Development of a National Repository of Digital Forensic Intelligence

Mark Weiser
Department of Management Science and Information Systems
Oklahoma State University
weiser@okstate.edu

David P. Biros
Department of Management Science and Information Systems
Oklahoma State University
david.biros@okstate.edu

Greg Mosier
Department of Economics and Legal Studies in Business
Oklahoma State University
greg.mosier@okstate.edu

ABSTRACT

Many people do all of their banking online, we and our children communicate with peers through computer systems, and there are many jobs that require near continuous interaction with computer systems. Criminals, however, are also "connected", and our online interaction provides them a conduit into our information like never before. Our credit card numbers and other fiscal information are at risk, our children's personal information is exposed to the world, and our professional reputations are on the line.

The discipline of Digital Forensics in law enforcement agencies around the nation and world has grown to match the increased risk and potential for cyber crimes. Even crimes that are not themselves computer-based, may be solved or prosecuted based on digital evidence left behind by the perpetrator. However, no widely accepted mechanism to facilitate sharing of ideas and methodologies has emerged. Different agencies re-develop approaches that have been tested in other jurisdictions. Even within a single agency, there is often significant redundant work. There is great potential efficiency gain in sharing information from digital forensic investigations.

This paper describes an on-going design and development project between Oklahoma State University's Center for Telecommunications and Network Security and the Defense Cyber Crimes Center to develop a Repository of Digital Forensic Knowledge. In its full implementation, the system has potential to provide exceptional gains in efficiency for examiners and

investigators. It provides a better conduit to share relevant information between agencies and a structure through which cases can be cross-referenced to have the most impact on a current investigation.

1. INTRODUCTION

Computer Forensics" is defined as "a sub-discipline of Digital & Multimedia Evidence, which involves the scientific examination, analysis, end or evaluation of digital evidence in legal matters" and "Digital Evidence" is defined as "Information of probative value that is stored or transmitted in binary form." [11] Taking these together or, "Digital Forensics" might be defined as "Scientific knowledge and methods applied to the identification, collection, preservation, examination, and analysis of information stored or transmitted in binary form in a manner acceptable for application in legal matters."

Digital forensics has become an indispensable tool for law enforcement. This science is not only applied to cases of crime committed with or against digital assets, but is used in many physical crimes to gather evidence of intent or proof of prior relationships. The volume of digital devices that might be explored by a forensic analysis, however, is staggering, including anything from a home computer to a video game console, to an engine module from a getaway vehicle. New hardware, software, and applications are being released into public use daily and analysts must create new and legally acceptable methods to address each of them.

Law enforcement agencies have widely varying capabilities to conduct forensics, sometimes enlisting the aid of other agencies or outside consultants to perform analyses. As new techniques are developed, internally tested, and ultimately scrutinized by the legal system, new forensic hypotheses are borne and proven. When the same techniques are applied to other cases, the new proceeding is strengthened by the precedent of prior case. Acceptance of a methodology in multiple proceedings makes it more acceptable for future cases.

Unfortunately, new forensic discoveries are rarely formally shared even within the same agency. Sometimes briefings may be given to other analysts within the same agency, although caseloads often dictate immediately moving on to the next case. Very little is shared between different agencies, or even between different offices of some federal law enforcement communities. The result of this lack of sharing is duplication of significant effort to re-discover the same or similar approaches to prior cases and a failure to take advantage of precedent rulings that may strengthen the admission of a certain process.

A need exists to create a "National Repository of Digital Forensic Information" to address these issues. Harrison, et. al., [7] proposed a repository for sharing information in 2002, but no such effort has been accepted by a significant

portion of the law enforcement community in a manner that allows previous discoveries to be best applied to future cases even within a single agency. Sharing of forensic knowledge between law enforcement agencies is almost entirely informal, and based on hearing about previous casework and contacting the case agent for more information.

We propose a design for such a repository that attempts to address many of the recognized impediments. The Center for Telecommunications and Network Security (CTANS) at Oklahoma State University is collaborating with the Defense Cyber Crimes Center (DC3) to implement a system prototype that we expect to make available to other cooperating law enforcement agencies. This paper outlines major elements of the working design and expected impediments to successful widespread implementation. Application of digital forensics extends far beyond criminal investigations. DC3, for instance, is a defense agency, so the structure of this model encompasses not only criminal matters [see Figure 1], but also forensic information for foreign intelligence and cyber needs. Approaches in media analysis and other forensic components overlap between these areas extensively, so a shared repository that can be applied in all areas will be of most benefit.

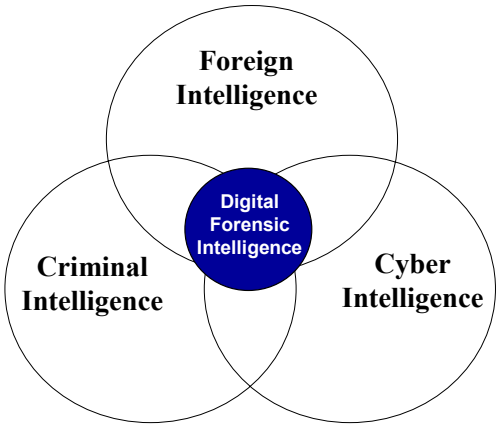

Figure 1: DC3 Digital Forensic Intelligence Model

2. WORKING DESIGN MODEL

Through interactions between CTANS and DC3, as well as other law enforcement agencies, a working design for the implementation has been developed. It allows for a modular implementation of features and a distributed structure that recognizes a varying willingness to share information between agencies. The major components are: 1) Digital Forensic Information Knowledge Base; 2) expert system and best practices for Forensic Investigations; 3) certified and available tools index; and 4) forensic case index. Each of these is briefly described below.

3. DIGITAL FORENSIC INTELLIGENCE KNOWLEDGE BASE

A "knowledge base" is typically a machine-readable repository of information. It goes beyond raw facts about a specific domain, but attempts to capture relationships between them and the context in which decisions were made. Each investigation and court proceeding are different from any that preceded them, although there are many potential commonalities. Given this, it is important to capture data, relationships, and contexts.

The knowledge base is at the core of this project. It is ultimately a type of case tracking system that stores all forensic discoveries related to a case from the time evidence is seized until the complete forensic analysis is returned to the responsible case investigator. Every law enforcement agency has slightly different procedures that they follow. Rules of evidence are similar across jurisdiction, however, so the basic process of one agency likely has more commonalities than differences with any other. Our design was modeled after the process employed by the Defense Cyber Forensics Laboratory (DCFL), which "provides digital evidence processing, analysis, and diagnostics for any DoD investigation that requires computer forensic support to detect, enhance, or recover digital media, to include audio and video. This includes criminal, counterintelligence, counterterrorism, and fraud investigations." [10]

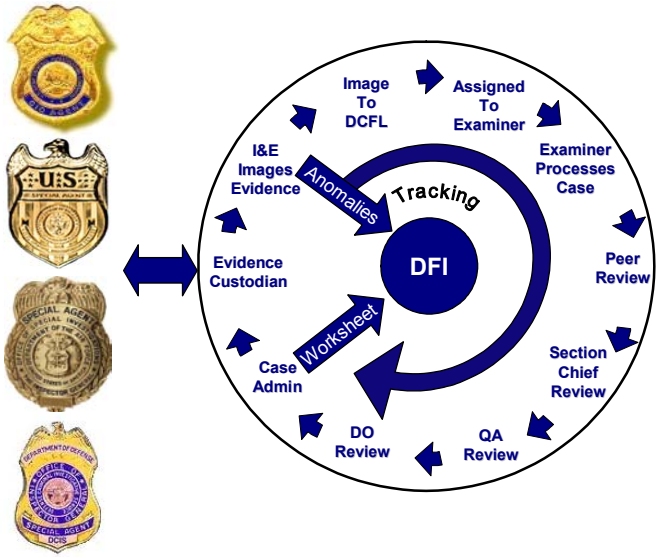

Figure 2: Cyber Forensics Investigation Model

Figure 2 graphically depicts this process. Because DCFL processes evidence for multiple agencies, they are often not involved in the seizure of that evidence, so the point of entry into their cycle is when the evidence custodian

receives the materials from any of the investigating agencies. Imaging, examiner assignment, media analysis, various reviews, and administrative actions follow.

Each of these steps is well documented and will be entered into the repository, along with scans of provided data. It will be indexed on the assigned case number, but will also have a full-text search capability to enable one method of locating related data from previous cases. A single case may now generate reams of paper reports, so a digital method to locate items within any of many reports and to eventually create an automatic cross-index of cases has great potential to aid future analyses.

4. EXPERT SYSTEM AND BEST PRACTICES

Newer examiners learn from the human experts in the lab, however, additional support is always welcome. An expert system would guide a user through more common forensic analyses with a series of questions, the answers to which will generate procedural documents and ask for input based on the results. This is not intended to replace human guidance, but may provide ideas about how to proceed in a specific case.

There are numerous articles that explain some best practices in forensics. These can then be modified and applied by an analyst as required by a particular investigation. There is no recognized central repository of best practices, although several exist, such as through the Scientific Working Group on Digital Evidence and the United States Secret Service. When these best practices are used in a case, or referenced by the expert system, they will become a part of the repository to fully explain the context and applied process for future examiners.

5. CERTIFIED AND AVAILABLE TOOLS INDEX

One of the three parts of DC3 is the Defense Cyber Crime Institute (DCCI). It provides legally and scientifically accepted standards, techniques, methodologies, research, tools, and technologies for computer forensics to meet DoD needs in counterintelligence, intelligence, information assurance, information operations, and law enforcement. A major part of that effort is to test tools and techniques in a realistic environment for their scientific validity and legal admissibility. This information is used to maintain a catalog of tools, along with the testing and analysis report for each. An independent validation of a tool prior to its application in an investigation provides enhanced credibility when presented in a legal proceeding.

This catalog is current available within the DoD and law enforcement community by request to DCCI. This prevents cyber criminals from exploiting weaknesses in forensic tools that are discovered in this process. Each item in the tools catalog has a testing and evaluation report that serves as partial justification for its use in any investigation. By including this in the repository,

a given object (along with the report) can be referenced in many different cases, without the need to include extensive and repetitive documentation across multiple cases.

There are also many tools that are available and not yet tested by DCCI. They may be used by law enforcement agencies, if the case dictates that. Each time a tool or technique is applied, that creates a record that supports its use or omission in similar future cases. Fully testing and reporting on any tool is a very time-consuming process and it is not always possible to wait for full vetting, due to time limitations on proceedings. The shared repository allows refinement and acceptability to be enhanced among many examiners and agencies, even before full testing.

6. FUSION, SEARCH AND RETRIEVAL CAPABILITIES

A shared repository is, in a sense, a database. The primary need of the repository is to build capability to fuse various cyber forensics cases into useful knowledge for the investigator. Information fusion is the process of intelligently combining the information (predictions) created and provided by two or more information sources (prediction models). Although there is an ongoing debate about the sophistication level of the fusion methods to be employed, there is a general consensus that fusion (combining forecasts and/or predictions) produces more useful information for decisions to be based upon [1]. It has been shown that fusion can improve accuracy, completeness, and robustness of information, while reducing uncertainty and bias associated with the individual predictors [3].

Once implemented, investigators can then use the repository as a data warehouse to quickly locate similar cases and capabilities. However, it is important to note that much of the information provided by investigators is in text format. Cyber forensics cases often include long written passage documenting the investigation process and the tools used. Because of this, text mining capabilities must be included in the repository.

Data Mining is the process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [6] stored in structured databases, where the data is organized in records structured by categorical, ordinal and continuous variables. However, vast majority of real world data is stored in documents that are virtually unstructured. According to a recent study by Merrill Lynch and Gartner 85 to 90 percent of all organizational data is stored in some kind of unstructured form (i.e., as text) [9]. This is where the text mining fits into the picture. Text mining is the process of discovering new, previously unknown, potentially useful information from variety of unstructured data sources including organizational documents.

Benefits of text mining are obvious in the areas where a large quantity of textual data is collected from organizational transactions. For example, free-

form text of user interactions and experiences allows trending over time in the areas of problems and complaints, which is clearly input to better equipment and system development. By not restricting the feedback to a codified form, the subject can present, in her own words, what she experiences and thinks about the domain of interest.

The common applications of text mining include Information Extraction (identifying key phrases and relationships within text by looking for predefined sequences in text via the process called pattern matching), Topic Tracking (by keeping user profiles and, based on the documents the user views, predicts other documents of interest to the user), Summarization (possessing and summarizing the document to its essence in order to save time on the part of the reader), Categorization (identifying the main themes of a document and doing so placing the document into a pre-defined set of topics categories), Clustering (grouping documents that are similar to each other without having a pre-defined set of categories), Concept Linking (connect related documents by identifying their commonly shared concepts and by doing so help users find information that they perhaps wouldn't have found using traditional searching methods), and Question Answering (deals with finding the best answer to a given question by knowledge driven pattern matching).

7. IMPEDIMENTS TO ADOPTION

There have been previous attempts to create centralized repositories for digital forensics. None have succeeded, except on a localized basis. The reasons most often cited are 1) a desire for discovering agency to completely control the data; 2) concerns about confidentiality or classification of data; 3) increased task load of entering data to support this initiative; and 4) concerns about unnecessary discovery provided to the defense or that more public information will help criminals avoid capture and/or prosecution;. This section overviews each of these concerns and provides an illustrative example of how our design for information structure leaves control of these important characteristics to the individual agencies.

7.1 Reluctance to Share Information Between Agencies

Jurisdictions of various law enforcement agencies overlap geographically. Within a single location, there may be a County Sheriff, City Police, State Police, and various federal agencies, any of which may investigate a crime depending upon the circumstances. There is a great sense of ownership of criminal case by investigators, so this overlap creates a kind of competition between the groups. Furthermore, law enforcement professionals and, more specifically, cyber security professionals tend to rely more on personal social networks rather than more formal repositories of information thus impeding information sharing in this domain [8].

This clearly extends to new systems. Individual investigators are very willing

to seek helpful information that is made available from any source, however, most have a great reluctance to release information beyond what is required. This is partially due to the aforementioned competitive nature, but also is done to protect their techniques from current and future criminals who may improve their skills with any knowledge that is available. Unfortunately, a knowledge repository will require wide input in order to leverage the knowledge of others, so this hurdle must be overcome.

The proposed system provides optional authorship recognition to investigators and agencies that contribute information that is used (and therefore linked) to another case. Cases that are repeatedly cited would be clearly recognizable as "critical" by their peers. The amount of information provided in that recognition would be up to the providing agency. However, recognition has proven to be a successful reward mechanism in the organization science literature [4]. Access to this system will be limited to DoD and law enforcement, except as is required by law. This mitigates the concern about criminals using the information to improve their own skills.

7.2 Classification Issues

Particularly in the DoD and Federal investigative agencies, some cases, or portions thereof, may be classified. In that case, the documents, evidence, and systems must be properly secured, and personnel with access must be appropriately cleared. An open sharing system is not an option in this case. Individual agencies, however, can implement instances of our system to create a knowledge base of their own classified projects, with access restrictions on a per-user basis. They may also access their own or separate systems to assist in the case on an unclassified system and network. Further, individual investigators in the organization may allow members of their personal social network to access their knowledge. The level of the access can be control by the sharing investigator.

7.3 Increased Task Load

Requiring members of investigative agencies to input data will increase their task load. The individual agencies already have information collection mechanisms. Any attempt to require investigators to input data into a central repository will increase their workload. As such, even those that would want to share information would not do it because they have other priorities. This is a problem often overlooked by well-meaning researchers who develop impressive data repositories and wonder why investigators will not contribute to their content. Initially, system data within the DC3 system is taken entirely from electronic worksheets that the analysts already use. As part of the normal case maintenance a clerk submits the entire file to the system, which automatically parses and indexes it. Our approach allows organizations to maintain their own data repositories and requires minimal increase in taskload.

7.4 Discovery Vulnerability

Reticence to share information across agencies can be driven by a variety of factors. One such factor is the concern over disclosure of practices and techniques that will be ultimately be nullified by a general awareness among the public and more specifically those committing offenses. While the security of such information may be easily protected with regard to casual observation, if disclosure is mandated as part of any court order or legal proceeding, the efficiency of some digital forensic science methodologies may be reduced. Initially, by observing appropriate protocols in the cataloging of information, this risk is minimized.

There are certain legal protections in place that also reduce the potential for disclosure of law enforcement techniques and methods including those that are related to digital forensics. For example, the Freedom of Information Act, 5 U.S.C. § 552 clearly exempts from disclosure "records or information compiled for law enforcement purposes, but only to the extent that the production of such law enforcement records or information….would disclose techniques and procedures for law enforcement investigations or prosecutions, or would disclose guidelines for law enforcement investigations or prosecutions if such disclosure could reasonably be expected to risk circumvention of the law…"

In court proceedings, discovery of digital forensic techniques by defendants in criminal cases may also be limited under the privilege recognized by the Eleventh Circuit court in United States v. Horn 789 F.2d 1492 (11th Cir. 1986). A subsequent case of United States v. Garey, 2004 U.S. Dist. LEXIS 23477, summarized that court's holding as "In general, the Eleventh Circuit and other courts applying the investigative techniques privilege have held that where the defendant has access to evidence, such as the product of the surveillance, from which a jury can determine the accuracy and validity of the surveillance equipment and techniques, the defendant has no need for the information that outweighs the government's interest in keeping it secret."

8. ANCHORED FLEXIBLE LOGICAL MESH STRUCTURE TO LIMIT IMPEDIMENTS

Every agency has different issues with data sharing and must be given the flexibility to determine the degree to which they will use data provided by others and/or contribute information about their discoveries to the community. Of course, the global benefit is maximized by everyone sharing all discoveries with all other groups, so there must be stimulus for that. Our model can be termed an "anchored, flexible, logical mesh." It is anchored on a core repository that will contain information made available to all authorized agencies without restriction. For example, the core repository may contain information on relevant laws and legal precedents that all forensics organizations may want to access. Ideally, it would house the common

knowledge that all organization would typically maintain and therefore remove the need for individual agencies to store and update the information themselves. Most participants will at least read information from the core repository. Relationships between servers are entirely flexible and up to the administrators of the servers themselves.

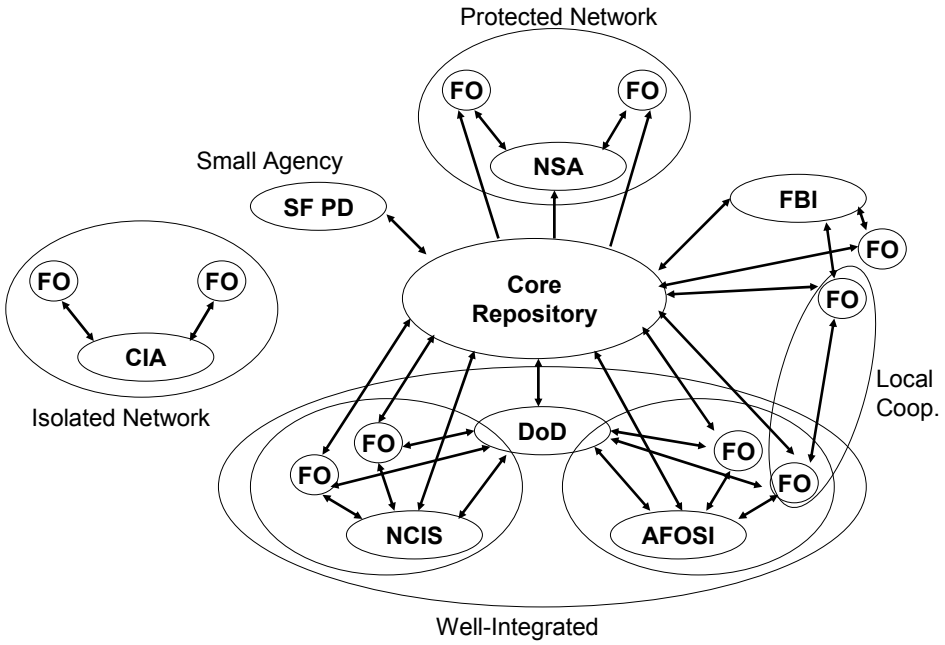

Figure 3: Anchored Logical Mesh of Repository Servers

Figure 3 shows several different examples of how this may be implemented. The diagram is not intended to reflect current or planned cooperative relationships between specific agencies that might participate in the repository. It is provided purely as a notional illustration:

- The DoD has a repository for storing information that they want to make available to their investigative agencies, but not outside the DoD, although the Naval Criminal Investigative Service (NCIS), the Air Force Office of Special Investigation (AFOSI), and their field offices can directly use and contribute to the core repository as well, or retain data only within their agency without elevating it even to the level of DoD.

- The FBI offices have a similar structure, but one of the field offices may cooperate extensively with one of the NCIS or AFOSI field

offices in the same city and liberally share new discoveries with each other. This creates a new "neighborhood" that is labeled "local coop" in the figure.

- Small agencies may have a single repository for their lessons learned, but they share with the core repository. In the extreme, there may be no local storage at all, but a web interface directly into the core repository. A small sheriff's office with a forensic capability can leverage the lessons learned in many other participating agencies with little investment.

- Some data is very sensitive. In the figure, the NSA is shown with a neighborhood among its own central node and field offices, but only as a consumer of data from the central repository. This will not benefit other agencies; however, some organization's requirements will prohibit sharing information.

- Finally, there will be some agencies that choose to be entirely isolated. They can neither benefit from the central repository nor enhance it, because of a logical and/or physical separation. The underlying system design, however, allows them to share among their own neighborhood, while retaining complete control of hardware, software, and data.

Although there are certain impediments to the building a National Forensic Repository, the literature suggests that many of these can be overcome by employing various strategies toward promoting information sharing, protecting internal investigative procedures, and providing a multi-level approach. The strategies should help mitigate agencies' concerns toward using such a system. Investigators may still rely on their social networks for information regarding a case investigation, however our approach offer a means of providing standardization to the process. It allows investigator to share information while preventing release of internally sensitive data.

9. CONCLUSION

Network technology available to the average consumer has rapidly expanded. Valuable information about many facets of our lives resides on computer systems and traverse public networks. The value of this information and the potential value of the misuse of that information create increasing motivation to criminals to commit cyber crime. Law enforcement agencies at all levels have met this challenge with new investigative techniques and digital forensic analysis to compliment their existing skills. An information repository that allows these geographically and bureaucratically diverse groups to share information about cyber crimes and digital investigation would aid every

agency in successfully and efficiently prosecuting a case.

An ongoing project between Oklahoma State University and the Defense Cyber Crimes Center aims to meet this growing need. The National Repository of Digital Forensic Information will provide a platform for tracking details of cases as they are handles and a reference system to previous investigations that might be related. It will also provide a relevant legal index to help gauge the success of various prior approaches in court and an expert system to assist investigators who are assigned to case types that are less familiar to them.

There are many non-technical impediments to widespread adoption of the system to make it most valuable. Although some of the recognized issues have been addressed in this paper, more work must be done in this area. A full cross-agency implementation of this system has the potential to greatly leverage existing examiner and investigator skills and to allow newer investigators to more quickly acquire the best approaches for successful legal proceedings.

10. REFERENCES

1. Armstrong, J.S. “Combining Forecasts”, in: J.S. Armstrong, Principles of Forecasting, Kluwer Academic Publishers, Norwell, MA., 2002, 418-439.
2. Blakeman, William. "Digital Forensic Intelligence (DFI) Project." Baltimore, MD, 15 February, 2006.
3. Chase, C.W. Jr., “Composite Forecasting: Combining Forecasts for Improved Accuracy,” Journal of Business Forecasting Methods & Systems, 2000,19, 2-22.
4. Cacioppe, R. “Using team – individual reward and recognition strategies to drive organizational success,” Journal of Leadership and Organization Development, 1999, 20 (6), pp. 322-331.
5. Defense Computer Forensics Laboratory (DCFL) website. http://www.dcfl.gov/dcfl/mission.htm. March 27, 2006.
6. Fayyad, U.M., G. Piatetsky-Shapiro and P. Smyth. “From Data Mining to Knowledge Discovery: An Overview,” in Advances in Knowledge Discovery and Data Mining, AAAI/MIT Press, 1996, 1-34.
7. Harrison, et al. “A Lessons Learned Repository for Computer Forensics,” International Journal of Digital Evidence. Fall, 2002, 1 (3).
8. Jarvenpaa, S.L., and Majchrzak, A. “Developing Individuals’ Transactive Memories of Their Ego-Centric networks to Mitigate Risks of Knowledge Sharing: The Case of Professionals Protecting CyberSecurity,” Proceedings of the International Conference on Information Systems, ICIS 2005
9. McKnight, W. “Building Business Intelligence: Text Data Mining in Business Intelligence,” DM Review, 2005, 21-22.

10. Presentation by the Defense Cyber Crime Center, March 2005
11. "SWGDE and SWGIT Glossary of Terms," Scientific Working Groups on Digital Evidence and Imaging Technology. Version: 1.0 , July 25, 2005.