one of the reasons that I propose trying to utilize sentences in the evidentiary corpus, as they tend to contain the semantic and narrative context of the evidence.

Fan, et.al., describe how the evaluation of sentences is a widely used strategy for summarization. Summarization seeks to reduce the amount of data to be processed and/or reviewed, in part, by attempting to identify the “main points and overall meaning.” However, as they describe it, the evaluation of sentences is done by statistically evaluating the sentences or extracting information found subsequent to a “key phrase” (79).

Their article provides a useful visual representation of the text mining process as shown in Figure 12.

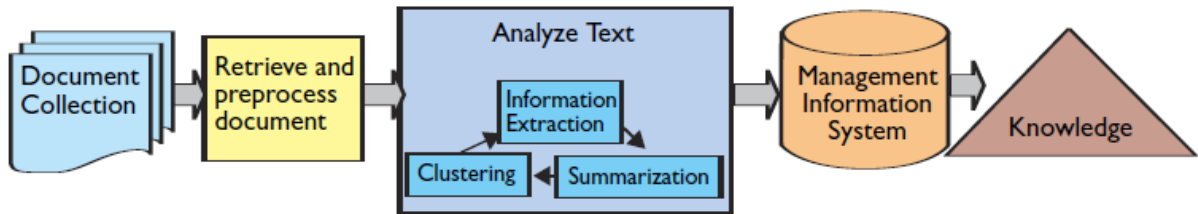


Figure 12: An Example of Text Mining. From Fan, et. al. Tapping the Power of Text Mining

Beebe and Clark focus on the text clustering modality suggested by Fan. They do so on the premise that “query precision (with respect to investigative objectives) can be improved by thematically grouping query result...due to the cluster hypothesis.” This, they describe, is from van Rijsbergen’s observation that “computationally similar documents tend to be relevant to the same query.” They extend this notion by utilizing Scalable Self-Organizing Maps (SSOM) to cluster the classified documents (Beebe & Clark S51-2). The preliminary results were reported at the 2007 Digital Forensics Research Workshop (DFRWS 2007), and suggested that clustering does provide assistance in the forensic analysis process, a conclusion that is further supported in their later work (Beebe et.al. Post-Retrieval Search 738-42). Fei, et. al., described, in 2006, how self-organizing maps could also be utilized to visualize forensic data.

Venter, Waal and Willers suggest, in their Specializing CRISP-DM for Evidence Mining, that it would be possible to build upon an already existing data mining process called the Cross-Industry Standard Process for Data Mining (CRISP-DM). They then proceed to modify the original model and produce a specifically forensic model they call CRISP-EM. One of the most useful parts of this work is the development of second and third level models to structure the process. Examples of these are provided in Figures 13 and 14.



Figure 13: CRISP-EM Level 2 From Venter, Waal & Willers

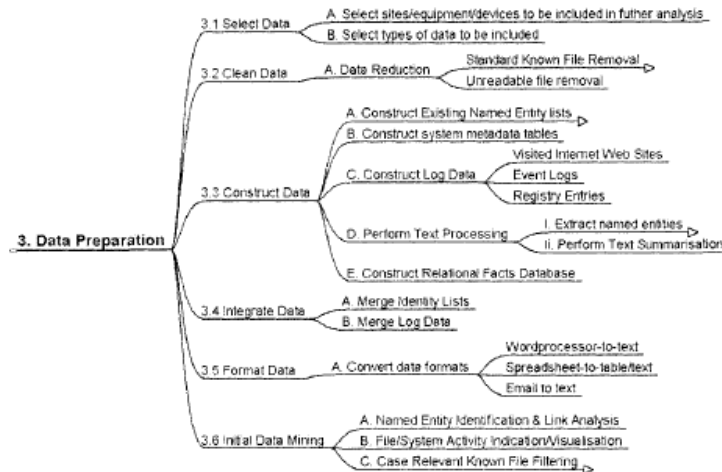


Figure 14: CRISP-EM Level 3 From Venter, Waal & Willers

Waal, Benter and Barnard, in their 2008 paper extend this work to a case study: “When one thinks of a text corpus as a collection of documents, it makes sense that each document has an underlying semantic context. This context develops as the document is generated, and refers to the intended meaning of the document.” Despite this, they suggest using a statistical analysis

of the words, in the text, to uncover the “latent semantic context.” In their paper, the term “latent semantic context,” appears to be synonymous with meaning. Further, they also recommend the “bag of words” approach (Waal, Benter & Barnard 117-8).

After applying their topic modeling approach, they offer some valuable lessons learned. The identification of “named entities,” nouns that signify a person’s name, a location or an organization, are very valuable, but they interfere with the pre-processing of data. Since the topic modeling algorithms include word frequency, the repetitive use of a named entity would skew the result. For example, if hundreds of email from Enron were processed, the number of occurrences of the word *Enron* would likely skew the distribution of more probative words. They suggest removing them for the pre-processing phase, and returning them to the analysis as concepts, not as words. Similarly, the use of “stemming,” a technique where words with a common root are represented by the “root stem” word, interfere with “the understanding of topic distributions. A third lesson is that the removal of words that only occur once in the corpus, has a disproportionate impact on the number of words removed from a corpus. In most corpora, this technique reduces the corpus by approximately 5%. In the forensic corpus utilized by the authors, the amount of data removed was approximately 50%. Given this fact, the authors suggest that it is likely that by removing these hapaxes, significant useful information will be discarded.” Lastly, they recognize that forensic data is often comprised of a variety of different document types (e.g.: letters, emails, etc.) and it may be useful to model each genre independently (123-5). These lessons seem to suggest that the traditional NLP processing steps, which result in the “bag of words,” may actually detract from the effectiveness of topic modeling. This is interesting, as topic modeling is more content-focused than many traditional NLP approaches. I would suggest that it also supports the notion of using sentences as the base structure for narrative analysis.

Carole Chaski's work has taken a very different approach to utilizing natural language processing in forensic analysis. Chaski, who is a forensic linguist, has focused on automated and semi-automated methods of authorship identification and assessment of risk. This work has its genesis in the investigative problem of assessing and attributing threatening communications. To this end, she has developed a commercial service, utilizing NLP software, to make these assessments. In her 2007 paper, she conducts an experiment that assesses the ability to discriminate among different authors in two different approaches. The first processes the documents as a whole, while the latter utilizes a five sentence "chunk." She reports that while discrimination was higher for longer documents, the level for the five sentence "chunks" was still high and suggests that small documents, such as blog posts, might be classifiable (Chaski 143-45).

In a later paper, "IntentFinder: A System for Discovering Significant Information Implicit in Large, Heterogeneous Document Collections," Chaski and her co-authors suggest an integrated approach that utilizes network mapping techniques, along with some sophisticated NLP techniques to process large collections of documents with the intent to extract "stories" from the data, based on a user query. In the paper, they discuss the shortfalls of topic mapping and clustering. They make an insightful observation that: "We actually want more fine-grained models of information, rather than topics." This suggests that the *relationship* between the subjects, the subjects and the topics, and the temporal context, may be at least as significant as the topic itself. The authors propose not merely to identify a story, but to be able to show the progression of the story over time (Ungar, Leibholz & Chaski 219-20).

The first element of the IntentFinder system is a document management system which appears to pre-process the corpora to extract phrases utilizing statistical n-gram techniques. The

paper doesn't explicate how document metadata is utilized, but they note: "Different document types have different metadata, which will also be collected: who wrote it, where, when, who was it sent to or where was it published, etc." (Ungar, Leibholz & Chaski 220-1).

The story extraction portion of IntentFinder utilizes a two-step process. The first is an algorithm that "associates entities and metadata," in effect, clustering potentially related material together. In the second phase, a "story" will be constructed from these clusters. The paper provides no specifics on how that is done, but they do discuss the literary theory that stories "generally follow a structure of exposition, climax and denouement." They do not explicitly describe how they codify or evaluate these structures in software, but careful reading of the paper seems to suggest that somehow, looking at the changes over time, in the network nodes described below, these narrative structures can be identified. They couple the notion of structure with a network analysis, where "each node represents some entity and connections between nodes denote the relations between entities and actions relating the entities." They describe this as a "schema." They summarize their approach by defining story extraction as fitting the information from the documents into the schema (Ungar, Leibholz & Chaski 221).

In the analysis phase, the authors describe the need to conduct analysis to determine the levels of significance and reputation. These topics, while interesting, are not of immediate application to this dissertation. Another element of IntentFinder is the lexical-semantic analysis. According to the authors the goals of this phase is to classify "(1) the document's role in the story (expository, climatic, denouement) and (2) the document's text type (such as promising, threatening, agreement, reportage.)" (Ungar, Leibholz & Chaski 221-2). This latter analysis is similar to sentiment analysis and is an example of the evolving field of "affective computing"

(Montoyo, Martinez-Barco, & Balahur). Similar work has been done by Chaski, and others, in connection with evaluating threat and suicide documents (Chaski, Howald, & Parker).

Liddy recognizes that the lower levels are primarily “rule-based” versus the higher levels, which require much more complex processing (8-9). She also defines the following classes of NLP applications: information retrieval, information extraction, question-answering, summarization, machine translation, and dialogue systems (12-13). The last two are beyond the scope of our current inquiry. And while summarization of all of the text contained in a hard drive would be phenomenal, it is currently beyond the capability of current technology. Similarly, having a robust ability to ask both broad and narrow questions of our evidence would be spectacular, however, at the present time, like summation, it is impractical and we will disregard both for the present. Most of the applications that make use of the higher levels of analysis utilize training corpora. But, as Jurafsky and Martin point out: “One implication of this is that the probabilities often encode specific facts about a given training corpus” (92). It is yet another example of technology mediating the text.

The difference between the simple “string searches,” that are the norm in current digital forensic practice and the use of NLP, is the difference between information retrieval and information extraction. In the former, we must know precisely what it is we are asking for and the result is nothing more than what we ask. We merely locate text. Information extraction is a richer and more complex process; it requires processing the raw text to identify and tag key elements. Given that there are numerous robust parsers that will identify parts of speech and basic sentence structure; information extraction would appear to be a potentially more fruitful approach to extract narratives from evidence.

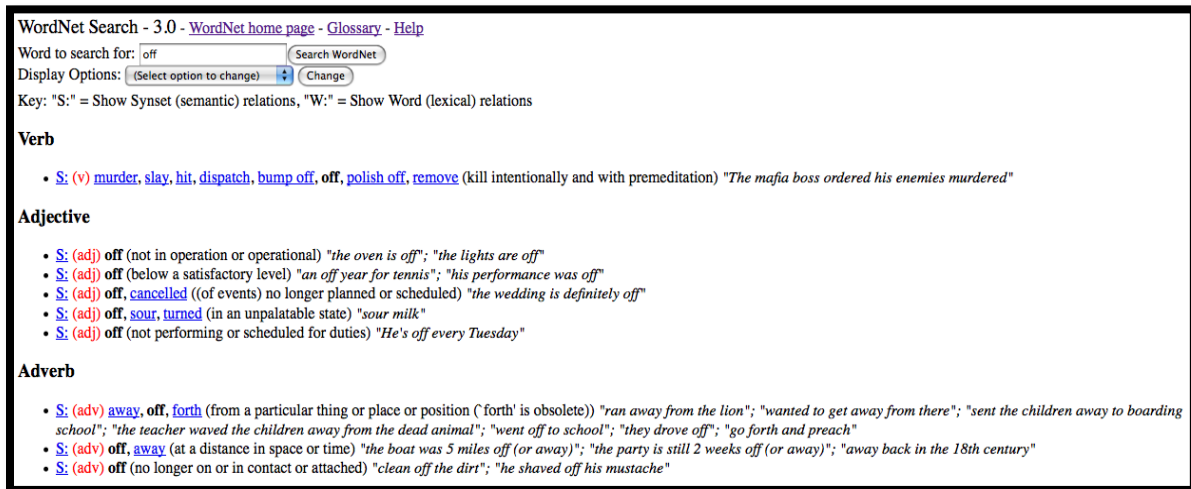
NLP Technologies

WordNet

An example of the type of application that may have utility is Wordnet, a project that combines part-of-speech tagging, along with lexical and conceptual mapping. WordNet is a project of the Cognitive Science Laboratory at Princeton University. It includes a database of English words, classified by their part of speech, and organized into sets of “cognitive synonyms” called synsets, which are “interlinked by means of conceptual-semantic and lexical relations.” It exists as both a library of routines and a website (Princeton).

The website and software utilize a set of index, data, and auxiliary files. Separate index and data files are created for each part of speech, and the auxiliary files serve to manage the searches, the output, and to handle exceptions. WordNet returns semantic and lexical relationships, as well as the parts of speech.

An example of a WordNet website search is shown in Figure 15.



WordNet Search - 3.0 - [WordNet home page](#) - [Glossary](#) - [Help](#)

Word to search for:

Display Options:

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Verb

- S: (v) [murder](#), [slay](#), [hit](#), [dispatch](#), [bump off](#), [off](#), [polish off](#), [remove](#) (kill intentionally and with premeditation) "*The mafia boss ordered his enemies murdered*"

Adjective

- S: (adj) [off](#) (not in operation or operational) "*the oven is off*"; "*the lights are off*"
- S: (adj) [off](#) (below a satisfactory level) "*an off year for tennis*"; "*his performance was off*"
- S: (adj) [off](#), [cancelled](#) ((of events) no longer planned or scheduled) "*the wedding is definitely off*"
- S: (adj) [off](#), [sour](#), [turned](#) (in an unpalatable state) "*sour milk*"
- S: (adj) [off](#) (not performing or scheduled for duties) "*He's off every Tuesday*"

Adverb

- S: (adv) [away](#), [off](#), [forth](#) (from a particular thing or place or position ('forth' is obsolete)) "*ran away from the lion*"; "*wanted to get away from there*"; "*sent the children away to boarding school*"; "*the teacher waved the children away from the dead animal*"; "*went off to school*"; "*they drove off*"; "*go forth and preach*"
- S: (adv) [off](#), [away](#) (at a distance in space or time) "*the boat was 5 miles off (or away)*"; "*the party is still 2 weeks off (or away)*"; "*away back in the 18th century*"
- S: (adv) [off](#) (no longer on or in contact or attached) "*clean off the dirt*"; "*he shaved off his mustache*"

Figure 15: WordNet Search Example

There is a stand-alone version of Wordnet available. But more usefully for this research, it has been integrated into a powerful, open-source natural language processing package called Natural Language Processing Toolkit (Bird, Klein and Loper 67-70)

Natural Language Processing Toolkit

The Natural Language Processing Toolkit (NLTK) was developed at the University of Pennsylvania in 2001. Originally designed as part of a computational linguistics course, it has now been widely used in both research and teaching. It exists as an extensive Python library which offers a wide variety of modules including parsing, tokenizing, stemming, part-of-speech (POS) tagging and chunking. NLTK has also integrated WordNet into its libraries. (Bird, Klein and Loper xiv; Perkins 7) NLTK is utilized in teaching, research and commercial products (Miller).

I will utilize NLTK for some of our experiments. That is not to say that this will be easy, as Wilks points out, “[scientists] have many formal achievements in print, they have had little success in producing any general and usable program to translate English to formal logic.”

Applying Natural Language Processing

In order to analyze and manipulate a collection of text (corpora), we must first break it up into semantic elements such as words, sentences, and paragraphs. This process is called tokenization. Software tools have been developed which perform this function. For example, the software understands that words are combinations of letters, “A” through “Z,” and are separated by a single space or punctuation. In turn, sentences are combinations of words and are separated by periods, exclamation points, question marks, and semi-colons, followed by a single or double blank space. Paragraphs are formed of one or more, sentences followed by a carriage return/new

line character. Each identified word or sentence is referred to as a “token” (Bird, Klein & Loper 8, 109, 234).

Once the words and sentences are identified, we can then process (parse) the text to identify words by their parts of speech, or in the parlance of NLP, word classes or lexical categories. This process is called part of speech tagging (POS tagging) and outputs both the token (the word) and a “tag,” which is an abbreviation for its part of speech. All of the tags utilized by a tagger are called a “tagset” (Bird, Klein, & Loper 179). Table 2 is an example, taken from the default tagset utilized in NLTK.

Table 2: Subset of NLTK Tagset

Tag	Meaning	Example
DT	determiner	all an another any both each either every half la many much nary neither no some such that the them these this those
NN	noun, common, singular or mass	common-carrier cabbage knuckle-duster Casino afghan shed thermostat investment slide humour falloff slick wind
PRP	pronoun, personal	hers herself him himself it itself me myself one oneseif ours self she thee theirs them themselves they thou thy us
VB	verb, base form	ask assemble assess assign assume atone attention avoid bake balkanize bank begin behold believe
VBD	verb, past tense	dipped pleaded swiped regummed soaked tidied convened halted registered cushioned exacted snubbed strode aimed

Words can, of course, have different parts of speech, depending on their usage. Taggers will often assign multiple POS tags for a single word. It takes a rather more sophisticated computation to disambiguate the particular use of a given word. Fortunately, many tokenizers have this facility built in, such as, in the example, taken from NLTK, in Figure 16. The original phrase, “She used a needle to needle him,” is parsed into tokens and tags. In the example, the

tokenizer is able to recognize that the first use of the word *needle* is a noun, while the second use is as a verb.

```
>>> text = nltk.word_tokenize ("She used a needle to needle him")
>>> nltk.pos_tag (text)
[('She', 'PRP'), ('used', 'VBD'), ('a', 'DT'), ('needle', 'NN'), ('to',
'TO'), ('needle', 'VB'), ('him', 'PRP')]
>>>
```

Figure 16: NLTK Tokenization and Tagging Example

While this is certainly impressive, the level of disambiguation achieved by any tokenizer is limited by the complexity of human communication. A further improvement can be achieved by utilizing statistical modeling of target text. For even better results, the statistical model can be refined by utilizing training data, from which the model can predict the likely meaning of a given word or phrase. In part, this is accomplished by recognizing what is called the Markov assumption. This theory posits that the likelihood that “we can predict the probability of some future unit without looking too far into the past” (Jurafsky and Martin 88). A simple example would be that the word “merry” is far more likely to precede the word “Christmas,” at least in American texts, than the word “happy.” Markov models accomplish this by looking at the relative frequency, in the training corpora, of tuples (ordered series of elements) of words preceding the target word. In the simplest form, as in the previous example, they look at the bigram: the target word and the word preceding it. Even higher accuracy can be achieved using more than one preceding word. The number of preceding words is usually represented with the letter N, hence the expression N-grams to represent these word groups (Jurafsky and Martin 88-113).

I previously interviewed Dr. Anne Diekema, acting Director of the Center for Natural Language Processing (CNLP) at Syracuse University. She confirmed much of the research that I

had earlier conducted. She, and the staff of the CNLP have undertaken a number of research projects, although none were focused on digital forensics. Diekema's view was that there are many well-developed software packages that provide syntactic and semantic functionality, but in her opinion, the state of NLP is not sufficiently mature to effectively deal with complex and rich material such as a suspect's hard drive. They are more effective with structured data or finite domains. She used, as an example, one of their projects that seek to create meta-data tags from library card catalogs. NLP works well in this context, as the card catalogs contain well-structured data.

In 2008, I interviewed Dr. Kevin Crowston, who was then the incoming Director at the CNLP. One of his research interests is developing ways of understanding group dynamics through the examination of communications. He has focused some of his current research on looking at the emails, blogs and other communications within the open source community. I discussed with him my focus on exploring ways in which to understand the hermeneutics of the hard drive in a forensic setting. His view was that NLP may provide some tools, but will not likely provide the level of knowledge desired at its current level of development (Crowston).

Diekema and Crowston point out that the less structured the text, the more difficult it is to extract information. This suggests two approaches. The first is, where possible, apply NLP to structured, or at least semi-structured texts. The extraction of inherently structured data, such as databases and spreadsheets is well-understood and routinely accomplished. But plain text, for the most part, is forensically viewed simply as a "string," or a "bag of words." Examining text files as genres, with a degree of organization, and comprised of semantic units (sentences), which communicate a story, will not solve all of the forensic analysis needs, but may be sufficiently narrow to provide a level of knowledge creation.

Jason Doyle, in his article "Mapping the World of Consumption" Computational Linguistics Analysis of the Google Text Corpus," wrote an interesting paper that while focused on marketing, has some very significant ideas that will bear on our digital forensic problem. The paper utilizes several premises. First is the notion that communications are "produced *by* people and *for* people," and as such, are "behavior." In turn, "behaviors define people and things." Another premise is that "people don't just write: they write *about* something, and to do so they tend to use words that relate to each other by being relevant to the topic in hand"(6).

For this study, Doyle focuses on actors, which he defines as consumers, stores, brands, manufacturers, etc. He then makes the assumption that "if people consistently describe an actor as behaving in certain ways, ... then this profile of actor-specific behaviors can be used to reveal the world's implicit view of that actor" (3). In digital forensics, while we may be interested in an external view of our investigative subjects or the subject's view of external parties, our primary goal for this research is to find out what our subject is thinking and doing. Doyle's assumption can be reframed; an actor's behavior, as demonstrated by his writing, can provide an actor-specific profile of the actor.

Doyle then takes these fairly intuitive principles and leverages them utilizing an interesting approach. He posits that the co-occurrence of verbs "with each actor describes the behavior of that actor." He furthers this approach by proposing to look at subject-verb occurrences to look at a subject's behavior and subject-object pairs which describe what "is done to the actors". He does this analysis by computing 2-grams for subject-verb pairs and verb-object and using both first and second order co-occurrence (6-9). The former is where the words co-occur within a given context, and the second is where the words occur with similar words in other contexts (Edmonds). Consider the following two sentences:

1. Dick ate a peanut butter and jelly sandwich for lunch.
2. Jane had a peanut butter and banana sandwich for supper.

In the example above, the words *peanut*, *butter*, and *sandwich* share a first order co-occurrence between the two sentences. *Banana* and *jelly* are considered a second order co-occurrence, within a 2-gram of the word *sandwich* and within a 3-gram of the word *butter*. As a result, we can say that bananas and jelly are something that goes on a sandwich and with peanut butter.

Liu, et. al. provide a good synopsis of the extensive NLP work that has been done in the biomedical field. As they describe it, most of the work has been done with a focus on fitting the text into an ontological framework. Instead of using an ontology, they suggest using an “ontology-independent semantic relatedness measure that uses second order co-occurrence vectors.” They argue that strict hierarchical structures do not evolve with changes in medical knowledge, while second-order co-occurrences provide a way for knowledge to evolve with additional texts (363-4). Likewise, Doyle argues that second order co-occurrence, “using 2-grams the method of choice” (8). Despite having said this, he makes another observation that: “In general, the smaller and more focused the context, the better first order co-occurrence performs in computing similarity... because if the context is too wide, the topic may have changed” (7).

This suggests that in the context of a forensic analysis, the use of second order co-occurrence may be useful in genres of text that are large in size, closely related, or have a known connection. In the case of small parcels of text, first order co-occurrence may be a more suitable approach. A third approach might be to leverage both, by examining first order co-occurrences within documents and second order co-occurrences in collections of documents.

I also want to credit Rusu, et al., for their paper “Triplet Extraction from Sentences”. While I did not find this paper until late in my research, and much of my approach was already focused on the use of nouns and verbs, their methodology and inclusion of the predicate phrase provided me with some useful insights.

The research described in this chapter strongly suggests that by leveraging NLP tools and using XML as a repository, it may be possible to analyze the content of at least some of the text found in digital evidence. Todorov suggests that the *subject* of a story is a noun and the *predicate* is a verb. Doyle suggests the tuple of *subjects* and their *objects* (which are equivalent to Todorov’s *predicate*) describe *behavior*. Behavior can be equated to Abbott and Bals’ notion of *events/actions*. Specifically, by looking for noun-verb pairs within sentences, we may approximate the core elements of the narrative fabula as described by Bal.

CHAPTER FOUR: DIGITAL FORENSIC THEORETICAL FRAMEWORK

In this chapter I will lay out two theories that seek to integrate the underlying theories of digital forensics, knowledge management, content analysis, narratology, and natural language processing. The first, which I call the Hermeneutic Theory of Digital Evidence, is an overarching theorem that attempts to define the search for meaning from digital evidence. The second, which I label the Narrative Theory of Digital Evidence, focuses on the use of narrative and natural language processing in the digital forensic process.

Hermeneutic Theory of Digital Evidence

Digital forensics, as we have seen, utilizes two, fundamental approaches: a computer science approach and an investigative approach. Throughout this dissertation, we have seen a tension between these approaches. On one hand, the “traditional” computer science forensic approach can be seen to focus on the technical aspects of the evidence, and seeks to produce reports and testimony that are scientifically defensible. For discussion purposes, I will label this approach as the technical approach.

In the technical approach, the digital forensic examiner deconstructs the physical evidence, consisting of captured and stored digital data, into a series of artifacts. These artifacts include the literal data as well as the operating system, file system, and/or network metadata. The forensic examiner seeks, by use of the forensic examination process, to answer the kind of questions described by Inman and Rudin in their forensic taxonomy: identification, classification/individualization, association, and reconstruction. Forensic examiners produce forensic reports and testimony that can be described as a meta-narrative combining the forensic process undertaken in the current examination, the operation of the technologies associated with the artifacts identified, and the artifacts themselves. These artifacts may be selected because they

are probative in either the case or legal narrative, but the examiner's forensic interest is in their presence and provenance, not in the content *per se*.

In contrast, the investigative approach seeks to discover the people, and the events, that constitute proof of either a crime or tort. They do so, often using a different sub-set of the evidence, to answer different questions. While the technical information and metadata obtained from the forensic examination may be used in the investigative analysis, it is principally the content of the files and fragments that form the majority of the analysis. I do not suggest that the products of the technical examination are not useful, nor commonly used by the analysis. These, in large measure, form the skeleton of the analysis, but it is the content of these artifacts that provides most of the material from which the investigative narratives are created. The person conducting the investigative/analytical role, regardless of whether they are an attorney, investigator, or forensic examiner, seeks to construct a meta-narrative that consists of two main elements: the case narrative and the legal narrative. The former is the narrative constructed by answering the investigative questions of *who, what, when, where, why, and how?* The case narrative takes the answers to the investigative questions, and instantiates them into the elements of the crime or tort, thus forming the legal narrative. The evidence utilized to document these elements are the physical evidence (which includes the digital evidence) and the testimony (including both lay and expert).

Despite any notions to the contrary, these two approaches have been interrelated since the very first digital forensic examination. The relationships have changed as the technology changed, the law adapted, and the processes evolved. In the earliest days of computer forensics, it was the investigative approach that drove the process. As practitioners gained experience, as the courts began to admit digital evidence, and as the technology grew more complex, the

technical approach became more important. In the mid-1990's, technology was advancing rapidly, but the ordinary citizen had little technical knowledge. To prove the reliability of digital evidence, practitioners relied more and more on the technical aspects in order to prove the legitimacy of their proffered evidence. Technical examinations were not, however, answering the investigative questions. In many investigators sought to perform the analysis part of the process and rely on the technicians merely to provide reliable data. As digital evidence datasets grew ever larger, their ability to conduct efficient analyses diminished. With the increased volume of evidence, the problem becomes one of knowledge management. The use of technologies, such as natural language processing and XML, may allow for technically-assisted knowledge management. However, these technologies will not, by themselves, provide much assistance. The transformation of data into information, and information into knowledge, is a cognitive process. While tools can help to present data or information to an analyst, they do not, on their own, produce information or knowledge. In order to transform data to information, it is necessary to not only examine the content, but to situate it in an investigative or forensic context. As a result, there is a need to go beyond tools which focus on the technology and apply content-focused methodologies that utilize technologies that assist the analyst to contextualize the data and information.

Table 3: Comparison of Technical and Investigative Approaches

	Technical Approach	Investigative Analysis Approach
Evidence	Artifacts (operating system, file system, and application metadata, plus content)	Content and communications from recovered artifacts
Process	Forensic Examination	Investigative Analysis
Context	Information technology system	Case and legal context
Answers	Forensic Questions	Investigative questions
Explicative Approach	Meta-narrative of forensic process, technology, and evidentiary artifacts.	Meta-narrative of case narrative and legal elements

Collectively, these two approaches can be viewed as an over-arching theoretical construct for digital forensics. Digital forensics cannot exist without both of these approaches, and thus they represent the core paradigm for the science. The key elements of these approaches can be depicted as shown in Table 3. This suggests a formal paradigm, or theory, for digital forensics. I would suggest the following:

Hermeneutic Theory of Digital Evidence:

1. The legal system utilizes data stored or transmitted in digital form as evidence.
2. The digital evidence must meet the reliability and authenticity tests required by the legal system.
3. Forensic examinations preserve, by use of technical means, the integrity of the original evidence.
4. The products of the examination process consist of artifacts, which can be subdivided into metadata and content.
5. The products of the digital forensic process are utilized to answer two, inter-related sets of questions: the forensic questions and the investigative questions.
6. The digital forensic examination process focuses on answering the forensic questions.
7. The digital forensic analysis process utilizes the products from the digital forensic examination process and generally focuses on answering the investigative questions.
8. The examination and analysis phases of the digital forensic process are both knowledge management processes. These processes should add value to the data and information developed during each phase.

9. The result of the forensic and investigative processing of evidence results in one or more meta-narratives, which are based on a combination of technical artifacts, meta-data, and content. These results form a synthesis of both the forensic and investigative processes.
10. The goal of the digital forensic process is to provide knowledge, in the form of actionable intelligence, investigative leads, testimony, and/or probative evidence.

The Narrative Theory of Digital Forensics

The last part of the theorem states, in effect, that all digital evidence is, at one or more levels, narrative. These narratives contribute to, and are parts of, both the technical and the investigative meta-narratives. It also states that, in order to create these meta-narratives, content is a key element of the examination and analysis process. Implicit in this notion is that content contains a narrative, contributes to a narrative, or is, in and of itself, narrative.

Since narrative is so important to the analysis of the evidence, how should we seek this crucial element? One potential solution is to utilize the core concepts of narratology. Despite a difference in terms, both Bal and Abbott agree that there are effectively three elements which define a narrative. Using Bal's terminology, these are: chronology/logic, event/action, and actors.

As Waal, Benter and Barnard; Toderov; and Doyle all suggest, it may be possible to search for narratives by utilizing an approach which attempts to identify narratives by utilizing the semantics of the content. Because this is primarily a semantic analysis, the use of sentences as a basic unit of analysis is appropriate.

As the goal of this semantic analysis is to develop knowledge of subjects/actors, actions/events, and chronology, the application of natural language processing should focus on the identification of, and relationships among, these elements. Since nouns and verbs can be used

as rough proxies for the subjects/actors and actions/events respectively, techniques which focus on these grammatical structures will likely improve the identification, efficiency, and comprehension of the narrative.

Beyond the identification of narratives, it is essential, given the vast quantities of digital evidence, to be able to identify particular narratives which are probative. At present, there do not seem to be technologies that can, by themselves, accurately evaluate the probative value of any given narrative. However, the ability of human analysts to evaluate data is limited by things like attention, fatigue, and pre-conceptions about the data. It would therefore be useful to find ways which simultaneously reduce the volume of data that needs to be reviewed by the analyst and enhance the ability of the analyst to identify the probative data. Reading a half million emails is not practical. Reducing that number, while increasing the likelihood that what is read is probative, should be a goal of knowledge management in the digital forensic context.

These aspects of the analysis of digital evidence suggest two hypotheses:

Hypothesis One

The identification of narratives, by automated means, can contribute to the efficiency and effectiveness of the forensic examiner/investigative analyst.

A second, dependent hypothesis might be:

Hypothesis Two

Any automated process that improves the ability of the forensic examiner/investigative analyst to quickly identify probative narratives will improve the efficiency and effectiveness of the process.

Hypothesis Three

The use of nouns and verbs will assist in the identification of both general and probative narratives more economically than reading complete texts.

Hypothesis Four

Natural language processing software can assist in the identification of probative narratives by use of lexical, grammatical, and semantic techniques.

CHAPTER FIVE: APPLYING THE THEORY TO DIGITAL EVIDENCE

Experimental Design

These experiments are preliminary and exploratory. They are not designed to be definitive, but merely explore the possible validity of the theories described in the previous chapter. Given the very general nature of the hermeneutic theory posited, I will focus on the second, narrative theorem, and in particular, on Hypotheses Three and Four. There is no doubt that even if these initial experiments appear relevant, much additional research will be needed to verify and validate these hypotheses.

Digital forensics, like natural language processing, deals with human communications “in the wild,” and therefore, it combines both structured and unstructured data. And while it would be desirable to develop a finished forensic software tool, while simultaneously proving the narrative theory on un-structured data, this is not realistic. Conversely, the ability to generalize a solution developed on carefully crafted “artificial” data, would not be terribly compelling. As a result, I have chosen to experiment, utilizing a semi-structured set of naturally occurring text — emails from the Enron corpus.

Earlier, I described the Natural Language Tool Kit (NLTK). This is a set of computer instructions, written in the scripting language Python, organized into a collection of routines, called libraries. I have utilized the native Python language, the libraries available through NLTK, and some code that I wrote, to process non-arbitrary files (emails) from the Enron corpus. It is important to note that I am not a programmer, nor am I an expert in either Python or NLTK. The code developed for this dissertation is functional, but is neither elegant nor efficient. There is no doubt that an expert programmer could duplicate my results with many fewer lines of code and process the text in far less time.

Earlier, I described how XML has great potential to represent the various elements of a digital forensic examination, as well as the content of the evidentiary texts. For the following experiments, I will utilize XML as a repository for the processed text. As the purpose of these explorations is not the development of a digital forensic software application, by merely extracting and presenting the data utilizing XML, I will not realize the full, rhetorical potential of XML. Clearly, this would be a fruitful extension of this work, but is beyond the scope of this dissertation.

The focus of these experiments is to explore the validity of the proposed narrative theory of digital forensics, specifically the identification of the three elements of the fabula: subject/actor, actions/events, and chronology/logic, by use of natural language programming.

The Enron Corpora

As previously mentioned, the criminal case involving Enron Corporation resulted in the public availability of part of the voluminous emails from the company. The decision to utilize these for experimental purposes was the result of a compromise. On one hand, these emails represent human communication “in the wild,” albeit in a particular cultural context. Much of the analysis of the Enron scandal focuses on the cultural environment, which facilitated the behaviors that were subsequently determined to be criminal. For an excellent insight into the complex cultural and criminal aspects of this massive case, I would suggest McLean and Elkind’s book, *The Smartest Guys in the Room*. As this book demonstrates, emails are an important part of the evidence utilized in complex cases. One advantage to using emails is that they have built-in metadata, in the form of the headers. These headers record, in addition to the sender and receiver, the source, destination, path, and other useful information about the email. The third advantage of utilizing emails for research purposes is that they form a genre of text,

one that has common characteristics and usages. In the case of the Enron corpora, they are also represented exclusively in ASCII characters. This makes the textual analysis straightforward, in comparison to having to interpret embedded formatting information.

On the other hand, it can be argued, for all these same reasons, that an email corpus is not representative of much of the non-email data found in items of digital evidence. While that is certainly true, the purpose of this experimentation is to identify potentially useful approaches, not to find a singular, “silver bullet” for digital forensics.

The term, Enron corpora, refers to several different corpora, all of which originate with their release by the Federal Energy Regulatory Commission in 2003. There are several versions which have been processed in a variety of ways. The initial, raw email files were purchased by Leslie Pack Kaelbling at MIT, and the emails were further processed by Melinda Gervasio and others at SRI to eliminate duplicates, attachments, and much of the spam. Further “cleaning” of the dataset was conducted by William Cohen and his colleagues at Carnegie Mellon University. The corpus used for the experiments in this dissertation is available from their [website](#). This corpus contains approximately 500,000 emails, organized into a folder for each user, each of which contains multiple folders for *sent*, *received*, *draft*, and other software and user-created topics (Cohen, William).

There have been a number of research projects that, utilizing the Carnegie Mellon (CMU) corpus, created additional corpora (Cohen, William). Ted Pedersen, from the University of Minnesota, conducted research which classified the content of the CMU corpus and is available as a separate corpus, called [Enron Email Corpus by Topic](#), with files listing the classifications obtained by his software (Pedersen). I reviewed this corpus and the topics were so broadly identified, such as business plans, schedules, legal matters, and general, that the classifications

did not appear to have sufficient granularity to assist in the identification of narratives, and I chose not to utilize this corpus. Another corpus, the [Enron Sent Corpus](#), was produced by the University of Colorado. This corpus has extracted all of the textual sentences found in the bodies of the emails (Styler). There is an online, searchable version of the corpus at <http://www.enron-mail.com>.

Initially, I will utilize selections from the CMU corpora to demonstrate the application of several NLP techniques. A sub-set of the emails will be processed, using the most useful NLP techniques, and the results will be stored in an XML file for further analysis.

Description of Experiments

The exploration of both the narrative theory and the use of NLP will be done through a series of software routines. The combination of Python and NLP tools will be utilized to extract the textual elements and write the data to an XML file. In the following sections, I will describe the experiments and the results.

For the initial testing I will use one particular email as an example. This email, identified as number 443, was arbitrarily extracted from the Inbox of Jason Wolfe, an Enron employee, and is part of the publically available Enron Corpus. The email, dated November 27, 2001, is a company-wide announcement of a merger between Enron and another company, Dynegy. The full text of this email is reproduced in Appendix A.

Email Header Extraction

The Enron emails are provided in an entirely ASCII text format and include the header. The email header contains the sender, receiver(s), date, time, subject, and some routing information. As mentioned earlier, this is very useful metadata, and as such, we wish to capture this data. The header consists of both internal and external metadata. The *sender*, *receiver(s)*, and

subject are selected by the creator of the message and embedded by the email application. The dates, times, and initial routing information are created by the originating computer. Additional routing information is appended to the header by all of the subsequent devices which route (forward) the email to its destination. In the case of the CMU corpus, the complete routing information is not present. Any entry in the email header which begins with an “X,” is non-standard, and is ignored by the routers. These entries are commonly a product of the email client and/or originating email system (University of Illinois). For purposes of these experiments, the “X-headers” will not be used. The remainder of the email is the textual “payload,” and is the product of the email’s author. Effectively, an email consists of two parts: header information and the body or payload as shown in Figure 17.

Email Header

Message-ID: <11730746.1075862738340.JavaMail.evans@thyme>
Date: Tue, 27 Nov 2001 18:47:00 -0800 (PST)
From: announcements.enron@enron.com
To: dl-ga-all_enron_worldwide1@enron.com
Subject: Enron/Dynegy Merger; Antitrust Issues
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Enron General Announcements
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=MBX_ANNCENRON>
X-To: DL-GA-all_enron_worldwide1 </O=ENRON/OU=NA/CN=RECIPIENTS/CN=DL-GA-all_enron_worldwide1>
X-cc:
X-bcc:
X-Folder: \JWOLFE (Non-Privileged)\Wolfe, Jason\Inbox
X-Origin: Wolfe-J
X-FileName: JWOLFE (Non-Privileged).pst

Email Body

As you know, Enron has signed a merger agreement by which Dynegy will acquire Enron. We expect the transaction to close following shareholder and regulatory approvals and various conditions to closing.

Even though Enron has entered into this agreement, U.S. and foreign antitrust laws require that Enron and Dynegy continue to operate independently of each other. In particular, to the extent that Enron and Dynegy are competitors in various businesses or markets, their respective activities must be undertaken at arm's length until the transaction has closed. Therefore, for antitrust purposes you should treat Dynegy as you would any other unaffiliated company notwithstanding the merger agreement.

In addition, all information, documents and communications related to the merger between Enron and Dynegy should be coordinated through and approved by Mark Muller, Lance Schuler, Robert Eickenroht, Mark Haedicke, Rob Walls or Greg Whalley of Enron. It is absolutely critical that this procedure be maintained. To the extent that information is required to be disclosed to Dynegy under the merger agreement, then such disclosure should be approved by one of the foregoing individuals.

If you have any questions concerning this notice, please contact Lance Schuler (713/853-5419), Robert Eickenroht (713/853-3155), Mark Haedicke (713/853-6544) or Rob Walls (713/646-6017). Thank you for your help in this matter.

Figure 17: Email Header and Body

As a result, the first phase in processing the emails is to separate the header from the body, capture the metadata from the header, and to isolate the email body for further analysis. Using a Python script, the email text file (original file) is loaded into the computer's memory, as

a string (an ordered series of characters), and thereafter read as individual lines. This is accomplished by separating the string (tokenizing) on carriage return/newline characters. These are the characters which tell the computer to display the data on a new line. This approach is utilized because header information is stored on individual lines. Once separated into lines, the program then looks for the header labels, such as: *From*, *To*, *Subject* at the beginning of each line. It then writes each item to the output XML file, embedding the data in the appropriate XML tags, as shown in Figure 18.

```
<?xml version="1.0" encoding="UTF-8"?>
- <Email file="c:/Python27/443.txt">
  - <Header>
    <Message-ID>1730746.1075862738340.JavaMail.evans@thyme</Message-ID>
    <Date>Tue, 27 Nov 2001 18:47:00 -0800 (PST)</Date>
    <From>announcements.enron@enron.com</From>
    <To>dl-ga-all_enron_worldwide1@enron.com</To>
    <Subject>Enron/Dynegy Merger; Antitrust Issues</Subject>
  </Header>
</Email>
```

Figure 18: Extracted Header in XML

The software continues, line by line, to extract the desired header information, until the end of the header is reached. This is identified by locating the *X-FileName* tag. It should be noted that only the *Message-ID*, *Date*, *From*, *To*, and *Subject* tags are extracted. This was done in an effort to select the most useful information, while minimizing the volume of data. All data, past the *X-FileName* tag, is stored as a string of characters, in an object labeled “body.”

I need to mention that emails often contain text “quoted” from pervious emails. As described by Styler in the “EnronSent Corpus Report”, this can present a problem from an analytical perspective (2). By having repetitive copies of the same text, any statistical analysis is skewed toward the content of the quoted text. Further, just from a volume and efficiency perspective, it adds volume to the analyzed text, without adding any additional value. If the

quoted material is either clearly marked as quoted, or it contains the header information, it is fairly straightforward to recognize that the material has a different source. This problem was identified by the compilers of the Enron corpora, and some efforts were made to eliminate some of this type of material, but some quoted material remains.

Separating quoted material from the content of the email is more difficult than might initially appear. Different email clients, such as Microsoft Outlook, Mozilla Thunderbird, and other, format quoted text differently. Users will often “cut and paste” parts of other emails, in an arbitrary manner, into the current email. Some email is composed and transmitted in HTML and later converted to ASCII text.

The value of the header information, in terms of setting a temporal and discourse context is obvious. There are a number of studies, using the Enron corpora, which identify network relationships, discourse, and organizational structure. Email header information is critical for these sorts of studies. One header tag that deserves discussion, in the particular context of this dissertation, is that of the *Subject*. In many cases, for example our 443 test email, the subject is both descriptive and accurate in characterizing the content of the email. Observation and reflection will remind us that this is often not the case. In some cases, authors of emails will fail to type anything in the *subject* line. Often, the *subject* line, especially in emails that are second or subsequent responses, will indicate a topic that is not directly addressed in the immediate email. This is also seen in forwarded emails, which often contain information that is tangential, or even unrelated, to the *subject* line. As a result, the analytical value of the subject line is highly variable and therefore, unreliable for characterizing content.

Processing the Email Body

Once the email body has been separated from the message file, the text must be processed, through a series of progressively more granular steps, to attempt to identify the narrative elements of the fabula. The processing proceeds through several phases, including sentence tokenizing, part of speech tagging, grammatical analysis, and summarization.

Sentence and Part of Speech Processing

Once the header information has been parsed and tagged, we begin the processing of the email body. Since one of the goals of the analysis is to identify the content of the document, it is necessary to disassemble the text into its component parts. Writers are commonly taught to use sentences, organized into paragraphs, to organize individual, complete thoughts. As such, identifying and parsing paragraphs, sentence structures, and parts of speech are necessary. Fortunately, NLTK has a number of tools which will allow us to do this.

The first step in processing the body is to identify paragraphs. This was done by looking for blank lines. Each group of text, contained between two blank lines, was treated as a single paragraph. The purpose of parsing the text into paragraphs was the notion that, when properly educated, we will encapsulate individual concepts into paragraphs. While this approach worked very well in the test message, providing good groupings of related information, it proved very troublesome when applied to a large quantity of emails. It turns out that not only do many people ignore the compositional best practices, they often write phrases, or sentences, on single lines. In some cases, people write emails in a “bulleted” fashion, one phrase per line, with no punctuation. This results in creating a large number of XML structures that do not add to the textual analysis. Another issue was the need to eliminate multiple blank lines, which were being tagged as paragraphs. While this later issue was eliminated by testing for the presence of content in the

structure, the former problem was more difficult. As a result, in later experiments, the tokenizing of paragraphs was abandoned, and sentences were tokenized directly. The output, shown in Figure 19, is from an early version of the software, and demonstrates these issues.

Once the paragraphs are identified, we parse the paragraphs into sentences. The standard sentence parser from the NLTK libraries, *sent_tokenizer*, was utilized. The individual sentences are then parsed, using the the Punkt Word Tokenizer, in order to create a list of words in each sentence. The tokenized sentences are, in turn, tagged for parts of speech, utilizing the NLTK *pos_tag* tagger. These tags are a defined set of abbreviations that represent nouns, verbs, and all the other parts of speech (See Table 2). The result is a list of tuples, containing the word and part of speech tag. The results of this process, is stored in the XML data file, as shown in Figure 19.

```
<?xml version="1.0" encoding="UTF-8"?>
- <Email file="c:/Python27/443.txt">
- <Body>
  - <Paragraph>
    <Para_num>1</Para_num>
  </Paragraph>
  - <Paragraph>
    <Para_num>2</Para_num>
    - <Sentence>
      <Sent_num>1</Sent_num>
      <Content>As you know, Enron has signed a merger agreement by which Dynegy will acquire Enron.</Content>
      <POS_Tags>[('As', 'IN'), ('you', 'PRP'), ('know', 'VBP'), (',', ','), ('Enron', 'NNP'), ('has', 'VBZ'), ('signed', 'VBN'), ('a', 'DT'), ('merger', 'NN'), ('agreement', 'NN'), ('by', 'IN'), ('which', 'WDT'), ('Dynegy', 'NNP'), ('will', 'MD'), ('acquire', 'VB'), ('Enron.', 'NNP')]</POS_Tags>
    </Sentence>
  </Paragraph>
```

Figure 19: Example of Parsed Sentence in XML

The part of speech tagger selected is a standard version included in the NLTK library. An even higher level of accuracy, in tagging words, may be possible utilizing a tagger trained with a training set. However, the purpose of this experiment is demonstrative and no attempts to optimize the software have been undertaken.

Grammatical Analysis

Once the parts of speech are tagged, it is possible to evaluate the words in order to identify sentence structure. In order to understand the meaning of a sentence, we must understand its grammar. Jurafsky and Martin lay out the key concepts of grammar, which they describe as comprising of three elements: constituency, grammatical relations and subcategorization (sic), and dependency. Constituency refers to the notion that words can work together as a group, such as noun or verb phrases. The grammatical relations include constructs, such as *subjects* and *objects*, while subcategorization (sic) and dependency refer to particular “relationships between words and phrases” (Jurafsky and Martin 385-6). One of the most common ways to deal with the grammatical aspect of text is to utilize a computer model, for any given grammar. The most commonly used are known as “context-free grammars,” which, according to Jurafsky and Martin, were formalized by Chomsky in 1956 and Backus in 1959 (387). These grammars evaluate the words, by part of speech and phrase, in an attempt to show the relationships between the words and phrases. This is often accomplished by utilizing NLP software to parse sentences, using a set of rules called a “grammar,” to diagram the sentence in a similar fashion to the way in which generations of students learned grammar, by diagramming sentences. The product of the software is often displayed in a “tree” structure, as shown in Figure 20. The labels *NP*, *VP*, and *PP* represent a noun phrase, a verb phrase, and a prepositional phrase respectively.

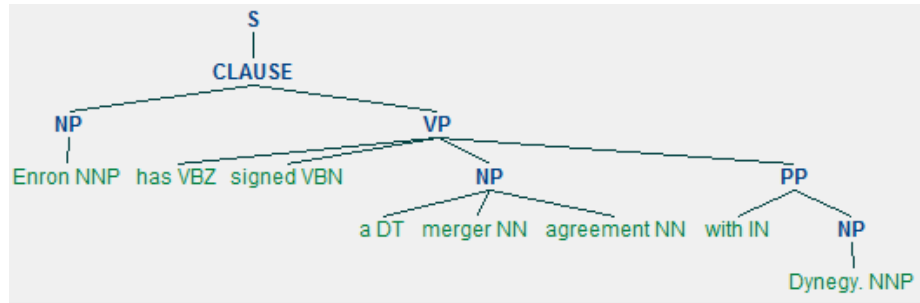


Figure 20 NLTK Parse Tree

Software is utilized to process the text, in order to make the grammatical structures, such as noun or verb phrases, that Bird, Klein, and Loper label “entities,” visible, utilize a number of techniques including “chunking” and “chinking.” The former, decides which words combine to form the structures defined in the grammar, while the latter removes words that are determined, by the grammar, to be outside of the defined grammar (263-9). The accuracy of the grammatical evaluation is dependent on the rules which define the model’s grammar. It is possible to manually create a set of rules, which form the grammar, which is used by the “chunker” software. It is also possible to utilize previously analyzed corpora to “train” the chunking software. Publically available tagged corpora, such as the CoNLL-2000 Chunking Corpus, allow us to both train our software and evaluate our grammars. The use of training allows for a much higher accuracy in determining the sentence entities. Examples in Perkin’s book, show accuracy from relatively simple grammars, to be in the 30-60 percent range, while the use of training corpora can result in accuracies in the 90-97% range (124-32). By utilizing even more complex (and computationally demanding) techniques, the accuracy can reach the 98-99 percent range. It should be noted that these results are achieved by evaluating part of the same corpus as training data, and another part as test data. As a result, the data is very similar. Results achieved by training against a known dataset, and then parsing a different set of data, such as that found in

digital forensic cases, is likely to be much less accurate. Similarly, the ability to “train” the grammars against the same forensic dataset as it is being analyzed is problematic. It is difficult to identify, decide upon, and separate training and analytic datasets from the original evidence. Another issue is that forensic datasets are often wildly disparate, internally, in format and content.

While the use of these advanced techniques for “understanding” text in the NLP context is well-known, and often utilized, their use in digital forensics may be counterproductive. These techniques require substantial computational processing, and usually result in data structures that are comparable in size to the original text. In order to efficiently transform, in the knowledge management sense, the large volume of data into more concentrated, and valuable information, we need to think about reducing the data to be reviewed. This returns us to Hypothesis Three: The use of nouns and verbs will more effectively and efficiently facilitate the identification of both general and probative narratives, than reading complete texts.

One approach would be to extract all nouns and verbs from a given text. The “443” test email is comprised of 244 words, of which, 78 are nouns and 33 are verbs. Even when eliminating duplicates, 21 nouns and 16 verbs remain. As Figure 21 shows, the results of this approach, which is akin to the “bag of words” approach, does not provide much assistance in understanding the content of the email. It should be noted that this particular process extracted all of the words in the noun class. If we were to limit our selection of nouns, to personal nouns, we reduce the number to 128 entries, which is still too many to be useful. It is relatively simple to further process the nouns and verbs, to eliminate duplicates, resulting in “sets” of unique personal nouns and verbs. This results in 21 and 16 entities, respectively, as shown in Figure 22.

Nouns:

questions,notice,please,contact,Lance,Schuler,Robert,Eickenroht,Mark,Haedicke,Rob,Walls,procedure,maintained.,In,addition,information,documents,communications,merger,Enron,Dynegy,Mark,Muller,Lance,Schuler,Robert,Eickenroht,Mark,Haedicke,Rob,Walls,Greg,Whalley,Enron.,In,extent,Enron,Dynegy,competitors,businesses,markets,activities,arm,length,transaction,closed.,Even,Enron,agreement,U.S.,laws,Enron,Dynegy,continue,other.,As,Enron,merger,agreement,Dynegy,Enron.,procedure,maintained.,In,addition,information,documents,communications,merger,Enron,Dynegy,Mark,Muller,Lance,Schuler,Robert,Eickenroht,Mark,Haedicke,Rob,Walls,Greg,Whalley,Enron.,In,extent,Enron,Dynegy,competitors,businesses,markets,activities,arm,length,transaction,closed.,Even,Enron,agreement,U.S.,laws,Enron,Dynegy,continue,other.,As,Enron,merger,agreement,Dynegy,Enron.,addition,information,documents,communications,merger,Enron,Dynegy,Mark,Muller,Lance,Schuler,Robert,Eickenroht,Mark,Haedicke,Rob,Walls,Greg,Whalley,Enron.,In,extent,Enron,Dynegy,competitors,businesses,markets,activities,arm,length,transaction,closed.,Even,Enron,agreement,U.S.,laws,Enron,Dynegy,continue,other.,As,Enron,merger,agreement,Dynegy,Enron.,Enron,agreement,U.S.,laws,Enron,Dynegy,continue,other.,As,Enron,merger,agreement,Dynegy,Enron.,Enron,merger,agreement,Dynegy,Enron.

Verbs:

have,concerning,is,be,related,be,coordinated,approved,are,be,undertaken,has,has,entered,require,operate,know,has,signed,acquire,is,be,related,be,coordinated,approved,are,be,undertaken,has,has,entered,require,operate,know,has,signed,acquire,related,be,coordinated,approved,are,be,undertaken,has,has,entered,require,operate,know,has,signed,acquire,are,be,undertaken,has,has,entered,require,operate,know,has,signed,acquire,has,entered,require,operate,know,has,signed,acquire,know,has,signed,acquire

```
[(';', 40), ('.', 1), ('CC', 31), ('CD', 8), ('DT', 34), ('IN', 60), ('JJ', 24), ('MD', 13), ('NN', 52), ('NNP', 128), ('NNS', 28), ('POS', 4), ('PRP', 9), ('PRPS', 4), ('RB', 8), ('TO', 12), ('VB', 20), ('VBD', 3), ('VBG', 1), ('VBN', 21), ('VBP', 16), ('VBZ', 17), ('WDT', 6)]
```

```
>>> len(pos_tags) 540
```

Figure 21: Email 443 Nouns and Verbs

```

>>> set(nouns)
set(['As',
'Dynegy',
'Eickenroht',
'Enron',
'Enron.',
'Even',
'Greg',
'Haedicke',
'In',
'Lance',
'Mark',
'Muller',
'Rob',
'Robert',
'Schuler',
'U.S.',
'Walls',
'Whalley',
'closed.',
'maintained.',
'other.'])
>>> len(set(nouns)) 21

>>> set(verbs)
set(['acquire',
'approved',
'are',
'be',
'concerning',
'coordinated',
'entered',
'has',
'have',
'is',
'know',
'operate',
'related',
'require',
'signed',
'undertaken'])
>>> len(set(verbs)) 16

```

Figure 22: Email 443 Noun-Verb Sets

The “set” of unique nouns and verbs, does substantially reduce the amount of material to review and at least subjectively, appears to contain most of the significant elements of the email. At the very least, it is more “dense,” in the sense that there is more information of potentially probative value in a lesser number of words. There are two additional observations of this particular process. First, is that many of the nouns, are proper nouns, i.e., personal or corporate names. This creates a potential problem in that, using this particular transformation, to link first and last names or a compound name for a company, such as *Exxon Mobil Corporation*. Secondly, there are several entries, in the “nouns” list, that are either not nouns, or are apparently duplicated (“Enron”). These issues are a result of the assumptions, made by the part of speech tagger. In the case of the misclassification of the parts of speech, accuracy, beyond a certain

point, requires much more complex processing, using an annotated corpus of similar material. As stated earlier, the purpose of this dissertation is exploratory, and no attempts have been made to optimize the software. Likewise, the apparent duplication of the Enron entry, is as a result of the sentence tokenizer including the period punctuating the sentence, into the token (word). As a result, the software thinks these are different words – one with a period and the other without.

Named Entity Extraction

Fortunately, there are additional tools that will allow us to refine our selection of proper nouns. Natural language processing has developed a notion, called “named entities,” which are used to label specific types of information, such as, persons (PERSON), organizations (ORGANIZATION), locations (LOCATION), and geo-political entities, such as cities, states, or countries (GPE). Named entities are identified through the use of a type of software called a “classifier.” NLTK has a previously trained classifier, for named entities, called *nltk.ne_chunk* (Bird, Klein and Loper 281-3).

Applying the NLTK named entity classifier to our test email yields a number of PERSON and GPE labeled entries. The resulting list is shown in Figure 23. The results of this extraction can be further refined, by creating a “set” of unique entities, and is shown in Figure 24. By set, we mean a list of terms, where all duplicates are removed. Note that Enron and Dynegy are both labeled as persons, not as organizations. Since it is likely that the training set utilized to train this classifier did not include either of those two companies, the software was, nonetheless, able to identify those words as proper nouns, and therefore, classify them as persons. It is also interesting to note that the classifier also labeled the word *Therefore* as a GPE. Again, this is likely a result of the particular training set. In a mature digital forensic NLP application, the tokenizers, parsers, chunkers and classifiers might be able to be trained utilizing sub-sets of the

evidentiary files. Additionally, application metadata, such as the email headers, could be utilized to train named entity classifiers.

The extraction of named entities appears to have substantial value in our forensic analysis. In traditional methodologies, searches for names and places could be conducted using string searches, but only if the search term was known. Often, the spelling of the term could substantially affect the search result. Further, the resulting “hits” were returned entirely out of context. The extraction of named entities reduces many of these problems. Since, for the moment, our goal is to identify the subjects/actors of the fabula, these persons and organizations identified can be utilized as subjects/actors for further analysis.

In a similar fashion, NLTK has libraries which identify time and date-related strings of data (*dateutil*) and for temporal relationships within text (*timex*) (Perkins 227-233). Several attempts to integrate them with the current process met with little success. No additional experiments were conducted with these tools. With additional coding processing, it might be possible to integrate these tools into the extraction of fabula elements.

```
(PERSON Enron/NNP)
(PERSON Dynegy/NNP)
(PERSON Enron/NNP)
(GPE U. S./NNP)
(PERSON Enron/NNP)
(GPE Dynegy/NNP)
(PERSON Enron/NNP)
(GPE Dynegy/NNP)
(GPE Therefore/NNP)
(PERSON Dynegy/NNP)
(PERSON Enron/NNP)
(PERSON Dynegy/NNP)
(PERSON Mark/NNP Muller/NNP)
(PERSON Lance/NNP Schuler/NNP)
(PERSON Robert/NNP Eickenroht/NNP)
(PERSON Mark/NNP Haedicke/NNP)
(PERSON Rob/NNP walls/NNP)
(PERSON Greg/NNP whalley/NNP)
(GPE Dynegy/NNP)
(PERSON Lance/NNP Schuler/NNP)
(PERSON Robert/NNP Eickenroht/NNP)
(PERSON Mark/NNP Haedicke/NNP)
(PERSON Rob/NNP walls/NNP)
```

Figure 23: Email 443 Named Entities List

```
set(['(GPE Dynegy/NNP)',
'(GPE Therefore/NNP)',
'(GPE U. S./NNP)',
'(PERSON Dynegy/NNP)',
'(PERSON Enron/NNP)',
'(PERSON Greg/NNP whalley/NNP)',
'(PERSON Lance/NNP Schuler/NNP)',
'(PERSON Mark/NNP Haedicke/NNP)',
'(PERSON Mark/NNP Muller/NNP)',
'(PERSON Rob/NNP walls/NNP)',
'(PERSON Robert/NNP Eickenroht/NNP)'])
```

Figure 24: Email 443 Named Entities Set

A set of tags, <Named_Entities>, was included in the email processing algorithm. The named entities are extracted and stored at the beginning of the email body section of the XML document as shown in Appendix B. An additional software tool was developed, which would process an entire folder of email, extracting the named entities from each email, writing them to XML, and then collecting all of them into a single set, with all the duplicates removed. The output of this program can be seen in Figure 25. An analysis of this process will be described in a later section.


```

<?xml version="1.0" encoding="UTF-8"?>
- <Data>
- <Email>
- <Header>
- <Filename>f:\Test2\2</Filename>
</Header>
- <Body>
<Named_Entities>set(['Delainey', 'McGowan', 'NE', 'MW', 'SO2', 'NOX', 'California', 'George', 'Genelle'])</Named_Entities>
</Body>
</Email>
- <Email>
- <Header>
- <Filename>f:\Test2\1</Filename>
</Header>
- <Body>
<Named_Entities>set(['Paul', 'Garrett', 'British Energy', 'Dana'])</Named_Entities>
</Body>
</Email>
- <Email>
- <Header>
- <Filename>f:\Test2\9</Filename>
</Header>
- <Body>
<Named_Entities>set(['Ontario', 'Wilson', 'ECT Subject', 'Power', 'Little League', 'Canadian', 'Robert', 'Municipal Affairs', 'Mr. Sutton', 'Clement', 'JoeMy',
'marketsJohn', 'Houston', 'Jim Wilson', 'Alberta', 'Tony Clement', 'John', 'Energy', 'Father Jim Wilson'])</Named_Entities>
</Body>
</Email>
<Collected_NE>set(['Wilson', 'Ontario', 'British Energy', 'Robert', 'Mr. Sutton', 'Alberta', 'John', 'Genelle', 'ECT Subject', 'Power', 'NE', 'SO2', 'Municipal Affairs',
'California', 'Tony Clement', 'Father Jim Wilson', 'Garrett', 'Canadian', 'McGowan', 'Delainey', 'Jim Wilson', 'Dana', 'Houston', 'Little League', 'Energy',
'Clement', 'JoeMy', 'MW', 'marketsJohn', 'NOX', 'Paul', 'George'])</Collected_NE>
</Data>

```

Figure 25: Named Entity Extraction and Collection

Semantic Analysis

Now that we have identified some of the subjects/actors, the next logical step is to identify the actions/events. As previously discussed, we will utilize verbs as proxies for these narrative constructs. While we have already compiled a list of verbs, it is necessary to link them to the nouns. As many NLP authors discuss, one of the challenges to understanding English grammar is its complexity. As we saw, with a very simple example of a tree (Figure 20), there can be numerous noun and verb phrases in a single sentence. Furthermore, they may be constructed in several different orders, often with similar meanings. Rather than trying to understand the detailed grammar, perhaps merely linking the verbs to the nouns will allow a “surrealist” reading of the text?

Linking verbs to the proper nouns can be done in a number of ways. The verbs can be listed in the order in which they appear in the sentence (syntactically) or in alphabetic order after the noun. The former preserves the positional relationship between the noun and verb phrases, possibly assisting in understanding the intended meaning. However, the result does not substantially decrease the volume of the processed text, nor does it greatly improve the “density”

of information provided to the analyst. Sorting the verbs alphabetically, and/or removing the duplicates, assists in reducing the volume, but may not improve the comprehension of the material. Some examples from the results of this experiment are shown in Figure 26.

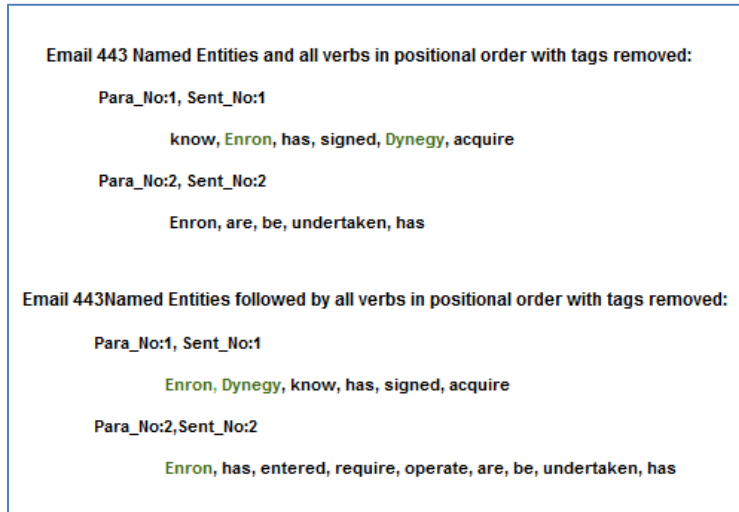


Figure 26: Email 443 Named Entities and Verbs

While these results are certainly “economical,” with respect to the volume, they lack apparent significance. What are missing are the objects of the sentences, which are other nouns. In the first example in Figure 26, the noun phrase “merger agreement,” is significant to the understanding of this sentence. If we modify our program to extract all of the nouns and verbs, without any of the other words, we get results as shown in Figure 27. The result is somewhat readable, contains only 78 words (from the original 244), and appears to provide some significant information.

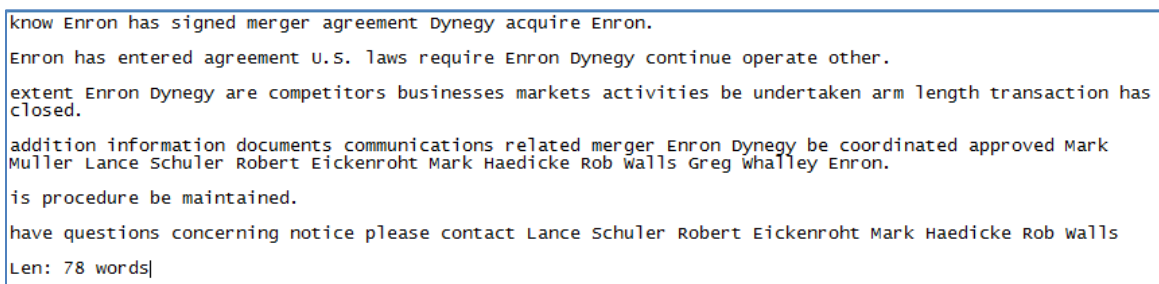


Figure 27: Email 443 Extracted Nouns and Verbs

Merely reading the parsed first sentence, it is clear that this email has something to do with the merger between Enron and Dynegy. This certainly provides much more information density, that is, the combination of significance, or probative value, with a minimum of textual volume. And while this is a reduction of over 68%, extended over perhaps hundreds of email, it is still a substantial analytic burden.

The use of statistical approaches to NLP is well established. One of the simplest techniques is to calculate a frequency distribution of words. Because many short words, such as *a*, *an*, *the*, and *for* are very common, they often must be eliminated prior to calculating a frequency distribution. To do otherwise would skew the distribution away from potentially more probative words. It is common to utilize a list of insignificant words, called a “stop list,” to be removed prior to calculating the frequency distribution. Usually, this is done with a large file of “stop words,” each of which has to be tested against every word in the analyzed text. Since, as can be seen in Figure 27, we have eliminated all words that are not nouns and verbs and substantially reduced the need for a “stop list.”

In the next experiment, I will calculate separate distributions for both nouns and verbs in the test email. The distribution is then displayed as a list of words, in order of decreasing frequency, as shown in Figures 28 and 29.

```

>>> fdistn.items()
[('Enron', 7),
 ('Dynegy', 6),
 ('agreement', 4),
 ('merger', 4),
 ('Mark', 3),
 ('Eickenroht', 2),
 ('Haedicke', 2),
 ('Lance', 2),
 ('Rob', 2),
 ('Robert', 2),
 ('Schuler', 2),
 ('Walls', 2),
 ('extent', 2),
 ('information', 2),
 ('transaction', 2),
 ...

>>> fdistv.items()
[('be', 5),
 ('has', 3),
 ('approved', 2),
 ('is', 2),
 ('acquire', 1),
 ('are', 1),
 ('close', 1),
 ('closed', 1),
 ('concerning', 1),
 ('coordinated', 1),
 ('disclosed', 1),
 ('entered', 1),
 ('expect', 1),
 ('following', 1),
 ('have', 1),
 ...

```

Figure 28: Noun and Verb Frequency Distribution

```

The Five Most Frequent Nouns (decreasing Order)
['Enron', 'Dynegy', 'agreement', 'merger', 'Mark']
()
The Five Most Frequent Verbs (decreasing Order)
['be', 'has', 'approved', 'is', 'acquire']

```

Figure 29: Email 443 Five Most Common Nouns and Verbs

At least based upon our one test email, the most frequent nouns contain the most significant information in the email: Enron has agreed to be merged with Dynegy. It is also noteworthy that it is the proper nouns that are significant. The verbs appear much less significant. It should be noted that the frequencies of the verbs were extremely low, with only a handful having multiple occurrences. Perhaps associating the verbs, with the most significant nouns, will provide additional context.

As an experiment, I wrote a program that searched each of the sentences in the test email. It searched for the existence of any, or all, of the five most common nouns. If it found one or

more of those nouns, it would print them out, indicating the number of the paragraph and the sentence where they were located. It then searched that sentence, to extract each of the verbs, in positional sequence, and printed them out. The results are seen in Figure 30. It does appear that associating the verbs with the nouns, at least in an un-ordered way, contributes to the comprehension. It does decrease the density. Perhaps we can modify this approach and combine it with our extracted nouns and verbs approach.

```
Paragraph No.1: Sentence No. 1
Nouns:Enron, Dynegy, agreement, merger,
Associated verbs: acquire, has, know, signed,

Paragraph No.3: Sentence No. 1
Nouns:Enron, Dynegy, agreement,
Associated verbs: entered, has, require, operate,

Paragraph No.3: Sentence No. 2
Nouns:Enron, Dynegy,
Associated verbs: be, has, are, undertaken,

Paragraph No.3: Sentence No. 3
Nouns:Dynegy, merger,
Associated verbs: notwithstanding, treat,

Paragraph No.5: Sentence No. 1
Nouns:Enron, Dynegy, merger, Mark,
Associated verbs: be, approved, related, coordinated,

Paragraph No.5: Sentence No. 3
Nouns:Dynegy, agreement, merger,
Associated verbs: be, is, required, disclosed, approved,

Paragraph No.7: Sentence No. 1
Nouns:Mark,
Associated verbs: concerning, have,
```

Figure 30: Verbs Associated with the Common Nouns

A review of the extracted text from both approaches reveals that there are still a number of “to-be” verbs present. One of the observations out of Dove’s experiment was that humans seem to have the ability to reconstruct a narrative sequence even when presented with non-linear elements, such as that created by surrealist techniques like montage (Ray). Since the value of these words is somewhat inferred in a “surrealist” reading of the text, it may be efficacious to remove those words from our summarization. Since there only about two dozen of these words, I will use 23 of them as “stop words.”

Summarization

I did a further experiment, where I combined the extraction of nouns and verbs, the sequential approach seen in Figure 27, the elimination of the “to-be” stop words, and the frequency analysis. The resulting program extracts all sentences, from the email body and identifies the five most common nouns in the email. It then extracts only the nouns and verbs from each sentence, while extracting any stop words. The remaining nouns and verbs, from each sentence, are called “summary snippets,” and abbreviated *SumSnip*. Each snippet is tested to see if any of the five most common nouns are present. If so, the *SumSnip* is written to the XML file. The results of this experiment are shown in Figure 31. The resulting 92 words clearly express the nature of the email, and are a substantial reduction in volume from the original 244 words.

It is noted that this total, 92 words, is higher than one of the previous experiments reported in Figure 27. It was determined, while running various scripts, that there were two basic ways of parsing for sentences. One was to read the input file as a series of lines, while the other was to read the entire body into a single string and then parse it for sentences. The former method is easier, and yields accurate results when the author utilizes “bullet points,” where phrases are each on a single line or when the author fails to utilize recognized punctuation. The latter method loses any contextual clues from the author’s use of paragraphs, but results in accurate parsing of sentences, if they are correctly punctuated. If not, the results can be problematic. In testing 266 files from a single (Kenneth Lay) user’s ‘sent’ folder, numerous files were parsed, and there were “sentences” exceeding 200 words. For experimental purposes, I chose the latter approach for the experiment reported in Figure 31. The difference, in methodology, can be seen by comparing Figure 27 and Figure 31. This accounts for the difference in the number of words reported.

```

<?xml version="1.0" encoding="UTF-8"?>
- <Report>
  - <Email file="f:\Test5\443.txt">
    <HF_Words>['Enron', 'Dynergy', 'agreement', 'merger', 'Mark']</HF_Words>
    - <Summary>
      <SumSnip LOC="S1">know Enron signed merger agreement Dynergy acquire Enron.
        </SumSnip>
      <SumSnip LOC="S2">expect transaction close following shareholder approvals
        conditions Enron entered agreement U.S. laws require Enron Dynergy continue
        operate other. </SumSnip>
      <SumSnip LOC="S3">extent Enron Dynergy competitors businesses markets activities
        undertaken arm length transaction closed. </SumSnip>
      <SumSnip LOC="S5">procedure maintained. </SumSnip>
      <SumSnip LOC="S6">extent information required disclosed Dynergy merger agreement
        disclosure approved foregoing questions concerning notice please contact Lance
        Schuler Robert Eickenroht Mark Haedicke Rob Walls </SumSnip>
      <SumSnip LOC="S7">Thank help matter. </SumSnip>
    </Summary>
  </Email>
</Report>

```

Figure 31: Summary Snippets from Email 443

Description of the Software Developed

As described earlier, the software developed for this dissertation was written in Python and borrowed heavily from the Natural Language Processing Toolkit, several texts, and open source code. As previously stated, I am not a programmer. As a result, the code is neither elegant, nor efficient. It does function, as required, for this research.

As is common with object-oriented languages, this prototype was built using a script which, in turn, calls modules. In conducting the experiments for this dissertation, I developed a number of specifically designed modules, some of which were merely exploratory, while some were incorporated into the main processing script. The complete listing of the main program is located in Appendix C. A brief explanation, of the modules, follows Figure 32, which outlines the process.

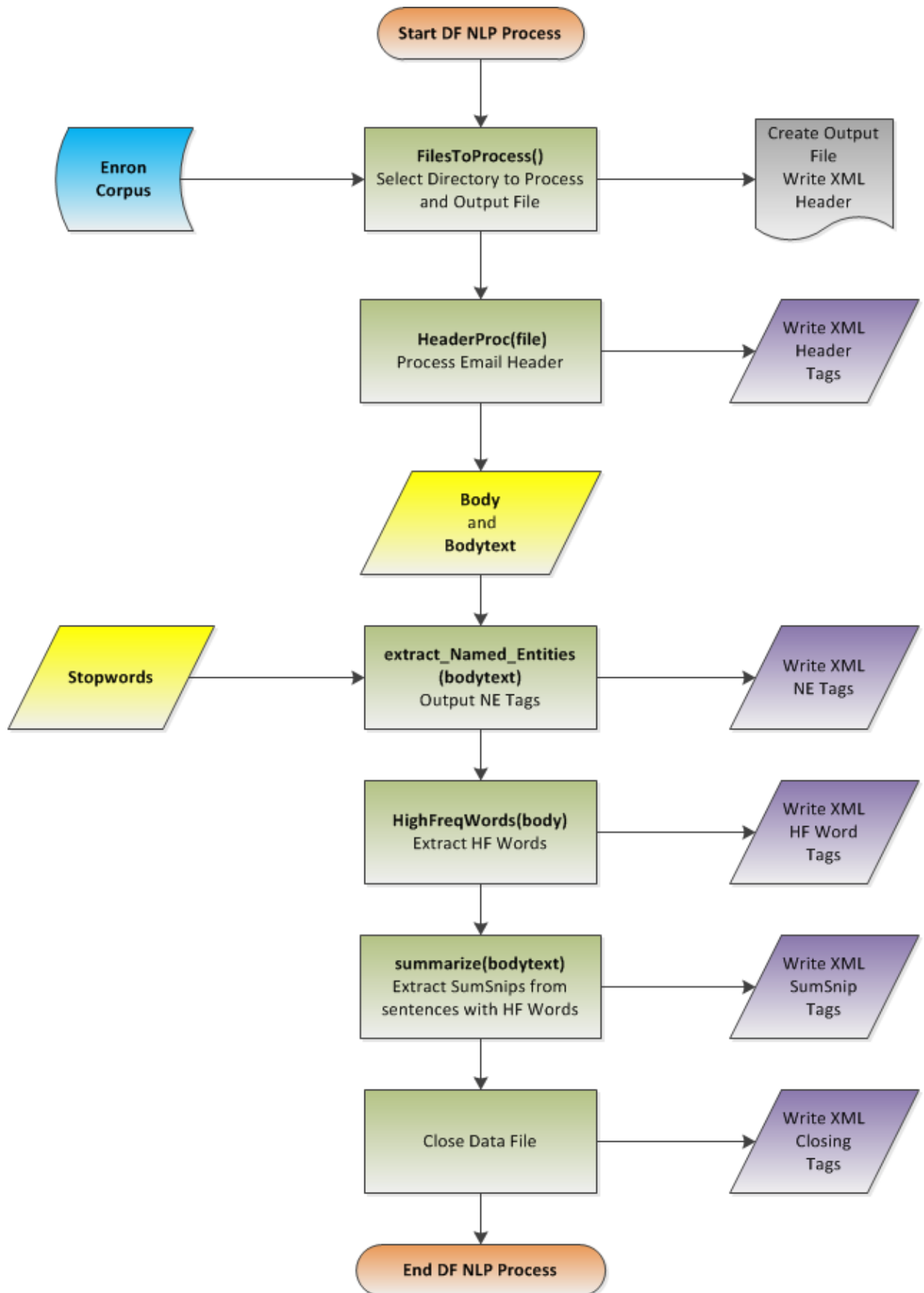


Figure 32: Digital Forensic NLP Program Flow

The first module, *FilesToProcess()*, requests user input to point to the directory containing the email files to process and the name and the location of the XML output file. It returns a list of file names, which is used as input, for the next module. The main program script parses the file list for a file name (*file*) and iterates through the entire file list, calling each of the following modules. It also opens the output file, and writes the XML file's header information.

The *HeaderProc(file)* module parses the file, line-by-line, until all of the desired metadata items are identified. These are then written to the output file in XML format. The module then processes the remainder of the file, as the body, and creates two objects. The *body* object is a list of lines in the body, while the *bodytext* object is a string of text, without carriage return or linefeed characters. These two objects are utilized by subsequent modules.

The *bodytext* is processed by the *extract_Named_Entities* module. It creates a set (a list of unique strings) of the identified named entities, within the entire email body. It then writes this set to the output file in XML format.

The *HighFreqWords(body)* module takes the lines from the body object, and creates a single, un-punctuated string. It then tokenizes it into a "bag of words," and tags the words by their part of speech. It then creates dictionaries for all of the nouns and verbs. It then does a frequency distribution on the nouns, and returns the five most common (high frequency) nouns. This object, *HF_Words*, is written, in XML format, to the output file, and is utilized by additional modules.

The *summarize(bodytext)* module is then called. This module takes the *bodytext* and parses it into sentences. Each sentence is then tokenized into words, and each word is tagged by part of speech. The tagged sentence is checked, against the list of *HF_Words*, and if one or more is present, all of the nouns and verbs (except for those in the object *stopwords*) are extracted and

written to a SumSnip in XML. This process repeats, until every sentence in the email is evaluated.

The main program script then closes the XML file tags and iterates to the next file until the file list is completed. The XML data tags are then closed, along with the output file.

Analysis and Discussion of Results

Email Header Processing

Most digital forensic software packages have at least basic capabilities to search emails by header information. The ability to conduct complex searches, using most digital forensic software, usually requires specialized programming in the forensic application's scripting language. The ability to perform both repetitive searches and complex ad hoc searches using an XML parser is a definite improvement. Email *Subject* lines, when accurately descriptive, can be very useful in identifying probative emails. Conversely, the content of progressive replies tends to shift away from the original content (Bekkerman). An additional advantage is that the information is stored in the same dataset as the content and can be utilized to provide context for the email content. For example, it might be possible to replace all of the "I" and "you" phrases with the names of the sender and receiver. Similarly, it might be possible to build a timeline, by utilizing the email's date and time as a base to plot the temporal expressions extracted using an NLP tool such as *timex*. The email header provides at least one data point in the chronology of the fabula. It may also provide one or more data points to establish the logic and/or events dimension of the fabula.

There are a number of potential shortcomings in utilizing the header information. The first is that, particularly with emails sent to large numbers of recipients, the amount of data stored in the XML file exceeds its analytical value.

Named Entity Extraction

From both a narrative and investigative perspective the extraction of named entities may be the most directly valuable product of the use of NLP techniques. From an investigative perspective, the ability to quickly identify subjects, witnesses, victims, organizations, and locations is crucial. The further ability to quickly identify emails mentioning these entities dramatically reduces the need to guess what search terms might be effective. Other researchers, such as Beebe, have utilized sophisticated techniques to cluster digital forensic data. And while those techniques have proved effective, the simple use of named entities might allow for “quick and dirty” clustering based on named entities. The research in this dissertation has not quantified this potential, but observations of the data would seem to support this notion.

John Lavorato was a senior executive with Enron. According to McLean and Elkind, he was actively involved in the fraudulent trading of energy futures and in defrauding the California Public Utilities Commission (215-6, 276,282). He is believed to have been named an “un-indicted co-conspirator” in the Federal Criminal case. Lavorato’s emails are part of the Enron Corpus and, as an experiment, the software, previously described, that collects named entities from an entire directory was run on the 17 emails contained in his email folder, labeled “california” (sic). The results of this test are shown in Figure 33.

```

- <Email>
  - <Header>
    <Filename>f:\Enron\lavorato-j\california\6</Filename>
  </Header>
  - <Body>
    <Named_Entities>set(['Sue Mara3', 'Update', 'Direct Access', 'SDG', 'Cal PX', 'Any News', 'Enron'])
    </Named_Entities>
  </Body>
</Email>
<Collected_NE>set(['Leslie Lawner Enron Summer', 'Edison', 'CFR', 'FloorLos Angeles', 'Hedges',
'extendedAnalysisAgain', 'Time', 'SB33X', 'Collateral', 'Enron', 'Sacramento', 'Update', 'Failure', 'LegislatureFor',
'Davis Siting Emergency Actions', 'x39510', 'News', 'Approved', 'Jim', 'Please', 'phillip_fantle', 'Michael', 'Mirant',
'ISO Tariffs', 'AB31xRuelte', 'Agenda', 'Cal', 'Very Brief Analysis', 'Mike Florio', 'Sandi', 'Any News', 'Steven',
'Electricity Oversight', 'Novell_GroupWise', 'Cliff Notes Version', 'CourtIf Davis', 'SBX6', 'Next', 'SoCal Ed', 'State',
'Mitigation', 'extendedAnalysis', 'Alan Connes7', 'Social', 'Kristen Bird', 'Energy Needs', 'MWs', 'Mike', 'Arizona',
'Enron Legislation Team Report', 'andrzej_kabarowski', 'Robert Johnston', 'Los Angeles', 'Consumer
OppositionHarvey Rosenfield', 'ISOREquires', 'Sam Wehn', 'McCutcheon', 'ENA', 'Minimize Outages', 'Jeff
Dasovich3', 'TRO', 'Susan', 'ConstructionThe', 'Davis', 'Stacey', 'Sue Mara3', 'Sandi McCubbin2', 'Reliant', 'Energy',
'Centralize Authority', 'Ginger', 'SBX32', 'CPUCGovernor Davis', 'Mike Smith3', 'Barry Ogilby', 'Solutions', 'Michael
Lifrak', 'Harry Oliver', 'Air Pollution Flexibility', 'Steps6', 'Expedite Siting', 'Team Reports Enron Model Legislative
Team Report', 'JAMES', 'Robert', 'districtsThe Board', 'Cal PX', 'ISO', 'Board', 'CDWR Credit Issues', 'CEQA', 'Local',
'JLHUEMOE', 'Assembly', 'William Urquhart', 'Kristin Walsh', 'GFergus', 'New Powerplants', 'Harvey Rosenfield',
'Water Resources', 'Control Board', 'California', 'Tammy', 'Sandra', 'Novell', 'JOHN', 'Association', 'Clean Air Act',
'CA', 'SMTP', 'LLP865 South Figueroa Street', 'katalin_kiss', 'Energy Commission', 'Harry', 'Ronald Carroll', 'Enron
Bankruptcy Response', 'Duke', 'Jeff', 'Federal Court', 'SB33x The Governor', 'PUC', 'GDB', 'California Power',
'Burton', 'Issues', 'FERC', 'Heather', 'Richard', 'New Generation', 'Direct Access', 'McCubbin', 'Maximize Output',
'BailoutBill', 'Electricity Oversight Board', 'Proposal', 'Rosenfield', 'Dydney', 'Jeffrey', 'Western Power', 'CPUC',
'Enron Interpretation', 'James', 'Require', 'Expedite', 'Update PX Credit Calculation', 'DWR', 'EES Petition', 'CERA
Study', 'Testimony', 'Order', 'x39934', 'Existing', 'IPPs', 'OutDavis', 'Richard Shapiro', 'CERA Special Report',
'Margaret Carson5', 'sterling_koch', 'SDG', 'Leslie', 'California State Power Authority', 'John Burton', 'SCE Court
Case', 'Summer Peak Season'])</Collected_NE>
</Data>

```

Figure 33: Named Entity Set from Lavorato's California Email Folder

A review of the collected set of named entities shows a large number of terms, names, and organizations that are described in the section of McLean and Elkind's book, which discusses the Enron's California activities (264-283). Doubtless, an investigator, reviewing this list, would recognize these terms, people, and organizations as playing a part in this aspect of the case. They would doubtless choose to read these emails in their original detail. While it is not possible, from reviewing only the named entities, to derive a narrative of the California events, one could clearly derive the subject/actor elements of the fabula.

Validation of the Use of Named Entities

In reviewing the data, it would appear that the identification of named entities provides a useful way to identify some of the actors in the fabula. Is there any way to measure the ability of this approach relative to the "actual" fabula? It occurred to me that perhaps there is, at least in this case. I observed, among many of the email accounts, that many of the users created topical folders for their email. A number of email recipients had folders with the word *California* in the

folder's label. McLean and Elkind's book, *The Smartest Guys in the Room*, contains a chapter called "Gaming California" (264-83). This chapter describes how Enron traders manipulated the electricity supply in the State of California, creating huge profits for the company, and causing a series of major power emergencies in the state. The chapter tells a compelling story and is itself a narrative containing sub-narratives. If the chapter and the emails are both narratives, "stories" in Bal's typology, about the same fabula, shouldn't they share most of the same narrative elements?

This would be analogous to two individuals simultaneously observing the same event. If they were later interviewed separately, they would tell different stories. The core "facts" about the people, places and activities would generally be the same for both witnesses, but they would construct, and deliver, the "story" in slightly different ways. To determine if they were explicating the same fabula, we can extract the elements of the fabulas of each witness's story and then compare the elements of one witness's elements to the other's. We are measuring the commonality between the two sets of fabula elements.

If we accept that the book is, at least arguably, the "real story," then we can measure the performance of our natural language processing against this "known" standard. Perhaps our natural language processing tools would allow us to make that comparison. I conducted the following experiment.

I scanned the book chapter's pages and used optical character recognition software (Adobe Acrobat Pro) to convert the printed book to ASCII text. I next wrote a routine that identified all of the email sub-directories (n=20), with the name California (both upper and lower cases) in them. I identified one additional directory, which contained *caliornia* (sic) in the directory name. Given that it was located between two other sub-directories named California

and included the word investigations, I chose to include it in the experiment. There were a total of 2385 emails, from nine users, in these directories.

All of the email directories were processed, recursively, using the named entity routine, and the output saved to an XML file. Similarly, the text from the book chapter was processed with the same routine. The two XML files were reduced to sets of unique entries using a Python routine. Another routine, using the intersection of the two sets, was used to identify the number of named entities found in the book chapter that were also present in the emails. The results of this experiment are shown in Table 4.

Table 4 Results of Named Entity Extraction from Ch. 17 and 2385 Emails

Source	Number of Unique Named Entities
McLean & Elkind Chapter 17	151
Enron Employee California Emails (n=2,385)	11,336
Intersection of Chapter & Emails	98

The 98 unique named entities, which were in both the book chapter and the emails, represent 64.9% of the 151 unique named entities in the chapter. In other words, just fewer than 65% of the named entities found in the book chapter were also found in the “California” emails. This ratio, which I will call the “commonality ratio” (C), can be represented, where NE^k is the set of named entities in the known text and NE^q is the set of named entities located in the questioned text, as:

$$C = \frac{NE^k}{NE^k \cap NE^q}$$

A review of the remaining 53 named entities from the book chapter that were not in the emails, revealed several issues. Seven of the entities, *Aside*, *BEWILDERED*, *Entire*, *GOOD LUCK*, *Inside Enron*, *Such and Thin*, were misclassified as named entities. An additional entity, *Ei Paso*, was determined, by reference to the original text, to be misspelled, likely due to an error in the optical character recognition phase of the experiment. Excluding these eight items, results in a commonality ratio of 98 out of 143, or 68.5%. Four entities, *Businessweek* (sic), *Energy Market Report*, *Motley Fool* and *PBS*, refer to media outlets quoted by the authors of the book, and are not part of the fabula, per se. When these are removed, the commonality ratio goes up to 70.5%. Given that the authors presumably relied on numerous sources besides the emails, the identification of over two-thirds of the pertinent named entities is significant.

Table 5 Results of Ch. 17 Experiment, After Data Scrubbing

	Entities in Book Chapter	Entities in Book and Emails	Commonality Ratio
Original Results	151	98	64.9%
Misclassified & Misspelled Removed	143	98	68.5%
External Literary References Removed	139	98	70.5%

It is possible, given the large number of named entities in the emails, the correspondence might be the result of chance. In any case, it would be useful to determine if the use of named entities is discriminatory in identifying particular actors in a given fabula.

I repeated the same experiment, except in the second instance, I used another chapter, from McLean and Elkind's book: "Andy Fastow's Secrets." This chapter, Chapter 11, is

approximately the same length as Chapter 16. Both were about 20 pages in length. After processing the text from Chapter 11, the NLP Named Entity Extractor identified 157 entities, compared to 151 for Chapter 16. Both chapters contained similar numbers of named entities. While Chapter 16 was about a particular fraud involving the State of California, Chapter 11 focused on the activities of one of the most senior executives in the Enron Corporation. Due to the nature of Chapter 11, the names of many of Enron’s senior executives, as well as a number of internal and external organizations, are included in the text. The results of this experiment are shown in Table 6.

Table 6 Results of Named Entity Extraction from Ch. 11 and Emails

Source	Number of Unique Named Entities
McLean & Elkind Chapter 11	157
Enron Employee California Emails (n=2,385)	11,336
Intersection of Chapter & Emails	62

A review of the named entities in Chapter 11, revealed that there were four that were misclassified as named entities: *Nor*, *Part*, *Securitization*, and *which*. There was one external reference to the Wall Street Journal, which was used as a source by the authors. After removing these from the chapter’s list of named entities, 152 remain. These results, and the calculated commonality ratio, are displayed in Table 7.

Table 7 Results after Data Scrubbing of Ch. 11 and Emails

	Entities in Book Chapter	Entities in Book and Emails	Commonality Ratio
Original Results	157	62	39.5%
Misclassified & Misspelled Removed	153	62	40.5%
External Literary References Removed	152	62	40.8%

Given that the large number of emails processed contained the names of people and organizations that were pertinent to a wide range of Enron operations, it is remarkable that the commonality ratio was so different. This is further reinforced by the fact that a number of major actors in the Enron saga are named in both chapters, thus increasing the ratio of both sets. Even in the worst case scenario, where the minimum commonality between the “California” emails and Chapter 16 is compared to the maximum commonality between the “California” emails and Chapter 11, the difference is 62.8%. This would seem to be a significant level of differentiation.

The analysis could taken one step further. If we assume that most of the named entities that are in common between Chapters 11 and 17 are not unique to the “California fabula,” they might be considered “false positives.” However, this ignores the overlapping nature of the fabula elements. Just because James Bond appears in both *Dr. No* and *Casino Royale*, does not mean that he is not a critical element of both stories.

Semantic Analysis

While the extraction of features, such as named entities, is reason enough to utilize NLP for digital forensic purposes, the goal of this research was to assist in the development of meaning from digital evidence. This aspect, which falls under the general topic of semantics, is

the most difficult and complex. This dissertation explored primarily two approaches to this problem — lexical and grammatical. A review of the literature revealed that there are a great many tools for performing grammar parsing. The results of these approaches, as demonstrated by the parse tree experiment, reinforced the tremendous difficulty of accurately extracting grammar, let alone meaning, from even well-formed, grammatically correct text. Since most people do not consistently write using proper grammar and punctuation, especially in abbreviated genres such as emails, I recognized that using a fully grammatical approach would not be feasible.

Given that the stated objective of the dissertation is to identify narratives, this presents a significant problem. However, narratology provides us with an approach that allows for a deconstruction of the text by reference to the elements of the fabula, rather than to the grammar. If actors, events, and chronology can be identified, then a formal grammar is not needed.

Reflecting on the notion of surrealism as a disjointed, abstract set of signs, and taking a clue from a script writer's technique of using only nouns in a comic exchange,(e.g.: Miranda: "I Tilly birthday party Sunday." Sally: "I think you dropped your verbs."), I chose to focus on the extraction of nouns and verbs, and to experiment with how their relationships might be utilized.

One of the experiments proved very interesting. The extraction of an email's high frequency nouns, with their associated verbs, appeared to have good potential. The initial experiment on the test email demonstrated, in the very first sentence, that the approach was capable of producing a useful synopsis.

The second set of experiments revolved around the collection of high frequency nouns from directories of emails. While this approach was usefully summative for small collections of topically organized emails, but the results became less selective, as the volume increased, and/or the emails were less topically organized. This is seen in Figure 34, in the difference between the

results from the *controls* ($n=6$), the *California* ($n=17$) and the *sent* directories ($n=647$). The first two provide some keywords that relate to the topic in the folder. The latter has very little content, and is comprised of words that are either associated with the author, such as his first name, “John,” his nickname, “Lavo,” the company name, and the abbreviations for morning and evening.

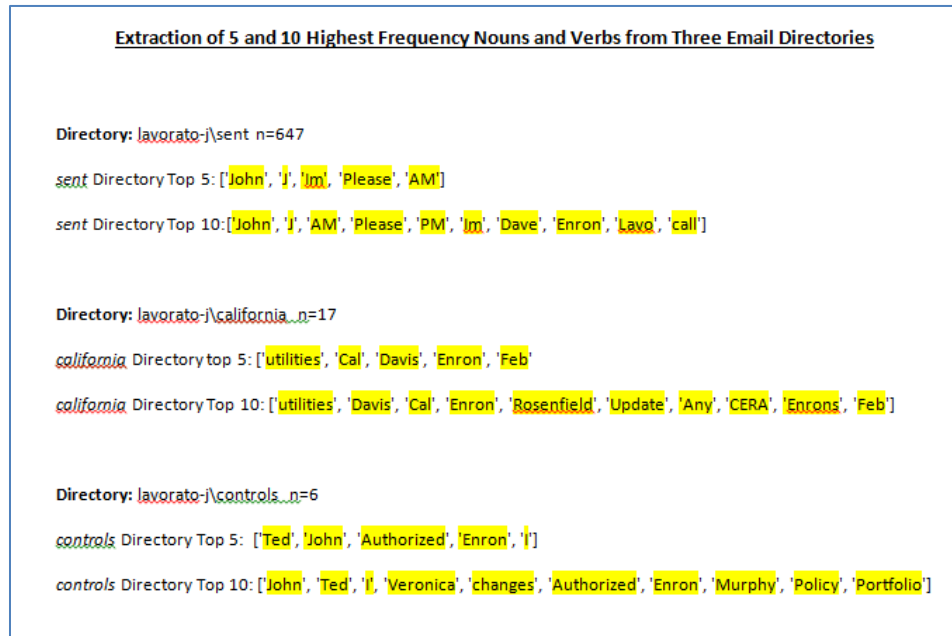


Figure 34: Five and Ten Highest Frequency Words from Email Directories

Summarization

The extraction of high frequency nouns, and associated verbs, was extended by parsing each sentence in an email to extract all of the nouns and verbs where at least one high frequency noun was contained in the sentence. This process produced a summary snippet, which I abbreviated as “SumSnip” for tagging purposes. This process produced a crude form of summarization, which, in some cases, was very useful and in others much less so.

A review of the summaries for a large number of emails revealed that, generally, when the text was well-organized, grammatically correct, text, the process produced useful results.

And while the results from emails comprise of sentence fragments and “bullet points” could be very useful, often the summary results were less useful, and sometimes meaningless.

This technique did provide some reduction in the volume of data to review: compression ratios averaged approximately 60% .When applied to large groups of emails, the methodology could still be applied, but required a great deal of reading by the analyst. It would appear that using the summary snippets as a secondary review, after selecting interesting emails utilizing the metadata and the named entities would be a more efficient approach.

Limitations

The contents of digital evidence can be wildly disparate. It can take the form of graphic images, audio, video, and a wide range of textual styles and genres. As a result no single analytical approach will suffice. The techniques explored in this dissertation are limited in their application. The header extraction is obviously limited to extracting data from emails. The lexical, grammatical, and semantic processing has been seen to be most effective on “well formed” text. If the text is not well formed, it appears that our ability to extract narratives is reduced.

What might happen if the subjects of a forensic analysis were to intentionally attempt to obfuscate their communications by techniques such as using code words to represent probative information? Would not these techniques fail? The answer is both yes and no. Certainly, coded text that is processed with the techniques described here might not extract the “true meaning” of the text. That is, the meaning intended by the sender to be understood by the recipient. If the subjects maintain the more or less proper use of English grammar, while the lexical element, such as the named entities and verbs, will be misrepresented, their relationships will not. It is likely, from an analytical perspective, that eventually these elements will conflict with the

external context. The substituting of “birthday party,” for the expression “assassination attempt,” will likely result, at least over time, in some rather peculiar text, thus raising suspicion. Since the relationships between the coded words keeps the same grammatical relationships as in un-coded text, once the meaning of the coded words is identified, then the NLP analysis can function effectively. In fact, it can be utilized to “decode” the coded messages. If the suspects decide to create their own grammar and vocabulary, it would be obvious that some form of coding has occurred and a cryptanalysis would be required. Interestingly, the frequency analysis of the words, and the ability of semantic parsers to be trained, provides useful tools for these cryptanalysis tasks.

CHAPTER SIX: CONCLUSIONS AND FUTURE WORK

Conclusions

In this dissertation I describe how several disparate approaches: knowledge management, content analysis, narratology, and natural language processing, can be combined in an interdisciplinary way to positively impact the growing difficulty in developing useful, actionable intelligence from the ever-increasing corpora of digital evidence. After describing how these techniques apply to the digital forensic process, I proffered two new theoretical constructs: The Hermeneutic Theory of Digital Forensics and the Narrative Theory of Digital Forensics. These link the existing theories of forensic science, content analysis, narratology, and natural language processing together in order to identify and extract probative narratives from digital evidence.

In order to demonstrate the latter theory, it was necessary to develop a series of experiments. The Enron email corpus was selected as the test data. Software was developed, using existing code from the Natural Language Processing Toolkit, along with additional routines, methods, and modules, to identify structural, lexical, and semantic elements of the test data. The NLP tools, developed for these demonstrations, were utilized to process selected emails. In order to store and manipulate the elements extracted from the corpus, it was necessary to expand the existing XML models for forensic data. The objectives for these demonstrations were to evaluate each of the four proposed hypotheses.

The selected emails and email directories, were processed to extract all of the internal metadata, such as sender, addressee, subject, dates, and times. These were recorded in XML structures. The body of the email was processed, utilizing standard NLTK tokenizers, taggers and classifiers, to extract named entities, high frequency nouns, and all verbs. No effort was

make to optimize for accuracy or repeatability. The extracted entities from each sentence were combined into a “summary snippet.”

After processing a number of test emails, a number of observations were made. Based on these observations, which will be discussed below, several experiments were formulated to measure the commonality between the extracted named entities from two sets of data. The first is the sub-set of the test data defined by directories labeled “California” by their owners, and the other being the extracted named entities from the chapter in McLean and Elkind’s book concerning fraudulent activities in California. This experiment revealed a strong commonality. In an effort to measure the discriminatory value of this methodology, the same experiment was conducted, using an un-related chapter of the same book. This experiment demonstrated a substantially lower correspondence to the processed emails.

Hypothesis One

Hypothesis One was stated as: The identification of narratives, by automated means, can contribute to the efficiency and effectiveness of the forensic examiner/investigative analyst.

Given the increasing size of digital storage devices, the need for increased efficiency is obvious. Similarly, the use of automation is an obvious choice to improve efficiency. The key part of this hypothesis is if an automated process for identification of narratives is possible. If it is, clearly the hypothesis is not rejected. We must evaluate whether the experiments described in this dissertation actually identify narratives.

For the purposes of this dissertation, I adopted Bal’s typology for defining a narrative. The three elements of a narrative, according to Bal, are the actors, actions/events, and the chronology. It can be said that through the use of named entities and nouns, the experiments succeeded in identifying the actors. Similarly, the verbs do identify most of the actions or events.

Between the date/time data stored in the email header and any chronological information identified in the body of the email, it can be very weakly argued that the chronological element has been captured. This was not, however, the intent. The objective was to identify the important and probative narratives, including all three of the elements, in an efficient way. The experiments undertaken in this dissertation did not succeed to that level. The manner in which the experiments used of verbs, while useful, did not assist in any material way towards identifying the probative narratives. The limited experiments conducted using the NLP tools *dateutil* and *timex*, did not yield useful results within the experimental process.

While the identification of the action/event and chronology elements of the fabula may not have been a complete success, the use of NLP tools to identify the actors of the fabula appears very promising. The correlation between the experimental processing of the emails and the fabula elements in the pertinent chapter of the non-fiction book were strong. Further, it appears that there is a substantially less correlation to non-pertinent chapter's text.

The identification of all the named entities in a text offers ways to potentially answer the *who*, *where*, and perhaps *what* investigative questions. These questions are often the most important questions, and the ability to identify them quickly, provides substantial value to the investigative analysis. Without the automated identification of precise words and phrases that have potential to be probative, conducting string searches of the evidence is a "trial and error" process, which might never search for the appropriate entities.

While failing to completely affirm the hypothesis, I would suggest that the results do not reject the hypothesis. The results of the experiments provide a strong indication that an automated process can improve the ability of an analyst to identify key elements of a fabula.

Hypothesis Two

Hypothesis Two was stated as follows: Any automated process that improves the ability of the forensic examiner/investigative analyst to quickly identify probative narratives will improve the efficiency and effectiveness of the process.

The difference between Hypotheses One and Two, concerns the differentiation between just any narrative and a probative one. The results of the experiments yielded mixed results. On one hand, the notion that an automated process can be of assistance was affirmed. By reducing the volume of information to be reviewed and summarizing the text, the process becomes less burdensome. This improvement, albeit nominal, affirms the hypothesis. Unfortunately, none of the experiments undertaken for this research independently identified which summaries were more probative than any other. Clearly, more research is warranted in this area.

Hypothesis Three

Hypothesis Three was stated as follows: The use of nouns and verbs will assist in the identification of both general and probative narratives, more economically than reading complete texts.

The summary snippets, produced by extracting named entities, high frequency nouns, and all verbs, generated a significantly reduced volume of data for review by a forensic practitioner or investigator. Some of the snippets provided clear and concise summaries that were clearly probative. However, there were many others which were not. One subjective observation was that there seemed to be some correlation between the length and genre of the text. That is, email bodies that were longer and contained well-formed paragraphs and sentences, appeared to produce more useful snippets, while those that were composed of short phrases or “bullet points,” produced more snippets, of less value. Additionally, where clear and concise snippets were

present, it subjectively appeared that they occurred in the beginning or end of the email body, typically within two sentences of the beginning or end of the email. One possible explanation for this would be that when writers are drafting well-formed text, they utilize rhetorical constructs as they were taught when they learned how to write. This observation suggests that additional research into the evaluation of the stylistics, organization and statistics of the text. Perhaps evaluating the text via some metric, such as Flesch-Kincaid or Gunning Fog indices, might be an indicator of the utility of a given type of natural language processing. Similarly, some form of automated genre classification system might be useful, to determine what techniques might be more efficient and effective.

Hypothesis Four

Hypothesis Four was stated as follows: Natural language processing software can assist in the identification of probative narratives by use of lexical, grammatical, and semantic techniques.

The experiments described in this dissertation clearly demonstrate that NLP software can, with a reasonable level of accuracy, identify lexical, grammatical, and semantic elements of digital evidence texts. While the software was not able to identify all of the elements of the fabula, nor identify the specifically probative narratives, it was sufficiently effective to affirm this hypothesis. The utility of identifying named entities is sufficiently robust to implement, with little modification, in current digital forensic processing. Further research into the comparison of extracted sets of named entities appears to be very likely to yield useful results with a modest level of effort. The results of these experiments suggest further research into more difficult and complex aspects of narrative identification, would be fruitful.

Future Work

Hermeneutic Theory of Digital Forensics

This dissertation has suggested an over-arching theorem for the emerging discipline of digital forensics. While proffering such a theorem may be viewed as ambitious, or even presumptuous, it is offered in the spirit of research. It can be evaluated, criticized, tested, disproved, and improved. Any discipline, to be called “scientific,” according to Kuhn, must have an underlying paradigm. This theorem is offered as a “shared example” (187-8) or “candidate paradigm,” in the Kuhnian sense:

History suggests that the road to a firm research consensus is extraordinarily arduous. ...In the absence of some candidate for paradigm, all the facts that could possibly pertain to the development of a given science are likely to seem equally relevant....Only very occasionally, as in the case of ancient statics, dynamics, and geometrical optics, do facts collected with so little guidance from pre-established theory speak with sufficient clarity to permit the emergence of a first paradigm. (15-15)

While there has been a fair volume of research done in the digital forensic field, a consensus has not developed as to a framework for future research. Some researchers have focused on the computer science aspects of the field, while a few, such as Beebe and Chaski, have focused on investigative and linguistic approaches. One of the goals, in setting out this theoretical construct, was to provide a "map" where researchers could situate their research and recognize the interrelationships between the different dimensions of the field. It may be that this construct is rejected by the community. If so, and an alternate framework is suggested, tested,

and accepted, then this research will have done its part in a "revolutionizing" the science of digital forensics.

The Narrative Theory of Digital Forensics

The results of the experiments reported in this dissertation support the notion that narrative is an effective way to view digital evidence. The current practice of digital forensics makes minimal use of the content within the files extracted from the examination process. The use of lexical, grammatical, and semantic elements to identify, or at least summarize, the narrative content has been shown.

At least with respect to the test data for these experiments, the length, structural organization, and degree of textual complexity seemed to have an impact on the ability to extract narrative elements and produce probative summaries. This suggests that additional work on automated means of identifying genres, organization, stylistics, and textual sophistication might be not only useful, but necessary to further the automated analysis of narratives. Perhaps a more advanced study of the elements of fabula, or an examination of the aspects of the story, such as sequential ordering, rhythm, frequency, characters space, and focalization, will suggest ways of utilizing natural language processing to further identify and evaluate narrative texts (Bal 75-163).

However, the use of the three elements of the fabula, the actors, the events/actions, and the chronology, has not been equally effective. The use of nouns, and particularly named entities, to identify the actors, has proven to be relatively straight-forward and produced obvious results. The identification of the events/actions, through the use of verbs, has not been demonstrated to any level of utility in these experiments. One issue identified in the experiments, which was also noted in some of the literature, is that the verbs tend to have a "flatter tree," meaning that there are fewer verbs used, and each verb is used with less frequency. As a result, techniques such as

frequency analysis do not seem to be particularly effective. The result is that it is difficult to identify which verbs are more significant to the narrative than others. Further investigation will be needed to identify potential techniques that will be able to identify which verbs are most significant. Similarly, the use of lexical, grammatical, and semantic tools to establish a narrative chronology will clearly require much additional work.

If perfected, these techniques could be applied to digital evidence in other fashions as well. The identification of authorship, as well as the identification of the subject actually at the keyboard, has always been a forensic question. Because the writing of a narrative is a sufficiently complex undertaking that produces a large volume of artifacts, a statistical and narrative analysis might be able to demonstrate that, a given text was written by a particular author. This notion is analogous to the identification of handwriting, where the repetitive details of a complex motor task (writing) leaves a consistent set of details. An author's use of particular fabula elements, in specific patterns might be amenable to statistical analysis.

XML

The previous work on the use of XML by Craiger and Garfinkel, combined with the work described in this dissertation, has demonstrated the flexibility and utility of using XML to store tagged artifacts of the digital forensic process. This suggests several further avenues of research. First, the utility of XML tags would be enhanced if there were some standardized terminology and definitions for the various common elements and tags. There is nothing "wrong" with the elements and tags developed by Craiger, Garfinkel, or myself. In order to be useful to the community, however, there must be a common, explicit definition of what each element and tag represents. This would need to be a community effort, perhaps led by an organization such as SWGDE or NIST.

One of the strengths of XML is the ability to manipulate the tagged data through the use of technologies, such as: parsers, XSL, XLink, and XPointer (Applen and McDaniel 162-96, 215-94). Research, focused on the notion of “visual rhetoric,” described by Applen and McDaniel, might in a digital forensic context, prove to be very enlightening (131-6). Is it possible to reify the complex digital forensic data in a more compelling and accurate way, by leveraging XML? How should we criticize, in a rhetorical sense, the products of our forensically mediated data?

Non-Forensic Applications for NLP Narrative Analysis

The use of natural language processing to identify narratives and their elements is not limited to forensic science. Extending the approaches and techniques described in this dissertation, it might be possible to utilize them for a wide array of textual analysis and manipulation.

Diagramming a sentence is a well-understood methodology to teach sentence structure. In a similar way, it is possible to “diagram” narratives. Such an approach could be used, for both analyzing how authors construct their writing, and teaching students how to write narratives. Perhaps, there might be a way to make this both visual and interactive. Such an approach might be useful in teaching individuals with cognitive disabilities.

If the automated identification of narratives can be achieved, it might be possible to utilize these techniques as a sort of “hypertext parser,” where textual material could be automatically extracted into HTML or XML elements which could then be re-used. For example, it would be possible to build custom textbooks comprised of data collected from many sources.

The automated identification of narratives has tremendous potential, specifically for the digital forensic field, and for textual studies in general. There is clearly enough potential to develop one, or more, research agendas.

This research has focused on the content of emails in the digital forensic context. In many ways these are “traditional” texts. They generally comport to the common use of sentence and paragraph structures, the organization of thought, and some modicum of proper grammar. However, electronic communications are becoming much less traditional, especially when coupled with electronic devices. The use of instant messaging, tweets, and the various social media, coupled with an increasing use of digital photography and videography, will make the sorts of analysis described in this dissertation much more difficult. If investigators and digital forensic practitioners wish to have any hope of successfully pursuing investigations in the future, there is an urgent need for research in the automated identification of narratives in digital evidence. At the 1997 meeting of the Association for Computing Machinery, Bran Ferren, the former Vice-President of Disney Imagineering, stated that “the computer is just barely getting good enough for storytelling.” He went on to suggest that the computer is not an information processing tool, but rather a storytelling tool (Ferren). Going forward, our ability to understand the evidence stored in electronic form will require our ability to understand “the story,” as told by the computer.

APPENDIX A—TEXT OF EXAMPLE EMAIL 443

Message-ID: <11730746.1075862738340.JavaMail.evans@thyme>
Date: Tue, 27 Nov 2001 18:47:00 -0800 (PST)
From: announcements.enron@enron.com
To: dl-ga-all_enron_worldwide1@enron.com
Subject: Enron/Dynegy Merger; Antitrust Issues
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Enron General Announcements
</O=ENRON/OU=NA/CN=RECIPIENTS/CN=MBX_ANNCENRON>
X-To: DL-GA-all_enron_worldwide1 </O=ENRON/OU=NA/CN=RECIPIENTS/CN=DL-GA-all_enron_worldwide1>
X-cc:
X-bcc:
X-Folder: \JWOLFE (Non-Privileged)\Wolfe, Jason\Inbox
X-Origin: Wolfe-J
X-FileName: JWOLFE (Non-Privileged).pst

As you know, Enron has signed a merger agreement by which Dynegy will acquire Enron. We expect the transaction to close following shareholder and regulatory approvals and various conditions to closing.

Even though Enron has entered into this agreement, U.S. and foreign antitrust laws require that Enron and Dynegy continue to operate independently of each other. In particular, to the extent that Enron and Dynegy are competitors in various businesses or markets, their respective activities must be undertaken at arm's length until the transaction has closed. Therefore, for antitrust purposes you should treat Dynegy as you would any other unaffiliated company notwithstanding the merger agreement.

In addition, all information, documents and communications related to the merger between Enron and Dynegy should be coordinated through and approved by Mark Muller, Lance Schuler, Robert Eickenroht, Mark Haedicke, Rob Walls or Greg Whalley of Enron. It is absolutely critical that this procedure be maintained. To the extent that information is required to be disclosed to Dynegy under the merger agreement, then such disclosure should be approved by one of the foregoing individuals.

If you have any questions concerning this notice, please contact Lance Schuler (713/853-5419), Robert Eickenroht (713/853-3155), Mark Haedicke (713/853-6544) or Rob Walls (713/646-6017). Thank you for your help in this matter.

APPENDIX B—FORENSIC XML CONTENT FILE EXAMPLE —EMAIL 443

```

<?xml version="1.0" encoding="UTF-8"?>
- <Data>
  - <Email>
    - <Header>
      <Filename>f:\Test5\443.txt</Filename>
      <Message-ID>1730746.1075862738340.JavaMail.evans@thyme</Message-ID>
      <Date>Tue, 27 Nov 2001 18:47:00 -0800 (PST)</Date>
      <From>announcements.enron@enron.com</From>
      <To>dl-ga-all_enron_worldwide1@enron.com</To>
      <Subject>Enron/Dynegy Merger; Antitrust Issues</Subject>
    </Header>
    - <Body>
      <Named_Entities>set(['Rob Walls', 'Robert Eickenroht', 'Mark Haedicke', 'Greg Whalley', 'U.S.', 'Dynegy', 'Mark Muller', 'Lance Schuler', 'Enron'])</Named_Entities>
      <HF_Words>['Enron', 'Dynegy', 'agreement', 'merger', 'Mark']</HF_Words>
    - <Summary>
      <Content>As you know, Enron has signed a merger agreement by which Dynegy will acquire Enron.</Content>
      <POS_Tags>[('As', 'IN'), ('you', 'PRP'), ('know', 'VBP'), (';', ';'), ('Enron', 'NNP'), ('has', 'VBZ'), ('signed', 'VBN'), ('a', 'DT'), ('merger', 'NN'), ('agreement', 'NN'), ('by', 'IN'), ('which', 'WDT'), ('Dynegy', 'NNP'), ('will', 'MD'), ('acquire', 'VB'), ('Enron.', 'NNP')]</POS_Tags>
      <SumSnip LOC="S1">know Enron signed merger agreement Dynegy acquire Enron.</SumSnip>
      <Content>We expect the transaction to close following shareholder and regulatory approvals and various conditions to closing.Even though Enron has entered into this agreement, U.S. and foreign antitrust laws require that Enron and Dynegy continue to operate independently of each other.</Content>
      <POS_Tags>[('We', 'PRP'), ('expect', 'VBP'), ('the', 'DT'), ('transaction', 'NN'), ('to', 'TO'), ('close', 'VB'), ('following', 'VBG'), ('shareholder', 'NN'), ('and', 'CC'), ('regulatory', 'JJ'), ('approvals', 'NNS'), ('and', 'CC'), ('various', 'JJ'), ('conditions', 'NNS'), ('to', 'TO'), ('closing.Even', 'JJ'), ('though', 'IN'), ('Enron', 'NNP'), ('has', 'VBZ'), ('entered', 'VBN'), ('into', 'IN'), ('this', 'DT'), ('agreement', 'NN'), (';', ';'), ('U.S.', 'NNP'), ('and', 'CC'), ('foreign', 'JJ'), ('antitrust', 'JJ'), ('laws', 'NNS'), ('require', 'VBP'), ('that', 'IN'), ('Enron', 'NNP'), ('and', 'CC'), ('Dynegy', 'NNP'), ('continue', 'NN'), ('to', 'TO'), ('operate', 'VB'), ('independently', 'RB'), ('of', 'IN'), ('each', 'DT'), ('other.', 'NNP')]</POS_Tags>
      <SumSnip LOC="S2">expect transaction close following shareholder approvals conditions Enron entered agreement U.S. laws require Enron Dynegy continue operate other.</SumSnip>
      <Content>In particular, to the extent that Enron and Dynegy are competitors in various businesses or markets, their respective activities must be undertaken at arm's length until the transaction has closed.</Content>
      <POS_Tags>[('In', 'IN'), ('particular', 'JJ'), (';', ';'), ('to', 'TO'), ('the', 'DT'), ('extent', 'NN'), ('that', 'IN'), ('Enron', 'NNP'), ('and', 'CC'), ('Dynegy', 'NNP'), ('are', 'VBP'), ('competitors', 'NNS'), ('in', 'IN'), ('various', 'JJ'), ('businesses', 'NNS'), ('or', 'CC'), ('markets', 'NNS'), (';', ';'), ('their', 'PRP$'), ('respective', 'JJ'), ('activities', 'NNS'), ('must', 'MD'), ('be', 'VB'), ('undertaken', 'VBN'), ('at', 'IN'), ('arm', 'NN'), (';', 'POS'), ('length', 'NN'), ('until', 'IN'), ('the', 'DT'), ('transaction', 'NN'), ('has', 'VBZ'), ('closed.', 'NNP')]</POS_Tags>
      <SumSnip LOC="S3">extent Enron Dynegy competitors businesses markets activities undertaken arm length transaction closed.</SumSnip>
      <Content>Therefore, for antitrust purposes you should treat Dynegy as you would any other unaffiliated company notwithstanding the merger agreement.In addition, all information, documents and communications related to the merger between Enron and Dynegy should be coordinated through and approved by Mark Muller, Lance Schuler, Robert Eickenroht, Mark Haedicke, Rob Walls or Greg Whalley of Enron.</Content>
      <POS_Tags>[('Therefore', 'NNP'), (';', ';'), ('for', 'IN'), ('antitrust', 'JJ'), ('purposes', 'NNS'), ('you', 'PRP'), ('should', 'MD'), ('treat', 'VB'), ('Dynegy', 'NNP'), ('as', 'IN'), ('you', 'PRP'), ('would', 'MD'), ('any', 'DT'), ('other', 'JJ'), ('unaffiliated', 'JJ'), ('company', 'NN'), ('notwithstanding', 'VBG'), ('the', 'DT'), ('merger', 'NN'), ('agreement.In', 'NN'), ('addition', 'NN'), (';', ';'), ('all', 'DT'), ('information', 'NN'), (';', ';'), ('documents', 'NNS'), ('and', 'CC'), ('communications', 'NNS'), ('related', 'VBN'), ('to', 'TO'), ('the', 'DT'), ('merger', 'NN'), (';', 'between', 'IN'), ('Enron', 'NNP'), ('and', 'CC'), ('Dynegy', 'NNP'), ('should', 'MD'), ('be', 'VB'), ('coordinated', 'VBN'), ('through', 'IN'), ('and', 'CC'), ('approved', 'VBD'), ('by', 'IN'), ('Mark', 'NNP'), ('Muller', 'NNP'), (';', ';'), ('Lance', 'NNP'), ('Schuler', 'NNP'), (';', ';'), ('Robert', 'NNP'), ('Eickenroht', 'NNP'), (';', ';'), ('Mark', 'NNP'), ('Haedicke', 'NNP'), (';', ';'), ('Rob', 'NNP'), ('Walls', 'NNP'), ('or', 'CC'), ('Greg', 'NNP'), ('Whalley', 'NNP'), ('of', 'IN'), ('Enron.', 'NNP')]</POS_Tags>
      <SumSnip LOC="S4">Therefore purposes treat Dynegy company notwithstanding merger agreement.In addition information documents communications related merger Enron Dynegy coordinated approved Mark Muller Lance Schuler Robert Eickenroht Mark Haedicke Rob Walls Greg Whalley Enron.</SumSnip>
      <Content>It is absolutely critical that this procedure be maintained.</Content>
      <POS_Tags>[('It', 'PRP'), ('is', 'VBZ'), ('absolutely', 'RB'), ('critical', 'JJ'), ('that', 'IN'), ('this', 'DT'), ('procedure', 'NN'), ('be', 'VB'), ('maintained.', 'NNP')]</POS_Tags>
      <Content>To the extent that information is required to be disclosed to Dynegy under the merger agreement, then such disclosure should be approved by one of the foregoing individuals.If you have any questions concerning this notice, please contact Lance Schuler (713/853-5419), Robert Eickenroht (713/853-3155), Mark Haedicke (713/853-6544) or Rob Walls (713/646-6017).</Content>
      <POS_Tags>[('To', 'TO'), ('the', 'DT'), ('extent', 'NN'), ('that', 'IN'), ('information', 'NN'), ('is', 'VBZ'), ('required', 'VBN'), ('to', 'TO'), ('be', 'VB'), ('disclosed', 'VBN'), ('to', 'TO'), ('Dynegy', 'NNP'), ('under', 'IN'), ('the', 'DT'), ('merger', 'NN'), ('agreement', 'NN'), (';', ';'), ('then', 'RB'), ('such', 'JJ'), ('disclosure', 'NN'), ('should', 'MD'), ('be', 'VB'), ('approved', 'VBN'), ('by', 'IN'), ('one', 'CD'), ('of', 'IN'), ('the', 'DT'), ('foregoing', 'NN'), ('individuals.If', '-NONE-'), ('you', 'PRP'), ('have', 'VBP'), ('any', 'DT'), ('questions', 'NNS'), ('concerning', 'VBG'), ('this', 'DT'), ('notice', 'NN'), (';', ';'), ('please', 'NN'), ('contact', 'NN'), ('Lance', 'NNP'), ('Schuler', 'NNP'), (';', 'Lance', 'NNP'), ('713/853-5419', 'CD'), (';', 'Robert', 'NNP'), ('Eickenroht', 'NNP'), (';', 'Robert', 'NNP'), ('713/853-3155', 'CD'), (';', 'Mark', 'NNP'), ('Haedicke', 'NNP'), (';', 'Mark', 'NNP'), ('713/853-6544', 'CD'), (';', 'Rob', 'NNP'), ('Walls', 'NNP'), (';', 'Rob', 'NNP'), ('713/646-6017', 'CD'), (';', 'Rob', 'NNP'), ('713/646-6017', 'CD'), (';', ';')]</POS_Tags>
      <SumSnip LOC="S6">extent information required disclosed Dynegy merger agreement disclosure approved foregoing questions concerning notice please contact Lance Schuler Robert Eickenroht Mark Haedicke Rob Walls</SumSnip>
      <Content>Thank you for your help in this matter.</Content>
      <POS_Tags>[('Thank', 'NNP'), ('you', 'PRP'), ('for', 'IN'), ('your', 'PRP$'), ('help', 'NN'), ('in', 'IN'), ('this', 'DT'), ('matter.', 'NNP')]</POS_Tags>
    </Summary>
  </Body>
</Email>
</Data>

```

APPENDIX C—DIGITAL FORENSIC NLP TOOLS PROGRAM

```
#-----  
# Name:      DFemailNLP Ver 3  
# Purpose:   Email Extraction - Includes NE extraction  
#  
# Author:    Mark Pollitt  
#  
# Created:   1/14/2013  
# Copyright: (c) Mark Pollitt 2013  
# Licence:   ALL Rights Reserved  
#-----  
  
def FilesToProcess():  
    import os  
    directorylist=os.listdir(source)  
    for x in directorylist:  
        file_name=str(x)  
        filelist.append(str(source)+'\\'+file_name)  
    return filelist  
  
def HeaderProc(file):  
    body=''  
    bc=0  
    count=0  
    linebody=[]  
    a=open(file)  
    lines=a.readlines()  
    linecount=len(lines)  
    data_out.write('<Header>'+'\n')  
    file_tag='<Filename>'+file+' </Filename>'+'\n'  
    data_out.write(file_tag)  
  
    while count < linecount:  
        s=lines[count]  
        # print s  
  
        if s[:13] == 'Message-ID: <':  
            ss='<Message-ID>'+s[14:-1]+'</Message-ID>'+'\n'  
            data_out.write(ss)  
  
        elif s[:5] == 'Date:':  
            ss='<Date>'+s[6:-1]+'</Date>'+'\n'  
            data_out.write(ss)  
  
        elif s[:5] == 'From:':  
            ss='<From>'+s[6:-1]+'</From>'+'\n'  
            data_out.write(ss)  
  
        elif s[:3] == 'To:':  
            ss='<To>'+s[4:-1]+'</To>'+'\n'  
            data_out.write(ss)  
  
        elif s[:8] == 'Subject:':  
            ss='<Subject>'+s[9:-1]+'</Subject>'+'\n'  
            data_out.write(ss)  
    # Finds end of header
```

```
elif s[:11] == 'X-FileName:':
    print s
    break

count=count+1
data_out.write('</Header>'+'\n')

for n in range(0,linecount):
    p=lines[n]
    if p[:11] == 'X-FileName:':
        bc=n+1
        break
    else: continue
for n in range(bc,linecount):
    p=lines[n]
    body+=p

bodytext=body.replace('\n',' ',999)
global bodytext
global body

def HighFreqWords(body):

# Translate Lines into un-punctuated string, word tokenize and POS tag
    tran = body.translate(string.maketrans("", ""), string.punctuation)
    str_words=word_tokenizer.tokenize(tran)
    str_pos=nlk.pos_tag(str_words)

#extract and count
    nouns = []
    verbs = []
    noun_count=0
    verb_count=0
    Counts=nlk.defaultdict(int)

#Identify all nouns and verbs
    for (word, tag) in str_pos:
        if tag[0:2] == 'NN' and word > "A" and word < "z":
            nouns.append(word)
            noun_count+=1
        elif tag[0] == 'V' and word > "A" and word < "z":
            verbs.append(word)
            verb_count+=1

# Create frequency dist for nouns & verbs and sort keys
    fdistn=FreqDist(nouns)
    nounvocab=fdistn.keys()

    fdistv=FreqDist(verbs)
    verbvocab=fdistv.keys()

# Export HF nouns
    HF_Words=nounvocab[0:5]
    global HF_Words
```

```
HFVstring=('<HF_Words>'+str(HF_Words)+'</HF_Words>')
data_out.write(HFVstring)

def extract_Named_Entities(bodytext):
    # Credit to Christopher Groskopf - https://gist.github.com/onyxfish
    sentences = nltk.sent_tokenize(bodytext)
    tokenized_sentences = [nltk.word_tokenize(sentence) for sentence in sentences]
    tagged_sentences = [nltk.pos_tag(sentence) for sentence in tokenized_sentences]
    chunked_sentences = nltk.batch_ne_chunk(tagged_sentences, binary=True)

    def extract_entity_names(t):
        entity_names = []
        if hasattr(t, 'node') and t.node:
            if t.node == 'NE':
                entity_names.append(' '.join([child[0] for child in t]))
            else:
                for child in t:
                    entity_names.extend(extract_entity_names(child))
        return entity_names
    entity_names = []
    for tree in chunked_sentences:
        entity_names.extend(extract_entity_names(tree))

    # Print unique entity names
    data_out.write('<Named_Entities>'+str(set(entity_names))+'</Named_Entities>'+'\n')

def summarize(bodytext):
    para_num = 1
    TotalWords=0
    HFlen=len(HF_Words)
    data_out.write('<Summary>')
    # Tokenize Bodytext to sentences
    sents=sent_tokenize(bodytext)
    bodylen=len(sents)
    for n in range (0,bodylen):
        l=sents[n]
        if len(l)>2:
            word_list=[]
            pos_tags=[]
            label=''
            outstring=''
            NounTest=0
            word_list=word_tokenizer.tokenize(l)

    #Label the Location where the snippet found.

        label=('S'+str(n+1))

    # Write full sentence
    data_out.write ('<Content>'+l+'</Content>'+'\n')

    #POS tag the tokenized sentence - then remove stop words

    pos_tags=nltk.pos_tag(word_list)
    data_out.write ('<POS_Tags>'+str(pos_tags)+'</POS_Tags>'+'\n')
```

```

# Extract nouns and verbs and remove stopwords
for (word, tag) in pos_tags:
    if tag[0] == 'V' or tag[0] == 'N':
        if word > "A" and word < "z":
            if word not in StopWords:
                outstring+=((word)+' ')
#Look for sentences with High Freq nouns in them.
#this produces "summary snippits"
for n1 in range(0,HFlen):
    sw=HF_Words[n1]
    swf=outstring.find(sw)
    if swf>=0:
        twords=word_tokenizer.tokenize(outstring)
        TotalWords+=len(twords)
        data_out.write('<SumSnip
LOC='+'\'+label+'\'+>'+outstring+'</SumSnip>'+'\n')
        break
    else: continue
else:
    continue

#output summary snippits with HF nouns
data_out.write('</Summary>')

# ##### Main Script #####

#Declarations
filelist=[]
entity_names=[]
# Input directories to process and output file
source = raw_input ('Input Directory to Parse: ')
outfile = raw_input ('Name of output file:')
# Import Libraries
import nltk
import nltk.data
import string
from nltk import FreqDist
from nltk.tokenize import PunktWordTokenizer
from nltk import sent_tokenize
tokenizer = nltk.data.load ('tokenizers/punkt/english.pickle')
word_tokenizer = PunktWordTokenizer()
StopWords = [ 'be', 'am', 'is', 'are', 'was', 'were', 'been', 'has', 'have', 'had', 'do', 'did',
'does', 'can', 'could', 'shall', 'should', 'will', 'would', 'may', 'might', 'must']

#Begin Procedures
FilesToProcess()
data_out=open(outfile, 'w')
data_out.write('<?xml version="1.0" encoding="utf-8"?>'+'\n')
data_out.write('<Data>'+'\n')
for file in filelist:
    data_out.write('<Email>'+'\n')
    HeaderProc(file)
    data_out.write('<Body>')
    extract_Named_Entities(bodytext)

```

```
HighFreqWords(body)
summarize(bodytext)
data_out.write('</Body>')
data_out.write('</Email>'+'\n')
data_out.write('</Data>'+'\n')
data_out.close()
```

APPENDIX D—ELSEVIER COPYRIGHT RELEASE – ROUSSEV ARTICLE

Title of your thesis/dissertation	THE HERMENEUTICS OF THE HARD DRIVE: USING NARRATOLOGY, NATURAL LANGUAGE PROCESSING AND KNOWLEDGE MANAGEMENT TO IMPROVE THE EFFECTIVENESS OF THE DIGITAL FORENSIC PROCESS
Expected completion date	May 2013
Estimated size (number of pages)	150
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.0 USD / 0.0 GBP
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the

combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

8. License Contingent Upon Payment: While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. Warranties: Publisher makes no representations or warranties with respect to the licensed material.

10. Indemnity: You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. No Transfer of License: This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. No Amendment Except in Writing: This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. Objection to Contrary Terms: Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. Revocation: Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. Translation: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. Website: The following terms and conditions apply to electronic reserve and author websites:

Electronic reserve: If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if:

This license was made in connection with a course,

This permission is granted for 1 year only. You may obtain a license for future website posting.

All content posted to the web site must maintain the copyright information line on the bottom of each image,

A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx> or the Elsevier homepage for books at <http://www.elsevier.com> , and

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. Author website for journals with the following additional clauses:

All content posted to the web site must maintain the copyright information line on the bottom of each image, and the permission granted is limited to the personal version of your paper. You are not allowed to download and post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed edition to create an electronic version. A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx> . As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier's online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article's Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. Author website for books with the following additional clauses:

Authors are permitted to place a brief summary of their work online only.

A hyper-text must be included to the Elsevier homepage at <http://www.elsevier.com> . All content posted to the web site must maintain the copyright information line on the bottom of each image. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version.

Central Storage: This license does not include permission for a scanned version of the

material to be stored in a central repository such as that provided by Heron/XanEdu.

19. **Website (regular and for author):** A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx> or for books to the Elsevier homepage at <http://www.elsevier.com>

20. **Thesis/Dissertation:** If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. **Other Conditions:**

v1.6

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500908782.

Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

APPEDNIX E—ELSEVIER COPYRIGHT RELEASE – WOOD ARTICLE

**ELSEVIER LICENSE
TERMS AND CONDITIONS**

Dec 02, 2012

This is a License Agreement between Mark Pollitt ("You") and Elsevier ("Elsevier") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Elsevier, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

Supplier	Elsevier Limited The Boulevard, Langford Lane Kidlington, Oxford, OX5 1GB, UK
Registered Company Number	1982084
Customer name	Mark Pollitt
Customer address	[REDACTED] Orlando, FL 32828
License number	3040760636309
License date	Dec 02, 2012
Licensed content publisher	Elsevier
Licensed content publication	Journal of Magnetism and Magnetic Materials
Licensed content title	Future hard disk drive systems
Licensed content author	Roger Wood
Licensed content date	March 2009
Licensed content volume number	321
Licensed content issue number	6
Number of pages	7
Start Page	555
End Page	561
Type of Use	reuse in a thesis/dissertation
Portion	figures/tables/illustrations
Number of figures/tables /illustrations	1
Format	electronic
Are you the author of this Elsevier article?	No
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	THE HERMENEUTICS OF THE HARD DRIVE: USING NARRATOLOGY, NATURAL LANGUAGE PROCESSING AND KNOWLEDGE MANAGEMENT TO IMPROVE THE EFFECTIVENESS OF THE DIGITAL FORENSIC

	PROCESS
Expected completion date	May 2013
Estimated size (number of pages)	150
Elsevier VAT number	GB 494 6272 12
Permissions price	0.00 USD
VAT/Local Sales Tax	0.0 USD / 0.0 GBP
Total	0.00 USD
Terms and Conditions	

INTRODUCTION

1. The publisher for this copyrighted material is Elsevier. By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at <http://myaccount.copyright.com>).

GENERAL TERMS

2. Elsevier hereby grants you permission to reproduce the aforementioned material subject to the terms and conditions indicated.

3. Acknowledgement: If any part of the material to be used (for example, figures) has appeared in our publication with credit or acknowledgement to another source, permission must also be sought from that source. If such permission is not obtained then that material may not be included in your publication/copies. Suitable acknowledgement to the source must be made, either as a footnote or in a reference list at the end of your publication, as follows:

"Reprinted from Publication title, Vol /edition number, Author(s), Title of article / title of chapter, Pages No., Copyright (Year), with permission from Elsevier [OR APPLICABLE SOCIETY COPYRIGHT OWNER]." Also Lancet special credit - "Reprinted from The Lancet, Vol. number, Author(s), Title of article, Pages No., Copyright (Year), with permission from Elsevier."

4. Reproduction of this material is confined to the purpose and/or media for which permission is hereby given.

5. Altering/Modifying Material: Not Permitted. However figures and illustrations may be altered/adapted minimally to serve your work. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of Elsevier Ltd. (Please contact Elsevier at permissions@elsevier.com)

6. If the permission fee for the requested use of our material is waived in this instance, please be advised that your future requests for Elsevier materials may attract a fee.

7. Reservation of Rights: Publisher reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment

terms and conditions.

8. **License Contingent Upon Payment:** While you may exercise the rights licensed immediately upon issuance of the license at the end of the licensing process for the transaction, provided that you have disclosed complete and accurate details of your proposed use, no license is finally effective unless and until full payment is received from you (either by publisher or by CCC) as provided in CCC's Billing and Payment terms and conditions. If full payment is not received on a timely basis, then any license preliminarily granted shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.

9. **Warranties:** Publisher makes no representations or warranties with respect to the licensed material.

10. **Indemnity:** You hereby indemnify and agree to hold harmless publisher and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.

11. **No Transfer of License:** This license is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.

12. **No Amendment Except in Writing:** This license may not be amended except in a writing signed by both parties (or, in the case of publisher, by CCC on publisher's behalf).

13. **Objection to Contrary Terms:** Publisher hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and publisher (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.

14. **Revocation:** Elsevier or Copyright Clearance Center may deny the permissions described in this License at their sole discretion, for any reason or no reason, with a full refund payable to you. Notice of such denial will be made using the contact information provided by you. Failure to receive such notice will not alter or invalidate the denial. In no event will Elsevier or Copyright Clearance Center be responsible or liable for any costs, expenses or damage incurred by you as a result of a denial of your permission request, other than a refund of the amount(s) paid by you to Elsevier and/or Copyright Clearance Center for denied permissions.

LIMITED LICENSE

The following terms and conditions apply only to specific license types:

15. Translation: This permission is granted for non-exclusive world **English** rights only unless your license was granted for translation rights. If you licensed translation rights you may only translate this content into the languages you requested. A professional translator must perform all translations and reproduce the content word for word preserving the integrity of the article. If this license is to re-use 1 or 2 figures then permission is granted for non-exclusive world rights in all languages.

16. Website: The following terms and conditions apply to electronic reserve and author websites:

Electronic reserve: If licensed material is to be posted to website, the web site is to be password-protected and made available only to bona fide students registered on a relevant course if:

This license was made in connection with a course,

This permission is granted for 1 year only. You may obtain a license for future website posting.

All content posted to the web site must maintain the copyright information line on the bottom of each image,

A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx> or the Elsevier homepage for books at <http://www.elsevier.com>, and

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

17. Author website for journals with the following additional clauses:

All content posted to the web site must maintain the copyright information line on the bottom of each image, and the permission granted is limited to the personal version of your paper. You are not allowed to download and post the published electronic version of your article (whether PDF or HTML, proof or final version), nor may you scan the printed edition to create an electronic version. A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx>. As part of our normal production process, you will receive an e-mail notice when your article appears on Elsevier's online service ScienceDirect (www.sciencedirect.com). That e-mail will include the article's Digital Object Identifier (DOI). This number provides the electronic link to the published article and should be included in the posting of your personal version. We ask that you wait until you receive this e-mail and have the DOI to do any posting.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

18. Author website for books with the following additional clauses:

Authors are permitted to place a brief summary of their work online only.

A hyper-text must be included to the Elsevier homepage at <http://www.elsevier.com>. All content posted to the web site must maintain the copyright information line on the bottom of each image. You are not allowed to download and post the published electronic version of your chapter, nor may you scan the printed edition to create an electronic version.

Central Storage: This license does not include permission for a scanned version of the material to be stored in a central repository such as that provided by Heron/XanEdu.

19. **Website (regular and for author):** A hyper-text must be included to the Homepage of the journal from which you are licensing at <http://www.sciencedirect.com/science/journal/xxxxx> or for books to the Elsevier homepage at <http://www.elsevier.com>

20. **Thesis/Dissertation:** If your license is for use in a thesis/dissertation your thesis may be submitted to your institution in either print or electronic form. Should your thesis be published commercially, please reapply for permission. These requirements include permission for the Library and Archives of Canada to supply single copies, on demand, of the complete thesis and include permission for UMI to supply single copies, on demand, of the complete thesis. Should your thesis be published commercially, please reapply for permission.

21. **Other Conditions:**

v1.6

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500908781.

Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

APPENDIX F—SPRINGER COPYRIGHT RELEASE – VENTER ARTICLE

December 17, 2012

Springer reference

Advances in Digital Forensics III

IFIP International Conference on Digital Forensics , National Center for Forensic Science, Orlando Florida, January 28-January 31, 2007

Series: IFIP Advances in Information and Communication Technology, Vol. 242

Volume package: Advances in Digital Forensics

Craiger, Philip; Shenoj, Sujeet (Eds.)

2007, XX, 358 p. 242 illus.

ISBN 978-0-387-73741-6

Figs 2 and 3

Your thesis

University: University of Central Florida

Title:

Print run: Up to 100 copies

Language: English

Territory: Worldwide

With reference to your request to reuse material in which Springer Science+Business Media controls the copyright, our permission is granted free of charge under the following conditions:

Springer material

- represents original material which does not carry references to other sources (if material in question refers with a credit to another source, authorization from that source is required as well);
- requires full credit (book title, year of publication, page, chapter title, name(s) of author(s), original copyright notice) is given to the publication in which the material was originally published by adding: "With kind permission of Springer Science+Business Media";
- may not be altered in any manner. Any other abbreviations, additions, deletions and/or any other alterations shall be made only with prior written authorization of the author and/or Springer Science+Business Media;
- may not be republished in Electronic Open Access.

This permission

- is non-exclusive;
- is valid for one-time use only for up to 100 copies
- includes use in an electronic form, provided it is password protected, on intranet or university's repository, including UMI (according to the definition on the Sherpa website: <http://www.sherpa.ac.uk/romeo/>), or CD-Rom/DVD, or E-book;
- is subject to courtesy information to the author (address is given in the book/chapter);
- is personal to you and may not be sublicensed, assigned, or transferred by you to any other person without Springer's written permission;
- is valid only when the conditions noted above are met.

Permission free of charge does not prejudice any rights we might have to charge for reproduction of our copyrighted material in the future.

Best regards,

Rights and Permissions



Springer-Verlag GmbH
Tiergartenstr. 17
69121 Heidelberg
Germany
E-mail: permissions.heidelberg@springer.com

APPENDIX G—ACM COPYRIGHT RELEASE – FAN ARTICLE

**ASSOCIATION FOR COMPUTING MACHINERY, INC. LICENSE
TERMS AND CONDITIONS**

Feb 06, 2013

This is a License Agreement between Mark Pollit ("You") and Association for Computing Machinery, Inc. ("Association for Computing Machinery, Inc.") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Association for Computing Machinery, Inc., and the payment terms and conditions.

License Number	3083280756134
License date	Feb 06, 2013
Licensed content publisher	Association for Computing Machinery, Inc.
Licensed content publication	Communications of the ACM
Licensed content title	Tapping the power of text mining
Licensed content author	Weiguo Fan, et al
Licensed content date	Sep 1, 2006
Volume number	49
Issue number	9
Type of Use	Thesis/Dissertation
Requestor type	Academic
Format	Electronic
Portion	figure/table
Number of figures/tables	1
Will you be translating?	No
Order reference number	
Title of your thesis/dissertation	THE HERMENEUTICS OF THE HARD DRIVE: USING NARRATOLOGY, NATURAL LANGUAGE PROCESSING AND KNOWLEDGE MANAGEMENT TO IMPROVE THE EFFECTIVENESS OF THE DIGITAL FORENSIC PROCESS
expected completion date	May 2013
Estimated size (pages)	150
Billing Type	Credit Card
Credit card info	Visa ending in 1658
Credit card expiration	05/2014
Total	8.00 USD
Terms and Conditions	

Rightslink Terms and Conditions for ACM Material

<https://rightslink.com/App/DispatchServlet>

1/4

1. The publisher of this copyrighted material is Association for Computing Machinery, Inc. (ACM). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the Billing and Payment terms and conditions established by Copyright Clearance Center, Inc. ("CCC"), at the time that you opened your Rightslink account and that are available at any time at).

2. ACM reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

3. ACM hereby grants to licensee a non-exclusive license to use or republish this ACM-copyrighted material* in secondary works (especially for commercial distribution) with the stipulation that consent of the lead author has been obtained independently. Unless otherwise stipulated in a license, grants are for one time use in a single edition of the work, only with a maximum distribution equal to the number that you identified in the licensing process. Any additional form of republication must be specified according to the terms included at the time of licensing.

*Please note that ACM cannot grant republication or distribution licenses for embedded third-party material. You must confirm the ownership of figures, drawings and artwork prior to use.

4. Any form of republication or redistribution must be used within 180 days from the date stated on the license and any electronic posting is limited to a period of six months unless an extended term is selected during the licensing process. Separate subsidiary and subsequent republication licenses must be purchased to redistribute copyrighted material on an extranet. These licenses may be exercised anywhere in the world.

5. Licensee may not alter or modify the material in any manner (except that you may use, within the scope of the license granted, one or more excerpts from the copyrighted material, provided that the process of excerpting does not alter the meaning of the material or in any way reflect negatively on the publisher or any writer of the material).

6. Licensee must include the following copyright and permission notice in connection with any reproduction of the licensed material: "[Citation] © YEAR Association for Computing Machinery, Inc. Reprinted by permission." Include the article DOI as a link to the definitive version in the ACM Digital Library. Example: Charles, L. "How to Improve Digital Rights Management." Communications of the ACM, Vol. 51:12, © 2008 ACM, Inc. <http://doi.acm.org/10.1145/nnnnn.nnnnnn> (where nnnnnnnnnn is replaced by the actual number).

7. Translation of the material in any language requires an explicit license identified during the licensing process. Due to the error-prone nature of language translations, Licensee must include the following copyright and permission notice and disclaimer in connection with any reproduction of the licensed material in translation: "This translation is a derivative of ACM-copyrighted material. ACM did not prepare this translation and does not guarantee that it is an accurate copy of the originally published work. The original intellectual property contained in this work remains the property of ACM."

8. You may exercise the rights licensed immediately upon issuance of the license at the end of the licensing transaction, provided that you have disclosed complete and accurate details of your proposed use. No license is finally effective unless and until full payment is received from you (either by CCC or ACM) as provided in CCC's Billing and Payment terms and conditions.
9. If full payment is not received within 90 days from the grant of license transaction, then any license (preliminarily granted) shall be deemed automatically revoked and shall be void as if never granted. Further, in the event that you breach any of these terms and conditions or any of CCC's Billing and Payment terms and conditions, the license is automatically revoked and shall be void as if never granted.
10. Use of materials as described in a revoked license, as well as any use of the materials beyond the scope of an unrevoked license, may constitute copyright infringement and publisher reserves the right to take any and all action to protect its copyright in the materials.
11. ACM makes no representations or warranties with respect to the licensed material and adopts on its own behalf the limitations and disclaimers established by CCC on its behalf in its Billing and Payment terms and conditions for this licensing transaction.
12. You hereby indemnify and agree to hold harmless ACM and CCC, and their respective officers, directors, employees and agents, from and against any and all claims arising out of your use of the licensed material other than as specifically authorized pursuant to this license.
13. This license is personal to the requestor and may not be sublicensed, assigned, or transferred by you to any other person without publisher's written permission.
14. This license may not be amended except in a writing signed by both parties (or, in the case of ACM, by CCC on its behalf).
15. ACM hereby objects to any terms contained in any purchase order, acknowledgment, check endorsement or other writing prepared by you, which terms are inconsistent with these terms and conditions or CCC's Billing and Payment terms and conditions. These terms and conditions, together with CCC's Billing and Payment terms and conditions (which are incorporated herein), comprise the entire agreement between you and ACM (and CCC) concerning this licensing transaction. In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions, these terms and conditions shall control.
16. This license transaction shall be governed by and construed in accordance with the laws of New York State. You hereby agree to submit to the jurisdiction of the federal and state courts located in New York for purposes of resolving any disputes that may arise in connection with this licensing transaction.
17. There are additional terms and conditions, established by Copyright Clearance Center, Inc. ("CCC") as the administrator of this licensing service that relate to billing and payment for licenses provided through this service. Those terms and conditions apply to each transaction as if they were restated here. As a user of this service, you agreed to those terms and conditions at the time that

you established your account, and you may see them again at any time at <http://myaccount.copyright.com>

18. Thesis/Dissertation: This type of use requires only the minimum administrative fee. It is not a fee for permission. Further reuse of ACM content, by ProQuest/LMI or other document delivery providers, or in republication requires a separate permission license and fee. Commercial resellers of your dissertation containing this article must acquire a separate license.

Special Terms:

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 48 hours of the license date. Payment should be in the form of a check or money order referencing your account number and this invoice number RLNK500951557. Once you receive your invoice for this order, you may pay your invoice by credit card. Please follow instructions provided at that time.

**Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006**

For suggestions or comments regarding this order, contact RightsLink Customer Support: customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.

Gratis licenses (referencing \$0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

REFERENCES

- Abbott, H. Porter. *The Cambridge Introduction to Narrative*. 2nd ed. Cambridge University Press, 2008. Print.
- Anderson, Terence, David Schum, and William Twining. *Analysis of Evidence*. 2nd ed. Cambridge University Press, 2005. Print.
- Appen, J.D., and Rudy McDaniel. *The Rhetorical Nature of XML: Constructing Knowledge in Networked Environments*. 1st ed. Routledge, 2009. Print.
- Bal, Mieke. "The Point of Narratology." *Poetics Today* 11.4 (1990): 727–753. Print.
- Bal, Mieke. *Narratology: Introduction to the Theory of Narrative*, Third Edition. 3rd ed. University of Toronto Press, Scholarly Publishing Division, 2009. Print.
- Barrionuevo, Alexei. "Judge Sentences Former Enron Chief to 24 Years in Prison - Business - International Herald Tribune - The New York Times." Web. 9 Aug. 2012.
- Barry, John. "Lawyers Say They Need a Year to Review Computer Files in Schenecker Case" *Tampa Bay Times*. Web. 4 Aug. 2012.
- Barthes, Roland. *Image, Music, Text*. New York: Hill and Wang, 1977. Print.
- Beebe, Nicole, and Jan Clark. "Dealing with Terabyte Data Sets in Digital Investigations." *Advances in Digital Forensics*. 2005. 3–16. Web. 12 Nov. 2008.
- Beebe, Nicole Lang and Jan Guynes Clark. "Digital Forensic Text String Searching: Improving Information Retrieval Effectiveness by Thematically Clustering Search Results." *Digital Investigation* 4.Supplement (2007): 49–54. Print.

- Beebe, Nicole. "Digital Forensic Research: The Good, the Bad and the Unaddressed." *Advances in Digital Forensics V*. Ed. Gilbert Peterson & Sujeet Sheno. Vol. 306. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009. 17–36. Web. 11 May 2011.
- Beebe, Nicole Lang, et al. "Post-retrieval Search Hit Clustering to Improve Information Retrieval Effectiveness: Two Digital Forensics Case Studies." *Decision Support Systems* 51 732–744. Print.
- Bekkerman, Ron. "Email Classification on Enron Dataset." Web. 8 Dec. 2012.
- Bird, Steven, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. 1st ed. O'Reilly Media, 2009. Print.
- Blackmun, William Daubert, Et Ux., Etc., Et Al., Petitioners V. Merrell Dow Pharmaceuticals, Inc. 1993. Web. 8 Nov. 2009.
- Brehaut, Ernest. *An Encyclopedist of the Dark Ages, Isidore of Seville*. New York: Franklin, 1964.
- Breton, André. "Surrealist Manifesto." *Surrealist Manifesto*. Web. 11 Oct. 2012.
- Carey, Sorcha. *Pliny's Catalogue of Culture : Art and Empire in the Natural History*. New York: Oxford University Press, 2003.
- Carrier, Brian. *File System Forensic Analysis*. Boston, Mass: Addison-Wesley, 2005. Print.
- Carvey, Harlan. *Windows Forensic Analysis Including DVD Toolkit*. Pap/DVD. Syngress, 2007. Print.
- Casey, Eoghan. *Digital Evidence and Computer Crime*. Academic Press, 2004. Print.
- "Category: Digital Forensics XML." *Forensics Wiki*. Web. 29 Oct. 2012.
- "Chandra Levy - Wikipedia, the Free Encyclopedia." Web. 2 Aug. 2012.

- Chaski, Carole, Blake Howald, and Judith Parker. "Text-Typing Threat Letters." Conference Papers — Law & Society. 2006. 1. aph. Web.
- Chaski, Carole. "The Keyboard Dilemma and Authorship Identification." *Advances in Digital Forensics III*. Ed. Philip Craiger & Sujeet Sheno. 2007. 133–146. edo. Web.
- Clancy, Tom. *The Hunt for Red October*. Reprint. Berkley, 2010. Print.
- Cohen, L. Jonathan. *The Probable and the Provable*. Oxford University Press, USA, 1977. Print.
- Cohen, William W. "Enron Email Dataset." Enron Email Dataset. Web. 9 Nov. 2012.
- Cornell Legal Information Institute."Wex."Web. 2 Aug. 2012.
- Craiger, Philip, Mark Pollitt, and Jeff Swauger. "Law Enforcement and Digital Evidence." *Handbook of Information Security*. Ed. Hossein Bidgoli. I. 3 vols. Wiley, 2005. II–49. Print.
- Craiger, Philip. "Digital Evidence Markup Language: An Object-Oriented, XML-based Model for Sharing Computer Crime-related Information." 2009. Web. 29 October 2012.
- Crowston, Kevin. Personal interview. 15 July 2008.
- Curran, James M. "Statistics in Forensic Science." *Wiley Interdisciplinary Reviews: Computational Statistics* 1.2 (2009): 141–156. Print.
- Davenport, Thomas H., and Laurence Prusak. *Working Knowledge*. Harvard Business Press, 2000. Print.
- Deehl, David L. "Demonstrative Evidence in Federal Court." Web. 11 Sept. 2012.

- de Waal, Alta, Jacobus Venter, and Etienne Barnard. "Applying Topic Modeling to Forensic Data." *Advances in Digital Forensics IV*. Ed. Indrajit Ray & Sujeet Sheno. 2008. 115–126. edb. Web.
- Diekema, Anne. Personal interview. 11 July 2008.
- Diesner, Jana, Terrill Frantz, and Kathleen Carley. "Communication Networks from the Enron Email Corpus 'It's Always About the People. Enron Is No Different'." *Computational & Mathematical Organization Theory* 11.3 (2005): 201–228. Web. 2 Aug. 2012.
- DisplaySearch. "Smartphone Shipments to Pass One Billion in 2016, According to NPD DisplaySearch - DisplaySearch." Web. 19 Sept. 2012.
- Doyle, John R. "Mapping the world of consumption: computational linguistics analysis of the Google text corpus." OAIster. Web.
- Edmonds, Philip. "Choosing the Word Most Typical in Context Using a Lexical Co-occurrence Network." *Proceedings of ACL-EACL '97, Student Session*. 1998. edsarx. Web.
- "EDRM XML White Paper « The Electronic Discovery Reference Model." *ERDM XML White Paper*. Web. 9 Nov. 2012.
- Eisenstein, Sergei. *The Film Sense*. Revised. Trans. Jay Leyda. Harcourt Brace Jovanovich, 1969. Print.
- Fan, Weiguo, Linda Wallace, and Stephanie Rich. "Tapping the Power of Text Mining." *Communications of the ACM* 49.9 (2006): 76–82. Print.
- Farmer, Dan, and Wietse Venema. *Forensic Discovery*. 1st ed. Addison-Wesley Professional, 2005. Print.

- Federal Bureau of Investigation. "Regional Computer Forensics Laboratory Program Annual Report 2011." Web. 2011.
- "Federal Rules of Evidence (LII 2009 Ed.)." Web. 8 Nov. 2009.
- Fei, B. et al. "Exploring Forensic Data with Self-Organizing Maps." International Federation for Information Processing -Publications- IFIP. Ed. M. Pollitt & S. Sheno. 2006. 113–126. edsbl. Web.
- Ferren, Bran. "The Future of Storytelling." ACM97. Web. 2013
- Fisher, Jim. *Forensics Under Fire: Are Bad Science and Dueling Experts Corrupting Criminal Justice?* New Brunswick, N.J: Rutgers University Press, 2008. Print.
- Foley, Mary Jo. "It's War: Google + Quickoffice Vs. Microsoft Office Everywhere | ZDNet." ZDNet. Web.
- Forensic Sciences Foundation. "Jurisprudence." Web. 2 Sept. 2012.
- Garfinkel, Simson. "Dfxml.dtd." Web. 29 October 2012.
- Garfinkel, Simson L. "Carving Contiguous and Fragmented Files with Fast Object Validation." *Digital Investigation* 4, Supplement.0 (2007): 2–12.
- Garfinkel, Simson.L. "Automating Disk Forensic Processing with SleuthKit, XML and Python." 2009 Fourth International IEEE Workshop on Systematic Approaches to Digital Forensic Engineering (2009): 73. Print.
- Garfinkel, Simson L. "Digital Forensics Research: The Next 10 Years." *Digital Investigation* 7, Supplement (2010): S64–S73. Web.
- Gibson, Jennifer. "Surrealism Before Freud: Dynamic Psychiatry's 'Simple Recording Instrument'." *Art Journal* 46.1 (1987): 56–60. Print.

- Hicks, Richard C., Ronald Dattero, and Stuart D Galup. "The Five-Tier Knowledge Management Hierarchy." *Journal of Knowledge Management* 10.1 (2006): 19-31. ABI/INFORM Global, ProQuest. Web. 23 Mar. 2012.
- Hijmans, Ellen. "The Logic of Qualitative Media Content Analysis: A Typology." *Communications* 21.1 (1996): 93–108. Print.
- Hilbert, Martin, and Priscila López. "The World's Technological Capacity to Store, Communicate, and Compute Information." *Science* 332.6025 (2011): 60–65.
- Hueske, Edward E. "Firearms and Toolmarks." *The Forensic Laboratory Handbook Procedures and Practice*. Ed. Ashraf Mozayani & Carla Noziglia. Humana Press, 2011. 227–264. Print.
- Inman, Keith, and Norah Rudin. *Principles and Practice of Criminalistics: The Profession of Forensic Science*. 1st ed. CRC Press, 2000. Print.
- Jones, Tom. "SCAFO Online Articles." Opinion vs. Conclusion. Web. 27 Sept. 2012.
- Jurafsky, Daniel, and James H. Martin. *Speech and Language Processing*. 2nd ed. Prentice Hall, 2008. Print.
- Kaplan, Ronald E. "Computer Forensics—What Is It Good For?" *Journal of Digital Forensic Practice* 2.2 (2008): 57–61. Print.
- Kappagoda, A. "The Use of Systemic-Functional Linguistics in Automated Text Mining." (2009): n. pag. Web. 12 Aug. 2012.
- Kent, Karen et al. "NIST Special Publication 800-86: Guide to Integrating Forensic Techniques into Incident Response." Aug. 2006.
- Krippendorff, Klaus H. and Mary Angela Bock. *The Content Analysis Reader*. Sage Publications, Inc, 2008. Print.

- Kruse, Warren G., and Jay G. Heiser. *Computer Forensics: Incident Response Essentials*. Addison-Wesley Professional, 2001. Print.
- Kuhn, Thomas S. *The Structure of Scientific Revolutions*. 3rd ed. University Of Chicago Press, 1996. Print.
- Lahneman, William J. "The Future of Intelligence Analysis, Volume I, Final Report." University of Maryland, 10 Mar. 2006. Web.
- Leitch, Vincent B., ed. *The Norton Anthology of Theory and Criticism*. 1st ed. New York: Norton, 2001. Print.
- Lesk, Michael. "How Much Information Is There In the World." Web. 3 Mar. 2012.
- Liddy, E.D. "Natural Language Processing." *Encyclopedia of Library and Information Science, 2nd Ed*. New York. Marcel Decker, Inc. 2003. Web. 9 July 2008.
- Liu, Ying et al. "Semantic Relatedness Study Using Second Order Co-occurrence Vectors Computed from Biomedical Corpora, UMLS and WordNet." Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. New York, NY, USA: ACM, 2012. 363–372. Web. 15 Nov. 2012. IHI'12.
- Mauer, Barry. Communication to author. April 2008. MS.
- McLean, Bethany, and Peter Elkind. *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron*. Portfolio Trade, 2004. Print.
- McLuhan, Marshall, and Lewis H. Lapham. *Understanding Media: The Extensions of Man*. Reprint. The MIT Press, 1994. Print.
- McLuhan, Eric, and Frank Zingrone, eds. *Essential McLuhan*. 1st ed. Routledge, 1997. Print.

- Menon, A.K., and B.K. Gupta. "Nanotechnology: A Data Storage Perspective." The Fourth International Conference on Nanostructured Materials (NANO '98) 12.5–8 (1999): 1117–1125.
- "Natural Language Processing - Microsoft Research." Web. 28 July 2008.
- Miller, George A. "WordNet: A Lexical Database for English." *Communications of the ACM*. 1995. Vol. 38, No. 11: 39-41.
- Montoyo, Andrés, Patricio Martínez-Barco, and Alexandra Balahur. "Subjectivity and Sentiment Analysis: An Overview of the Current State of the Area and Envisaged Developments." *Decision Support Systems* 53.4 (2012): 675–679. Print.
- Mozayani, Ashraf, and Carla Noziglia, eds. *The Forensic Laboratory Handbook Procedures and Practice*. 2nd ed. Humana Press, 2010. Print.
- "National Commission on Terrorist Attacks Upon the United States." Web. 25 Mar. 2012.
- National Research Council (U.S.). *Strengthening Forensic Science in the United States: A Path Forward*. Washington, D.C: National Academies Press, 2009. Print.
- Neuendorf, Kimberly A. *The Content Analysis Guidebook*. 1st ed. Sage Publications, Inc, 2001. Print.
- O'Barr, William M. *Linguistic Evidence: Language, Power, and Strategy in the Courtroom*. Academic Press, 1982. Print.
- Ong, Walter J. *Orality and Literacy*. 2nd ed. Routledge, 2002. Print.
- Parker, Donn B. *Crime by Computer*. 1st ed. Charles Scribner's Sons, 1976. Print.
- Parker, Donn B. *Fighting Computer Crime: A New Framework for Protecting Information*. Wiley, 1998. Print.
- Pedersen, Ted. "Enron Email Corpus by Topic." Web. 21 Feb. 2012.

- Pemberton, J. Michael. "Knowledge Management (KM) and the epistemic tradition." *Records Management Quarterly*; Jul98 vol. 32 Issue 3, 58, 4.
- Perkins, Jacob. *Python Text Processing with NLTK 2.0 Cookbook*. Packt Publishing, 2010. Print.
- Plumlee, Rick. "Wichita Man, 26, Convicted of Producing Child Porn | Wichita Eagle." Web. 2 Aug. 2012.
- Pollitt, Mark M. "An Ad Hoc Review of Digital Forensic Models," Second International Workshop on Systematic Approaches to Digital Forensic Engineering, 10-12 April 2007. 43-54.
- Pollitt, Mark. "Digital Forensics as a Surreal Narrative." *Advances in Digital Forensics V*: : Fifth IFIP WG 11.9 International Conference on Digital Forensics, Orlando, Florida, USA, January 26-28, 2009, Revised Selected Papers. Ed. Gilbert Peterson & Sujeet Sheno. Springer Boston, 2009. 3-15. Print.
- Pollitt, Mark. "A History of Digital Forensics." *Advances in Digital Forensics VI*. Ed. Kam-Pui Chow & Sujeet Sheno. Vol. 337. Springer Boston, 2010. 3–15. Web. 20 Apr. 2012. IFIP Advances in Information and Communication Technology.
- Princeton University. "About WordNet — WordNet — About WordNet." Web. 26 July 2008.
- Public Broadcasting System. "Oppression and Malice: The O.J. Simpson Civil Trial | PBS NewsHour | Feb. 5, 1997." Web. 2 Mar. 2012.
- Ray, Robert B. *The Avant-Garde Finds Andy Hardy*. Cambridge, Mass: Harvard University Press, 1995. Print.
- Reagan, Brad. "The Digital Detectives." *Popular Mechanics* May 2006 : 84–134. Web.

- Reith, Mark, Clint Carr, and Gregg Gunsch. "An Examination of Digital Forensic Models." *International Journal of Digital Evidence* 1.3 (2002).
- Roberts, Carl G. "The 2006 Discovery Amendments to the Federal Rules of Civil Procedure." *Law Practice Today*. Web. 9 Nov. 2012.
- Roussev, Vassil, and Candice Quates. "Content Triage with Similarity Digests: The M57 Case Study." *Digital Investigation* 9, Supplement.0 (2012): S60–S68. Web. 19 Sept. 2012.
- Rusu, Delia et al. "Triplet extraction from sentences." (2007): n. pag. OAster. Web.
- Saferstein, Richard. *Criminalistics: An Introduction to Forensic Science*. 10th ed. Prentice Hall, 2010. Print.
- Saussure, Ferdinand. "Course on General Linguistics." *The Norton Anthology of Theory and Criticism*. Ed. Vincent B. Leitch. Norton, 2001. 956–977. Print.
- Slay, Jill et al. "Towards a Formalization of Digital Forensics." *Advances in Digital Forensics V: Fifth IFIP WG 11.9 International Conference on Digital Forensics, Orlando, Florida, USA, January 26-28, 2009*, Revised Selected Papers. Ed. Gilbert Peterson & Sujeet Sheno. Springer, 2009. 38–47. Print.
- Stockwell, Foster. *A History of Information Storage and Retrieval*. Jefferson, N.C.: McFarland, 2001.
- Styler, Will. "The EnronSent Corpus. Technical Report 01-2011. University of Colorado at Boulder Institute of Cognitive Science, Boulder, CO." 2011. Web. 8 July 2012.
- "SWGDE_SWGIT Glossary V2.0.pdf." Web. 1 May 2009.
- Tilstone, William J. *Forensic Science: An Encyclopedia of History, Methods, and Techniques*. Santa Barbara, Calif: ABC-CLIO, 2006. Print.

- U.S. Dept. of Justice Office of the Inspector General. "The Federal Bureau of Investigation's Efforts to Combat Crimes Against Children." Web. 16 Oct. 2012.
- United States V. Timothy James McVeigh and Terry Lynn Nichols*. CR 95-110. Western District of Oklahoma. Web.
- Van Voris, Bob. "Facebook Cites 'Smoking Gun' Proof of Fraud by Man Claiming Company Stake - Bloomberg." Web. 2 Aug. 2012
- Venter, Jacobus, Alta de Waal, and Cornelius Willers. "Specializing CRISP-DM for Evidence Mining." *Advances in Digital Forensics III*. Ed. Philip Craiger & Sujeet Sheno. 2007. 303. edo. Web.
- Wax, Dustin. "Toward a New Vision of Productivity, Part 5: Drowning in Information - Stepcase Lifehack." Web. 3 Jan. 2009.
- Wecht, Cyril H., and John T. Rago. *Forensic Science and Law*. CRC Press, 2006. Print.
- Weedn, Victor W. "DNA Analysis, in Forensic Science and Law: Investigative Applications in Criminal." *Civil, and Family Justice* (2006): 418. Print.
- Wilks, Yorick. *The History of Natural Language Processing and Machine Translation*. For a new edition of the *Encyclopedia of Language and Linguistics*. Web. 22 July 2008.
- Wood, Roger. "Future Hard Disk Drive Systems." *Current Perspectives: Perpendicular Recording* 321.6 (2009): 555-561.
- Ungar, L., S. Leibholz, and C. Chaski. "IntentFinder: A System for Discovering Significant Information Implicit in Large, Heterogeneous Document Collections and Computationally Mapping Social Networks and Command Nodes." 2011 IEEE

- International Conference on Technologies for Homeland Security (HST) (2011): 219. Print.
- University of Illinois at Chicago. "Headers of a Legit Email Message." Web. 20 Jan. 2013.
- Venter, Jacobus, Alta de Waal, and Cornelius Willers. "Specializing CRISP-DM for Evidence Mining."
- Yeo, Richard R. *Encyclopaedic Visions : Scientific Dictionaries and Enlightenment Culture*. New York: Cambridge University Press, 2001.

ADDITIONAL SOURCES

- Johnson-Eilola, Johndan, and Stuart A. Selber. *Central Works in Technical Communication*. illustrated edition. Oxford University Press, USA, 2004. Print.
- Lyman, Peter, and Hal R. Varian. "How Much Information?" Web. 2 Aug. 2012.
- McKasson, Stephen C. *Speaking as an Expert: A Guide for the Identification Sciences from the Laboratory to the Courtroom*. Springfield, Ill: Charles C. Thomas, 1998. Print.
- Peterson, Gilbert, and Sujeet Sheno. *Advances in Digital Forensics VII: 7th IFIP WG 11.9 International Conference on Digital Forensics, Orlando, FL, USA, January 31 - February 2, 2011, Revised Selected Papers*. Springer. 2011. Print.
- Selfe, Cynthia L., and Richard J. Selfe. "The Politics of the Interface: Power and Its Exercise in Electronic Contact Zones." *College Composition and Communication* 45.4 (1994): 480–504. Web. 31 Dec. 2012.
- Tilstone, William J. *Forensic Science: An Encyclopedia of History, Methods, and Techniques*. Santa Barbara, Calif: ABC-CLIO, 2006. Print.

Webster, William H. "Final Report of the William H. Webster Commission on the Federal Bureau of Investigation, Counterterrorism Intelligence, and the Events at Fort Hood, Texas, on November 5, 2009." Federal Bureau of Investigation, Web. 20 July 2012.