available at www.sciencedirect.com

ScienceDirect

journal homepage: www.elsevier.com/locate/diin

Digital Investigation

Bringing science to digital forensics with standardized forensic corpora

Simson Garfinkel[a,b,*], *Paul Farrell*[a], *Vassil Roussev*[c], *George Dinolt*[a]

[a]*Graduate School of Operational and Information Sciences, Department of Computer Science, Naval Postgraduate School, Monterey, CA 93943, USA*
[b]*Harvard University, USA*
[c]*University of New Orleans, USA*

Keywords:
Forensics
Human subjects research
Corpora
Real data corpus
Realistic data

ABSTRACT

Progress in computer forensics research has been limited by the lack of a standardized data sets—*corpora*—that are available for research purposes. We explain why corpora are needed to further forensic research, present a taxonomy for describing corpora, and announce the availability of several forensic data sets.

1. Introduction

Much of the work to date in digital forensics has focused on data extraction and for presentation in courts. Researchers have developed technologies for copying data from subject hard drives, storing that data in a disk image file, searching the disk image for document files, and presenting the documents to an examiner.

As both the variety and scale of forensic investigations increase, forensic practitioners need tools that do more than search and present: they need tools for reconstruction, analysis, clustering, data mining, and sense-making. Such tools frequently require the development of new scientific techniques in areas such as text mining, machine learning, visualization, and related fields.

One of the hallmarks of science is the ability for researchers to perform controlled and repeatable experiments that produce reproducible results. Science is based on the principle that phenomena can be observed and results can be reproduced by *anyone*—there are no privileged experimenters or observers (given sufficient training and financial resources, of course).

Sadly, much of today's digital forensic research results are not reproducible. For example, techniques developed and tested by one set of researchers cannot be validated by others since the different research groups use different data sets to test and evaluate their techniques.

1.1. Why forensic corpora are needed

Having a reference set of representative corpora enhances the scientific evaluation of forensic methods beyond the obvious benefits of providing ready test data and enabling direct comparison of different approaches. Namely, it allows for the ground truth to be established using manual or otherwise time-consuming methods. Such results can then be used as a baseline to evaluate the success of new tools and methods using objective metrics.

In the digital forensics field there have been sporadic efforts to produce standardized corpora, mostly in the form of forensic challenges. The main goal of these challenges has been the development of practical tools for problem areas in need. Since 2005, DFRWS has issued an annual challenge

* *Corresponding author*. Graduate School of Operational and Information Sciences, Department of Computer Science, Naval Postgraduate School, Monterey, CA 93943, USA.
E-mail address: slgarfin@nps.edu (S. Garfinkel).
1742-2876/$ – see front matter
doi:10.1016/j.diin.2009.06.016

Table 1 – Size of the US and Non-US drive corpus.

Corpus	HDs	Flash	CDs	GB[a]
US Corpus	1258			2939 GB
Non-US Corpus:				
BA	7			38 GB
CA	46	1		420 GB
CN	26	568	98	999 GB
DE	37	1		765 GB
GR	10			6 GB
IL	152	4		964 GB
IN		66		29 GB
MX	156			571 GB
NZ	1			4 GB
TH	1	3		13 GB
Total Non-US:	1056	643	98	3723 GB

a Uncompressed.

focused on specific topics: Windows memory analysis, Linux memory analysis, and data carving. The DoD Cyber Crime Center (DC3) has also been issuing annual challenges, which have had a broader scope, including data recovery from damaged hardware and media.

These challenges have spurred research and development in the focus areas and have brought excitement and tangible results to the field. Yet the scope of these challenges is much too limited to support digital forensics research and tool validation on a larger scale. In particular, it would be difficult to argue that a particular method has undergone rigorous evaluation just because it is able to solve a specific challenge.

Today the results from mainstream commercial tools are frequently accepted based solely on the reputation of the vendor, which in turn is frequently based solely upon name recognition. Although anyone can perform an independent evaluation of today's tools, such work is challenging without the availability of test data that can be readily shared.

Looking forward, the deluge of data that must be analyzed will continue to grow for the foreseeable future. This will likely necessitate the development of statistical and other approximation techniques. It is imperative that we have a large, representative sample of data that has been processed with exact techniques and well-characterized so that we can have confidence in these approximations. This ultimately benefits society as a whole, given the increasing importance of digital evidence in legal proceedings: by ensuring that the interpretation of digital evidence is grounded in facts and solid science and not simply upon opinion.

Another benefit of reference corpora is that they can become of the focus of investment for the entire community. By becoming community property, reference sets enables more efficient use of limited research resources. Such sets also give funding agencies a framework for advancing the field as a whole, whether through constructed ''challenges'' or by using the corpora to help establish and quantify milestones and reference capabilities.

From a training and educational perspective, it is difficult to overstate the need for realistic data sets. Anyone who has been on the instructors side of the process will testify to the huge investment of time that goes into creating realistic forensic scenarios. Much of this work is not shared broadly and that is clearly inefficient and wasteful in a relatively small field with limited budgets. The creation of common corpora can start and stimulate the process of accumulating and sharing such data.

1.2. *Contributions and paper outline*

With this paper we present justification for the creation of large-scale standardized forensic corpora (Section 2), and introduce a taxonomy for understanding the corpora that have been created to date (Section 3). We announce the availability of four corpora for research and educational purposes (Section 4). We share lessons learned (Section 5). Finally we present related work (Section 6) and conclude.

2. Forensic reproducibility

Despite the importance of reproducibility for the scientific process, there have been few attempts to enable digital forensics researchers to produce reproducible results. We suggest that this lack of attention to reproducibility is a result of the manner in which digital forensics has evolved and the nature of forensics data.

2.1. *Reproducibility in science*

In recent years the popular media has portrayed reproducibility as the primary means by which scientists validate each other's results and combat scientific fraud. While these are important benefits, reproducibility has a far more mundane though important role in day-to-day lives of scientists.

Fundamentally, the reproducibility of scientific results makes it possible for groups of scientists to build upon the results of others. This is especially true in experimental sciences, where observations can frequently outstrip the ability of theory to explain them. Reproducibility makes it possible for one researcher or research group to validate that they have mastered a technique and then to go off in a different direction. In biology, reproducibility is so important that researchers will routinely trade cell lines and DNA samples, and even apprentice in each others' labs, so that techniques and knowledge can be diffused throughout the field.

Reproducibility also has an important role in the development, sale and use of scientific instrumentation. Reproducibility allows equipment from different vendors to be calibrated against objective measures. Here the need for reproducibility goes hand-in-hand with the commercial availability of *scientific standards*—for example, weights of known mass, solutions of known concentration, and sealed glass vials of known composition.

Reproducibility has a critical role in education as well. Students learn and validate their mastery of scientific techniques by performing experiments with known outcomes. Without reproducibility there can be no objective evaluation of student work.

2.2. *Reproducibility and forensic practice*

Digital forensics has evolved over the past two decades to solve the real-world needs of criminal investigators. As discussed above, much emphasis, until now, has been on *evidence preservation* and *data presentation*. The legal standard that forensic tools must pass in order to be usable in this context is the Best Evidence Rule, and specifically Rule 1001(3) of the US Federal Rules of Evidence, which holds that the information shown to the Court (or used as the basis for testimony by an expert) must ''reflect data accurately (Federal rules of evidence, 2008)''.

This requirement for ''accuracy'', together with traditional forensic notions of evidence preservation, is largely responsible for the forensic community's standardization on cryptographic hash functions (e.g. SHA1) to detect possible alternations in the evidence. It is also responsible for the practice of reporting ''the precise address (physical cluster, sector offset, etc.)'' of data recovered from slack space on a disk when making a forensic report to a court or opposing counsel (Patzakis, 2001).

This ''accuracy'' standard is surprisingly low: for most purposes it is sufficient to show that a tool does not alter evidence and that it faithfully reports the precise addresses in order to certify the tool for use in a court room. Likewise, it is assumed that experts working for the defense would be given an ''accurate'' copy of the prosecution's evidence so that the scientific conclusions of the forensic analysis could be replicated. But the forensic community has been exceedingly slow to adopt performance requirements and standards for forensic software. For example, the NIST Computer Forensic Tool Testing Program did not create a draft ''Forensic String Searching Tool Requirements Specification'' until January 2008 (Computer Forensic Tool Testing Program, 2008), and the final version of this specification has yet to be adopted.

2.3. *Reproducibility and forensics research*

Establishing reproducibility of research findings is considerably more difficult than establishing the reproducibility of a forensic investigation in a specific criminal or civil case. Forensic tools and techniques, by their very nature, are operate on data sets that are large, generated by people, and are intensely personal. This leads to several practical and legal problems:

- Because the data sets are generated by people, their use by any research that is funded by the US Government or takes place within most universities is governed by the HHS Common Rule (45 CFR 46), Institutional Review Boards (IRBs), Research Ethics Committees, or some other form of institutional oversight.[1] By design, such oversight creates additional administrative barriers which must be satisfied prior to the use of human subject data.
- Rather than go to the problem of collecting data from research subjects and working within the formalized institutional oversight process, many researchers simply use their own data (network packets, disk images, *etc.*) in their experiments.[2] This data lacks the diversity and unpredictability of real data, compromising the research findings.
- Because research data typically contains personal or proprietary information, experimenters are typically not willing to share their experimental data with others. Because the data is not shared, there is no way for other experimenters to replicate the results.
- Experimenters who *do* work through the IRB process necessarily face procedural hurdles and administrative delays when they seek to share their data sets with others—especially when partnering with researchers at organizations that do not have IRBs.

Each of these issues has negatively impacted the progress of digital forensics research.

Consider the file identification problem. Starting with McDaniel's (2001) master's thesis, there have been slightly more than a dozen papers that have concerned themselves with the problem of identifying files using headers, footers, or fragments taken from the middle. For example, Moody and Erbacher (2008) report an accuracy rate of 72% for JPEG files; they state that this is an improvement over the 43.83% that McDaniel et al. report for the same problem. Karresand and Shahmehri report 97.90% true positives and 99.99% true negatives. Calhoun and Coles report accuracy rates ranging from 83% to 99% (Calhoun and Coles, 2008). But none of these results are comparable because none of them used the same data sets! Worse, because the set used by each group is not publicly available, anyone seeking to re-implement and improve the algorithms is handicapped—there is no way to tell if the algorithm is properly implemented!

Another example is Deolalikar and Laffitte's system for combining file system metadata with content analysis to automatically determine when source documents were edited to create second-generation documents (Deolalikar and Latte, 2009). The documents used for the published paper were proprietary documents from Hewlett Packard Labs. As a result, the author's paper and presentation had to be sanitized—removing critical information that would have allowed better evaluation and analysis—and it is not possible for other researchers to obtain the same corpus to see how other reconstruction techniques compare to the experimenters' published results.

2.4. *Digital forensics education*

The lack of readily available data sets has also been problematic for digital forensics education. Without standardized data collections, educators are forced to spend significant time creating their own data sets or to instruct students to get their own data to by analyzing their own systems, the systems of friends, by making purchases of used storage media on the second hand market, or (in the case of network forensics) by eavesdropping on open Wi-Fi access points.

[1] This paper uses the term IRB to denote any formalized institutional oversight process.

[2] Note that these researchers implicitly assume that self-experimentation on their own human subject data does not require IRB approval. However there is no exemption in the IRB regulations for self-experimentation.

After several years of teaching computer security and digital forensics at the undergraduate and graduate level, we have concluded that it is inappropriate to use real data such as this in a classroom environment:

- Real data may contain information that is privacy-sensitive and may even be legally protected, including personal email, financial records, academic records, and stored passwords. Using this data in the classroom environment—even for homework assignments—creates the possibility that confidential information may be inappropriately disclosed.
- Although the real data itself may not be protected, it may be illegal for students to obtain the real data.
- Real data may contain content that is itself illegal (e.g. obscene images and child pornography) or that is illegal to distribute to minors (e.g. pornography).
- Because personally owned computer systems contain highly confidential data, a professor cannot ethically ask students to share their data with one another. Even students working in groups risk divulging to fellow group members highly personal data (email messages, photos, passwords to websites, and event graded assignments) to other group members if they use their own hard drives for analysis.
- Students performing an analysis on their own computers know in advance what they are going to find—there is no element of surprise. Additionally, a single student computer lacks diversity. Thus, students who are limited to analyzing their own systems suffer a compromised educational experience.
- Because the data is generated by human beings, use of real data in research is classified as human subjects research under 45 CFR 46 (the ''Common Rule'') and requires approval by an Institutional Review Board (Garfinkel, 2008). Such approval typically takes weeks or months and cannot be reasonably performed within the context of most undergraduate courses.

These scenarios are not the result of idle speculation. Garfinkel has purchased more than 2000 devices on the secondary market. One of these devices contained 30,000 patient billing records from a medical facility in Florida. Another hard drive was previously used in an ATM machine and still had several thousand transactions on it. Several devices purchased on the secondary market contained pornography (Garfinkel and Shelat, 2002). Later, in a class that Garfinkel taught at Harvard Extension School, students were invited to create forensic reports of USB devices borrowed from friends; in one case a device that was purchased as ''new'' contained photographs from a previous user—apparently the ''new'' device had been previously sold, used, returned, and re-sold. Fortunately the device did not contain any illegal content.

It is theoretically impossible to review a hard drive in advance and determine if that drive contains content that is inappropriate or illegal—if it were possible to do this, Digital Forensics would be a solved problem! Consider a USB storage device purchased on the secondary market that contains a single photograph. A professor examining the file might see a flower. But a student examining the same photo might discover that the file contains a hidden image that is protected with steganography and cryptography. Upon performing a keyword search the student discovers the password—and upon decrypting the embedded image, the student discovers illegal content. There is simply no way that an educator can protect students by pre-screening forensic data sets acquired in the wild.

Undergraduates working on year-long research projects or graduate studies have the time to be properly trained in the handling of human subject data. It may be appropriate to give these students access to controlled corpora of real data from real users. But there is no reason why the average undergraduate enrolled in an introductory or upper-level computer forensics course needs to be working with data for which the content is not known.

3. Corpora characterization

Our proposal is to establish curated, standardized corpora that will be available for use in research, tool testing and education.

3.1. *Corpora modalities*

We envision many different kinds of corpora:

Disk Images are the most fundamental kind of forensic corpora because of their long-established use in the field of forensics and because of their general usefulness.

Memory images are urgently needed for the development of both forensic tools and forensic training. Ideally a memory image corpus would include images from multiple versions of multiple operating systems.

Network Packets have already been productively included in corpora. Packet corpora can consist of traffic from one or more individual systems or networks.

Files can be productively collected and distributed as corpora. As mentioned in Section 2.3, there has been considerable work on file and file fragment identification which would have benefited from standardized corpora of files. Work on metadata and text extraction would also benefit from such corpora. Although files can certainly be extracted from disk images, distributing files as stand-alone corpora significantly simplifies the effort for the intended users.

In many cases it is useful for a corpus to contain *time sequences*—for example, a single disk that is repeatedly imaged during the course of operations. It is also useful to have a multi-modality corpora—for example, disk images and matching network packets or memory images.

3.2. *Corpora sensitivity*

In addition to differing modalities of corpora, we believe that there is a need for corpora containing both sensitive and non-sensitive information. To this end we have developed the following taxonomy:

Test Data is specifically constructed for the purpose of demonstrating a specific forensic issue or testing a specific feature in a tool. An example of this is the Russian Tea Room floppy disk image (Lyle, 2008b) in the NIST Computer Forensic Reference Data Sets (Lyle, 2008a), which is used to validate Unicode search and display capabilities. Test data should be contain sufficient data to demonstrate or validate the specific item being tested but be otherwise simple and uncluttered with additional information. Test data can be freely distributed on the Internet without any controls.

Sampled data is obtained by selecting a subset of a large data source, such as the Internet, using some kind of randomized process. The essential idea is to eliminate bias that may come from the use of a researchers own data collection (e.g. documents or images from the researcher's personal hard drive). However, if true random sampling technique is employed, it becomes difficult to publish the set as it is impractical to ascertain that none of the files have any legal restrictions on redistribution.

Realistic data is similar to what a forensic investigator might encounter in an investigation, but the data set was in fact artificially constructed. Realistic data is typically created by performing clean installations of software on wiped machines. At this point the experimenter can run programs, perform basic operations, or even engage in sophisticated role play with other experimenters. Examples of Realistic data are boomer-win2k (Kornblum, 2007), a memory image of a Windows 2000 SP0 computer that is part of the Digital Forensics Tool Testing Images (Carrier, 2007), and the DARPA Intrusion Detection Data Sets (2000). Although there should be no privacy concerns when distributing realistic data, there may be copyright concerns.

Real and Restricted Data is created by actual human beings during activities that were not performed for the purpose of creating forensic test data. Access to this data should be controlled: it should not be placed on the Internet for anonymous download. Real data is typically subject to restrictions because of privacy or copyright concerns.

Real but Unrestricted data sets can be (or have been) made available for unrestricted access. For example, the Enron Email Dataset (Klimt and Yang, 2004) is a corpus of 619,446 real email messages from the 158 users inside the Enron Corporation. These email messages were entered as evidence in a court case by the US Government and, as a result, became publicly available without restriction. Another example of real but unrestricted data are photos that can be downloaded from the Flickr photo sharing website and user profiles on Facebook.

3.3. *Restrictions on corpora use*

Whether or not the distribution of a corpus is ''restricted'', the use of the corpus may still be legally governed within an organization as a result of the Common Rule. Although 45 CFR 46 specifically exempts ''observation of public behavior'' and research of ''existing data, documents [and] records'', most universities require that the determination of exemption be made by the IRB, not by the individual experimenter. Additionally, the regulations do not allow exemption if the data subjects can be identified. This is a problematic distinction, as the identifiably of data subjects is not just a function of the data in question, but also a function of additional databases available to the researcher and the researcher's technical skill.

3.4. *Describing corpora with metadata*

Currently there is no standardized metadata or schema to describe forensic corpora or elements within a corpora. For example, the HoneyNet Project has distributed 34 different ''Scans'' of disk images, memory images, packet traces, and other information. Not only are each of the scans distributed in different formats—there is not even a consistent schema for talking *about* the Scans. Instead, each ''scan of the month'' has a web page, and even these pages lack consistency (Honeynet, 2009).

One approach for characterizing corpora and corpora objects is to use the schema developed by the Dublin Core Metadata Initiative (2009). The Simple Dublin Core Metadata Element Set (DCMES) is a set of 15 elements: Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language, Relation, Coverage and Rights. Although many of these elements may not be appropriate to all forensic corpora, using DCMES for corpora seems a reasonable alternative to having the forensic research community develop its own, incompatible metadata framework. For example, we note that the National Science Digital Library specifies the use of Dublin Core for contributed data sets (The National Science Digital Library, 2009).

4. Available corpora and data sets

This section describes a number of data sets that we have developed and are making available for digital forensics research. For each set we explain the motivation, the content, and whatever restrictions are being imposed on the distribution and use.

4.1. *A real but unrestricted file corpus*

In recent years a significant amount of forensic research has involved the analysis of files or file fragments. In the absence of such corpora, researchers and students who wish to work with files first need to collect files—a surprisingly difficult task if one wishes a large number of files of many types from a variety of sources. Although many files can be freely downloaded from the web, building and running a high performance document discovery and downloading tool is not a trivial task. Once files are downloaded they need to be analyzed, characterized and curated. Finally, many corpora that might be assembled cannot be easily redistributed due to privacy or copyright concerns.

For these reasons, we have created and released a corpus of 1 million documents that are freely available for research and may be (to the best of our knowledge) freely redistributed. These documents were obtained by performing searches for words randomly chosen from the Unix dictionary, numbers randomly chosen between 1 and 1 million, and randomized combinations of the two, for documents of specified file types

that resided on web servers in the .gov domain using the Yahoo and Google search engines.

Each file in the corpus is presented as a numbered file with the original file extension (e.g. 0000001.jpg). They are distributed as a set of 1000 directories, with 1000 files in each directory.

We are making this corpus of 1 million documents available in several forms:

- As a set of files in an EXT2 file system delivered on a 1TB SATA internal hard drive.
- As a set of 1000 ZIP files (1000 files in each ZIP file) that can be downloaded from our web server.
- As a set of 10 subsets (''thread 0'' through ''thread 9''), each containing 1000 randomly chosen documents. These subsets were specifically created for to facilitate pilot studies and student research projects with the rationale that it's easier to work with 1000 files than with 1 million. Students are encouraged to use one subset for development and another subset for testing.
- As a set of 1 million files that can each be downloaded from our web server using a file-specific URL.

The following metadata is provided for each of the files:

- The URL from which the file was downloaded.
- The date and time of the download.
- The search term that was used.
- The search engine that provided the document.
- The length and SHA1 of the file.
- A Simple Dublin Core for the file. An example of such a record appears in Fig. 1.

The metadata is distributed as a tab-delimited file and as an SQL database dump. We have also created a Simple Dublin Core record for the entire corpus.

The entire corpus can be downloaded from our web server, http://domex.nps. edu/corp/files/govdocs1/.

4.2. *Test disk images*

We have created and have made available two test disk images for computer forensic tool testing and education:

nps-2009-hfsjtest1[3] A test image of a journaled HFS + system in which the data from a previous version of one of the files can only be recovered from the HFS + journal. Although the presence of this data can be verified on the disk using a hex editor, no forensic tool that we are aware of can attribute the data to the file from which it came.
nps-2009-ntfs1 A test image of an NTFS file system including unfragmented and highly fragmented files stored in raw, compressed, and encrypted directories. The decryption key for the encrypted files is provided in the root directory of the NTFS file system.

These images are available for download from our Test Disk Corpora website, http://www.digitalcorpora.org/.

[3] The full name of the file system has been blinded for review.

4.3. *Realistic disk images*

We have also created and have made available four realistic disk images:

nps-2009-canon2 is a set of six FAT32 forensic images created during a typical use of a Canon PowerShot SD800IS digital camera. The images were made by placing an SD card into the camera, taking photos, removing the card, erasing select photos, imaging the card, and then repeating the process. Some of the JPEGs are fragmented, some are not. Some are resident in the file system, some are deleted but recoverable, and several have data present but no longer have any file system metadata and can only be recovered through carving. Of these carvable JPEGs at least two are fragmented. This image was created to test and teach basic file recovery, fragmented file recovery, and file carving.
nps-2009-ubnist1 is a set of images made from a USB memory stick that contains a bootable copy of Ubuntu 8.10 Linux. Over the course of several weeks the image was repeatedly booted in Linux, used to browser several US Government websites, and then shut down and imaged. This image contains a boot loader and a FAT32 file system.
nps-2009-casper-rw The ext3 file systems extracted from the nps-2009-ubnist1 USB images. Although these file systems can be extracted from the FAT32 file systems by the user, it is somewhat easier to have the EXT3 file system provided as a separate disk image. (The word ''casper'' is the name of the file on in the FAT32 system that houses this file).
nps-2009-domexusers This is an NTFS file system of computer running Windows XP containing two user accounts. Over a course of several days, an experimenter playing the role of two users exchanged instant messages and emails with a third user that resided on a separate system. The two accounts received, edited and saved office document files as well as various media files. Some of these files were then deleted. Email and instant messenger conversations were saved locally on the system. The accounts also visited web pages for news and webmail. Details of the precise method by which this disk image were prepared can be found in another publication (Farrell, 2009). This image has been redacted with a special redaction tool (Garfinkel, 2009a) that scrambles the instructions from the Microsoft Windows executables but leaves the strings untouched. This allows analysis of the DLLs but prevents the image from being used to run Windows without a license, which believe is sufficient redaction for the purpose of distributing the disk image under the ''fair use'' provisions of the US Copyright Act.

These images are also available for download.

4.4. *The real data corpus*

The Real Data Corpus (RDC) is a collection of raw data extracted from hard drives, flash memory cards, and other data-carrying devices that were purchased on the secondary market around the world (Table 1). Many studies have shown that hard drives, cell phones, USB memory sticks, and other data-carrying devices are frequently discarded by their original users without the data first being cleared or purged. By

```
<?xml version="1.0" ?>
<metadata
  xmlns="http://domex.nps.edu/corp/files/govdocs1/"
  xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
  xsi:schemaLocation="http://domex.nps.edu/corp/files/govdocs1/
                      http://domex.nps.edu/corp/files/govdocs1/schema.xsd"
  xmlns:dc="http://purl.org/dc/elements/1.1/">

  <dc:description>
    Search term: Tenino.
    Search engine: http://www.yahoo.com
    Filesize: 40960.
  </dc:description>
  <dc:date>    2009-02-06 01:12:26 GMT. </dc:date>
  <dc:format> doc </dc:format>
  <dc:identifier> 000001 </dc:identifier>
  <dc:source> http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-2340A1.doc </dc:source>
</metadata>
```

Fig. 1 – The Simple Dublin Core record for File #000001 in the million file corpus.

purchasing these devices and extracting their data, we have created a data set that has much of the diversity of drive data that exists in the real world. For example, drives in the RDC come from many operating systems, but they are predominantly from Windows-based computers. There is a wide range of Windows variants, as well as a wide selection of application programs that were used to create the data files. Many of the programs are from off-the-shelf and shrinkwrapped applications, but there is also a large selection of custom applications. Some of the disks contain default installations of Windows and not much else; others are awash in personal information.

There are, nevertheless, important differences between the RDC drive images and those in the real world. First, while drives seized during the course of police investigations tend to be working, a significant number of hard drives sold on the secondary market are malfunctioning in some way—otherwise why would they have been sold? We thus see much higher failure rates with drives in the RDC that those in typical police work. As a result, many of the disk images are incomplete—many have data at the front of the disk and at the back of the disk, but are missing data in the middle where presumably there was some kind of disk failure.

Some of the disks in the RDC contain all of the data that was on the drive when it was taken out of service. On others there was some attempt made at sanitization—in some cases files were deleted, in other cases the file system was formatted. In some cases the entire file system was actually erased or blanked. Rather than purge the data set of these devices, we keep them as part of the set for external validity. For example, having disks that have had various sanitization attempts allows us to develop software that diagnoses the manner in which sanitization was attempted.

4.4.1. *RDC size: US vs. non-US*

Because of restrictions imposed on some researchers within the US Government, we make available two different versions of the Real Data Corpus. The US Persons Real Data Corpus contains images of disk drives purchased inside the United States, while the Non-US Persons Real Data Corpus contains data from devices that have been purchased outside the US. The term ''Real Data Corpus'' (RDC) is used to describe the union of the two corpora.

4.4.2. *XML index*

Each image file in the RDC is distributed with an XML file that contains information about the disk from which the image file was created, the partitions that were found on the disk, and all of the files in the partition that can be recovered using SleuthKit.

Fig. 2 shows the first 32 lines of the XML file generated from disk image ubnist1.gen0.raw discussed earlier. All of the XML is located inside an <fiwalk> block (fiwalk stands for ''file and inode walk). The XML starts with tags that describe which version of fiwalk, Sleuthkit, and AFFLIB were used to create the XML file; this allows new XML files to be automatically generated by our system when the tools are upgraded. This outer XML block can also contain information about the disk itself, such as the serial number of the ATA disk from which the image was made.

The <volume> block is repeated for each volume that is discovered inside the disk image. Typically there is one volume per file system. File system parameters such as the block size, file system type, and size (in blocks) is reported.

Finally, a <fileobject> block is reported for each file that can be recovered. The information in this block is the information that is extracted by SleuthKit. The primary advantage of having this information in XML description is that more people know how to read XML than know how to either read SleuthKit's text output formats or who know how to link the SleuthKit library in to their applications. Furthermore, unlike SleuthKit, the information is designed for extreme usability: this is why the <byte_runs> tags, which reports the location of each fragment in the file, are reported from *both* the beginning of the file system and the beginning of the physical disk image.

Using the information in the XML description it is possible for another program to determine which files are present in the disk image and to directly extract the contents of the files without relying on additional programs such as SleuthKit, EnCase or FTK. (Note: it is currently not possible to extract files that are compressed without using SleuthKit's icat command,

```
<?xml version='1.0' encoding='ISO-8859-1'?>
<fiwalk>
 <metadata>
    <dc:type>Disk Image</dc:type>
 </metadata>
 <Imagefile>/corp/ubnist1.gen0.raw</Imagefile>
 <fiwalk_version>0.5.1</fiwalk_version>
 <Start_time>Sun Mar 8 22:13:10 2009</Start_time>
 <tsk_version>3.0.0</tsk_version>
 <aff_version>3.3.4</aff_version>
 <volume offset='32256'>
  <Partition_Offset>32256</Partition_Offset>
  <block_size>512</block_size>
  <ftype>8</ftype>
  <ftype_str>fat32</ftype_str>
  <block_count>4114340</block_count>
  <first_block>0</first_block>
  <last_block>4114339</last_block>
  <fileobject>
   <id>1</id>
   <filesize>14607</filesize>
   <partition>1</partition>
   <flags>5</flags>
   <ALLOC>1</ALLOC>
   <USED>1</USED>
   <inode>4</inode>
   <type>1</type>
   <mode>73</mode>
   <nlink>1</nlink>
   <uid>0</uid>
   <gid>0</gid>
   <mtime>1230525210</mtime>
   <atime>1230451200</atime>
   <crtime>1230525210</crtime>
   <filename>ldlinux.sys</filename>
   <byte_runs>
     <run file_offset='0'
     fs_offset='4127744'
     img_offset='4160000' len='14607'/>
   </byte_runs>
   <md5>a40ba2f7239bdae2193dfd1089856f38</md5>
  </fileobject>
```

Fig. 2 – The first few lines in XML file created from ubnist1.gen0.raw; lines have been indented for clarity.

and SleuthKit does not currently support files that are encrypted using Microsoft EFS).

The XML files make it dramatically easier to work with the disk images, since it is easy to scan the XML to see if a file is present or absent. The use of XML, in preference to SleuthKit's native vertical-bar delimited format, allows the XML-generating tools to be upgraded and the XML to be annotated without modifying tools that ingest the XML.

We have also used the XML to support a remote access methodology. We have made the XML files available on a password-protected secure web server. These files can then be downloaded by an intended consumer of the files. The consumer can scan the files for files of a specific name or hash code. The consumer can then issue an XMLRPC call to our secure server and request specific blocks of a disk image. Using this methodology one of our research partners has searched the RDC for specific files and downloaded just the XML metadata files and then the specific files within the disk images that were of interest.

4.4.3. *RDC uses*

To date the RDC has been used for a number of projects, including:

1 Developing and validating forensic and data recovery tools. (Numerous bugs in The SleuthKit have been discovered by processing all of the RDC disk images with SleuthKit.)
2 Exploring and characterizing real-world computing practices, configuration choices, and option settings.
3 Studying the storage allocation strategies of file systems under real-world conditions.
4 Developing novel computer forensic algorithms.

4.4.4. *Access, availability and restrictions*

The RDC is available to qualified research collaborators as a set of encrypted AFF files. Encryption is with AES 256 and can be based on either a pass phrase or X.509 PKI using AFF encryption (Garfinkel, 2009b). The corpus can be obtained through a variety of modalities, including:

1 Disk images can be downloaded over the Internet from a secure server using SSL by authorized researchers.
2 Individual files from the corpus can be copied onto a 3.5'' SATA hard drive (Mac HFS or EXT2 format).
3 Researchers can be given an account on a multi-user Linux computer on which all of the corpora resides.
4 The remote access methodology can be used to access individual files in the corpus.

Because the information in RDC comes from real people, we require that all intended users obtain approval from their IRBs and provide us with a copy of both the IRB application and the approval letter.[4]

5. Lessons learned

This project ended up being much harder than we original suspected.

The first and most difficult aspect of this project has been working with the large size of forensic files. Although these days a 1TB hard drive can be purchased for less than $100, it is still quite difficult to work with a large number of disk files in the 10–100 GB range. Simply moving the files from system to system was a slow and tedious process, compounded by slow data transfer rates, failing hard drives, minor data corruption issues, and constantly running out of space on target devices. It would be very nice to have a high-availability persistent file store which offered a globally addressable name space and high performance access speeds, but no such system currently exists.

[4] Strangely, one potential collaborator was told by the legal department at his university that he could not share his IRB application with us because it was ''university property''. Because the approval letter simply said that the protocol had been approved without explaining the protocol that had been approved, we were unable to work with the collaborator.

We have adopted the following strategy for working with disk images which seems to work quite well:

- Whenever possible, a single disk image should be stored in a single file.
- We have one master server which has the master copy of each disk image.
- No two disk images should have the same file name, even if they are in different directories.
- When files are moved from system to system, the path names should not change. This allows the same scripts to be run on every system without change.
- Instead of using the rm command, we wrote a Python script that only erases a file if there is already a copy of the file on the master server.

In working on the million document corpus, we were frustrated by the decision of the .gov administrator to open the domain up to US States and Local governments, but then to refuse our requests for a list of non-federal domains that had been admitted. As a result, we were forced to manually reviewed all of the domains and removed documents from non-federal web servers. Of course, due to the size of the corpus, it was not possible to manually review each document.

We were also frustrated by web servers which claimed to be offering files up using one MIME type but actually delivered a document that was coded in another. We discovered that we needed to scan for duplicates at all stages of processing—for example, suppressing duplicate URLs, but also computing the SHA1 of each document and dropping it from the database if another document with the same SHA1 was already present. (Typically, this happened because web servers were configured to give HTML error pages served without a 403 error codes).

Finally, we were frustrated by the Yahoo search API, which uses a different API for searching for documents than for images, and by the inability of Yahoo's API to search for arbitrary document types.

6. Related work

With substantial funding from the Defense Advanced Research Projects Agency, MIT Lincoln Labs created a test network that simulated a US Air Force Base and the external Internet. Several hundred megabytes of packets (compressed) were captured representing both normal traffic and attacks. The results were used as the basis of the DARPA 1998, 1999 and 2000 Intrusion Detection Evaluation programs (Cunningham et al., 1999).

The MAWI Working Group of the WIDE Project has created a Traffic Archive with many packet traces of the trans-Pacific data links (Mawi working group traffic archive, 2009). This archive is of limited use since the IP traces are ''scrambled by a modified version of tcpdpriv''. The data payloads have also been removed. Nevertheless the authors warn that ''actions that trespass upon users' privacy are prohibited''. One of the most useful corpora to have been released to the forensics community is the Enron Corpus (Klimt and Yang, 2004). This data set is useful because of its depth and because, unlike other corpora, it is largely unredacted. A list of more than 20 different corpora that can be of use in forensics research, including the corpora from the Text REtrieval Conference, the American National Corpus Project, and the CALLFRIEND database of voice recordings, can be found on the Forensics Wiki at http://www.forensicswiki.org/wiki/Forensic_corpora.

Other research communities have established corpora for the purpose of enabling research; indeed, the creation of corpora has come to be regarded as a worthy scientific pursuit in its own right. For example, GenBank is a database of genetic sequences operated by the National Institutes of Health (National Center for Biotechnology Information, 2008).

Some schools have attempted to address the problem of exposing information security students to sensitive information by requiring that they sign written agreements. For example, George Washington University requires that students ''students entering Certificate, Masters or Doctoral programs in information assurance management'' to sign an agreement stating that they will comply with the school's Code of Conduct, a draconian document that threatens expulsion from the program for any infraction of the ethical or legal rules (Rayan and Rayan).

7. Conclusion

In this work, we argued that the development of representative standardized corpora for digital forensics research is essential for the long-term scientific health and legal standing of the field. We developed a baseline taxonomy of such corpora and outlined the legal and ethical hurdles that complicate their development. And we present a number of data sets that attempt to cover the spectrum of scenarios and have made them openly available to researchers. Special care has been taken to document the source of the data, as well as to avoid as many legal restrictions on its distribution as possible.

It is our hope that the community will support this effort and will adopt the provided sets for education, testing and research. Over the long run, it will be important to extend the scope of the corpora and to update it frequently to keep up with the pace of technology development. To that end, feedback from researchers will be essential in improving the collection methodology. We also hope that the sheer volume of data will challenge tool developers to come up with new techniques for processing huge amounts of data. In that case, the corpora can serve a target for performance evaluation studies.

The corpora we are presenting here are limited to corpora of files and disk images. There is also a real and pressing need for corpora of network packet captures and memory images. We hope that our work here will serve as an inspiration to others. We are happy to host the data from other experimenters on our web servers, so that there is ''one-stop shopping'' for forensic students and researchers.

Recently the National Research Council issued a scathing reporting on the status of forensic science, research, and practice in the United States. The NRC report devotes little space to the computer forensics, noting that much of today's

forensic practice originated in police departments, not forensic laboratories, and stating that only 25% of the forensic laboratories in the US have any computer forensics capability at all. Nevertheless, the report's concerns apply equally well to digital forensics: ''substantive information and testimony based on faulty forensic science analysis may have contributed to wrongful convictions of innocent people… Moreover, imprecise or exaggerated expert testimony has sometimes contributed to the admission of erroneous or misleading evidence (Committee, 2009)''.

If digital forensic science is truly a science, then the research community needs to adopt a culture of rigor and insistence on the reproducibility of results. Standardized forensic corpora will go a long way to making such desires a reality.

Acknowledgements

Funding for the initial Real Data Corpus was personally funded by Simson Garfinkel and Beth Rosenberg. Additional funding was provided by Basis Technology Corp. and NSF Award 0730389. Additional funding for the work described in this paper was provided in part by the National Institute of Standards and Technology.

REFERENCES

Calhoun William C, Coles Drue. Predicting the types of file fragments. In: Digital Investigation: The Proceedings of the eighth annual DFRWS conference, vol. 5; 2008.

Carrier Brian. Digital forensics tool testing images, http://dftt.sourceforge.net/; 2007.

Committee on Identifying the Needs of the Forensic Science Community. Strengthening forensic science in the united states: a path forward, February 2009.

Computer forensic tool testing program. Forensic string searching tool requirements specification, http://www.cftt.nist.gov/ss-req-sc-draft-v1_0.pdf; January 24 2008.

Cunningham Robert, Lippmann Richard P, Fried David J, Garfinkel Simson L, Graf Isaac, Kendall Kris R, Webster Seth E, Wyschogrod Dan, Zissman Marc A. Evaluating intrusion detection systems without attacking your friends: the 1998 DARPA intrusion detection evaluation. In: Third conference and workshop on intrusion detection and response; 1999.

MIT Lincoln Laboratory. DARPA intrusion detection data sets, http://www.ll.mit.edu/mission/communications/ist/corpora/ideval/data/; 2000.

Deolalikar Vinay, Latte Hernan. Provenance as data mining: combining file system metadata with content analysis. Usenix, http://www.usenix.org/events/tapp09/tech/full_papers/deolalikar/deolalikar.pdf; February 12 2009.

DMCI. Dublin core metadata initiative, http://www. dublincore.org; March 2009.

Farrell Paul. A framework for automated digital forensic reporting. Master's thesis, Naval Postgraduate School; 2009.

Federal rules of evidence, article X, rule 1003: admissibility of duplicates, http://www.law.cornell.edu/rules/fre/rules.htm; 2008.

Garfinkel Simson L. IRBs and security research: myths, facts and mission creep. In: Usability, psychology and security 2008 (co-located with the 5th USENIX symposium on Networked Systems Design and Implementation (NSDI '08)), http://www.simson.net/clips/academic/2008.UPS2008. pdf; April 2008.

Garfinkel Simson L. Automating disk forensic processing with sleuthkit, xml and python. In: Proceedings of the fourth international IEEE workshop on systematic approaches to digital forensic engineering. IEEE; 2009a.

Garfinkel Simson L. Providing cryptographic security and evidentiary chain-of-custody with the advanced forensic format, library, and tools. The International Journal of Digital Crime and Forensics 2009b;1.

Garfinkel Simson, Shelat Abhi. Remembrance of data passed. IEEE Security and Privacy January 2002.

Honeynet. Scan of the month, http://old.honeynet.org/scans/; 2009.

Klimt Bryan, Yang Yiming. Introducing the enron corpus. In: Conference on Email and Anti-Spam (CEAS). CEAS, http://www.ceas.cc/papers-2004/168.pdf; 2004.

Kornblum Jesse. Boomer-win2 k, http://dftt.sourceforge.net/test13/; 2007.

Lyle Jim. The cfreds project, http://www.cfreds.nist.gov/; 2008a.

Lyle Jim. Unicode string searching—Russian text, http://www.cfreds.nist.gov/utf-16-russ.html; 2008b.

Mawi working group traffic archive, http://tracer.csl.sony.co. jp/mawi/; 2009.

McDaniel Mason. Automatic file type detection algorithm. Master's thesis, James Madison University; 2001.

Moody Sarah J, Erbacher Robert F. SÁdi statistical analysis for data type identification. In: Third international workshop on systematic approaches to digital forensic engineering; 2008. pp. 41–54.

National Center for Biotechnology Information. Genbank overview, http://www.ncbi.nlm.nih.gov/Genbank/; April 2 2008.

Patzakis John. The best evidence rule. EnCase Legal Journal 2001: 31–8.

Ryan Julie JCH, Ryan Daniel J. Institutional and professional liability in information assurance education. working paper; http://www.danjryan.com/papers.htm; 2009.

The National Science Digital Library. Metadata guidelines, http://nsdl.org/collection/type.php; 2009.

Simson L. Garfinkel is an Associate Professor at the Naval Postgraduate School in Monterey, California, and an associate of the School of Engineering and Applied Sciences at Harvard University. His research interests include computer forensics, the emerging field of usability and security, personal information management, privacy, information policy and terrorism.

L.T. Paul Farrell was a graduate student at the Naval Postgraduate School when this work was done, and is now serving overseas with US forces.

Vassil Roussev is an Associate Professor at the University of New Orleans. His research interests include Distributed systems—computer supported cooperative work (CSCW), on-the-spot digital forensics, mobile devices. Software engineering—pattern-based techniques, component and service based models, agile development methods.

George Dinolt is a Professor of the Practice at the Naval Postgraduate School in Monterey, California. His research interests include formal methods, network security, and high-performance cryptography.