Digital Acquisition Learning Laboratory: A White Paper

Christopher A. Lee and Kam Woods
School of Information and Library Science
University of North Carolina at Chapel Hill
November 2011

This white paper is the result of a project funded by the Mellon Foundation called the Digital Acquisition Learning Laboratory (DALL). With the widespread adoption of digital information and communication technologies, collecting institutions – archives, libraries, museums – have unprecedented opportunities to document the lives of individuals. In order to seize these opportunities, information professionals must be prepared to extract digital materials from removable media in ways that reflect the rich metadata and ensure the integrity of the materials. Fortunately, there is an entire industry – digital forensics – that has established principles, technologies and methods for extracting data and associated metadata.

We have created, administered and implemented of a learning laboratory for the application of digital forensics to the acquisition of digital materials. This has included training, hardware and software to run practical exercises within courses at the School of Information and Library Science (SILS) at the University of North Carolina. DALL has provided the seed for a sustainable program of digital forensics education at SILS and several other professional development activities. We hope that the information and lessons conveyed in this white paper will help to inform similar efforts elsewhere.

Contents

1. Rationale

Collection institutions – libraries, archives and museums (LAMs) – are increasingly called upon to move born-digital materials that are stored on removable media into more sustainable preservation environments. This can involve media that are already in their holdings (e.g. disks stored in boxes along with paper materials), as well as materials that they are acquiring for the first time from individual donors or other producers.

The literature on digital archives tends to place a great emphasis on the "virtual" (i.e. intangible) nature of electronic resources. Computer systems have "an illusion of immateriality by detecting error and correcting it,"[1] but it is essential to recognize that digital objects are created and perpetuated through physical things (e.g. charged magnetic particles, pulses of light, holes in disks). This materiality brings challenges, because data must be read from specific artifacts, which can become damaged or obsolete. However, the materiality of digital objects also brings unprecedented opportunities for description, interpretation and use.[2] There is a substantial body of information within the underlying data structures of computer systems that can often be discovered or recovered. Recovery of data from physical media has been a topic of discussion in the professional library and archives literature for several years.[3,4]

There is also a large and expanding industry associated with digital forensics, which focuses on the discovery, recovery, and validation of information from computer systems that is often not immediately visible to common users. Several authors have recently investigated the use of forensic tools and techniques for acquiring digital collections in libraries and archives.[5,6,7] The Prometheus and PERPOS projects have developed software for data extraction, focusing on needs of specific collecting contexts.[8,9,10] A project funded by the Mellon Foundation called "Computer Forensics and Born-Digital Content in Cultural Heritage Collections" hosted a symposium and generated a report,[11] which provided a significant contribution to this discussion.

A very relevant project has been Personal Archives Accessible in Digital Media (PARADIGM), which investigated "issues involved in preserving digital private papers through gaining practical experience in accessioning and ingesting digital private papers into digital repositories, and processing these in line with archival and digital preservation requirements." PARADIGM's most visible product has been a Workbook on Digital Private Papers.[12,13] The Digital Lives project also investigated "personal digital collections and their relationship with research repositories."[14,15] Another recent project generated a white paper based on a series of site visits and meetings of those working with the born-digital components of three significant literary collections.[16] Born-Digital Collections: An Inter-Institutional Model for Stewardship (AIMS),[17] also funded by the Andrew W. Mellon Foundation, has explored and documented workflows that include digital forensics elements.

There have been several recent initiatives to create digital sandbox or laboratory environments to support the education of information professionals, though none have addressed the acquisition of data from removable media. The Institute for Museum and Library Services (IMLS) funded "Improving Student Learning of Advanced Digital Technologies in an Online Laboratory: A Research Approach" at the University of Arizona, Tucson, School of Information Resources and Library Science; it is exploring "project-based hands-on learning using on-line laboratory environments as a strategy to enhance the teaching of advanced library technologies." Another IMLS-funded project is "Curriculum, Cooperation, Convergence, Capacity - Four C's for the Development of Cultural Heritage Institutions" at the Simmons College Graduate School of Library and Information Science, which is incorporating museum informatics and data stewardship into the existing program of study in GSLIS, in order to develop a curriculum that will educate students to be cultural heritage professionals; this includes a prototype online space for organizing and disseminating resources to facilitate hands-on learning experiences. The National

Endowment for the Humanities (NEH) Preservation and Access Education and Training Program is also funding a two-year project at the School of Information at the University of Michigan to develop and implement a virtual laboratory featuring digital access and preservation tools to be integrated into five master's level courses in two SI specializations: Preservation of Information, and Archives and Records Management. The Humanities Advanced Technology and Information Institute (HATII) at the University of Glasgow has created a graduate program in Computer Forensics and E-Discovery, which addresses many of the tasks and skills that DALL will address, but catered specifically to students who plan to work in security and law enforcement jobs.

2. DALL Project Description

The Digital Acquisition Learning Laboratory (DALL) project was designed to build connections between the field of digital forensics and the education of information professionals. It has involved the creation, administration and implementation of a learning laboratory for the application of digital forensics to the acquisition of digital materials. This has included training, hardware and software that provide the capacity for running numerous practical exercises within courses at the School of Information and Library Science (SILS) at the University of North Carolina. This project has provided the seed for a sustainable program of digital forensics education at SILS. It has also involved significant engagement with both professionals in collecting institutions and individuals involved in digital forensics research and development.

DALL was a one-year project conducted from June 2010 to July 2011, with funding from the Andrew W. Mellon Foundation, and under the direction of Principal Investigator, Christopher (Cal) Lee. The grant funded specialized hardware, software, training and technical support for the development and implementation of a hands-on learning environment for the application of digital forensics methods to the acquisition of born-digital materials. The technical components of DALL included a dedicated workstation for developing, testing and benchmarking various processes (FRED Station); licensing of commercial software – Encase from Guidance Software and Forensic Toolkit (FTK) from AccessData -- for use in educational laboratory projects and exercises; USB write blockers; and SILS computer lab support. DALL also supported specialized training for Lee and SILS IT Director, Scott Adams. Lee and Adams used a one-year "All-Access Pass" to take several online classes from AccessData, and they took on-site FRED (Forensic Recovery of Evidence Device) training from Digital Intelligence in New Berlin, Wisconsin.

The curriculum development and implementation activities of DALL have focused on preparing students to:

- ensure reliable data extraction from digital media and computing devices,
- understand the composition of and interrelationships between storage hardware and filesystems,
- identify and manage metadata generated in the acquisition process, metadata stored by the filesystem, and metadata within particular file formats,
- identify and manage private and sensitive information located on digital media, and
- provide context-aware packaging, distribution, and access to processed digital materials.

The digital forensics industry is dominated by two main products: the Forensic Toolkit (FTK) from AccessData and EnCase from Guidance Software. Digital forensics professionals are routinely trained to use these two applications, and they are expected to be installed on the computers used as forensics workstations. While there are numerous open-source tools on the market that perform specific digital forensics functions, they require extensive expertise and effort to install and administer. FTK and EnCase are integrated suites of software, which allow a user with relatively little previous computer training to work through the entire evidence acquisition process. Both vendors provide an extensive suite of training opportunities and certification programs. These two applications are also being used by the British Library, Stanford and Oxford, so their use in the DALL project has served as a valuable point of reference in further conversations and collaborations among these institutions. Purchase of FTK and EnCase provided students with an opportunity to use these packaged software suites to learn about the overall principles and issues of digital forensics without getting overwhelmed by the underlying technical details of software installation and configuration. By using both FTK and EnCase, the students, PI and SILS computing staff have learned about the relative merits and limitations of the two packages in collecting institution scenarios. FTK and EnCase are both normally several thousand dollars per seat, but Lee was able to obtain licenses to the applications at reduced academic rates.

DALL has established a learning environment to be used in support of both course work (offered to undergraduate and graduate students) at SILS and continuing professional education offerings. Throughout the summer and fall of 2010, Lee and the SILS computing staff worked closely with each other to gain sufficient application-specific expertise and setup a sufficient technological infrastructure within the school to administer educational exercises that integrate digital forensics into LAM (library, archives and museum) acquisition tasks. In the Spring 2011 semester, Lee administered an integrated set of exercises within a course (Electronic Records Management), which he team-taught with Richard Marciano, who is a Professor at SILS and Director of the Sustainable Archives and Leveraging Technologies (SALT) Lab at UNC. Lee and Kam Woods have also developed and administered a new special-topics course at SILS called "Acquiring Information from Digital Storage Media," which introduced students to digital forensics concepts and methods, using both commercial and open-source software. In addition to building significant hardware, software and procedural capacity for future digital forensics teaching and research, the DALL project has provided SILS personnel with valuable practical lessons about the likely opportunities and challenges of LAM professionals using digital forensics tools.

Students have been taught how to plan for and implement a workflow for digital acquisition that includes leading-edge digital forensics tools and methods. In addition to the logistical issues,[18] there are also institutional and ethical issues,[19,20] which they explored and clarified in ways that would not be possible without direct experience with the technology.

One of the fundamental goals of DALL is for aspiring information professionals to understand the characteristics of and be able to apply operations to *disk images*. A disk image is a sector-by-sector copy of the data that was stored on a physical medium. It is a "snapshot" of the medium's content, including not only allocated files, file names, and other metadata, but also portions of deleted files and a variety of other trace information. Forensic acquisition of data from a medium typically involves the generation of a full disk image. Once a disk image has been generated, it is then stored as a single file or set of files. Forensic investigation and analysis is usually performed on disk images (i.e. copies of the medium

content) rather than on the original storage medium. Disk images are created using special software that accesses the physical through low-level input-output operations, rather than relying on the host computer's file system as an intermediary. The most widely available tool for creating disk images is the UNIX "dd" command. However, forensic investigators tend to use more specialized applications that generate record metadata (e.g. the name of the investigator, time that the image was created, and notes) and integrity information (e.g. checksums or hashes). The commercial applications that we have used in DALL for imaging disks are EnCase and FTK. We have also used FTK Imager, which is a free utility from AccessData, the producers of FTK.

Once one has created a disk image, there are a variety of tasks that he or she may wish to perform on the data. Common examples include searching for keywords, identifying files of certain types, analyzing timestamp information to reconstruct sequences of events, and recovering user account information. EnCase and FTK are suites of tools that were designed to administer, perform, manage, and document such activities on disk images. While these forensic applications were not designed for professionals working in collecting institutions such as libraries and archives, their functionality can be applied in such contexts.

3. The Setting: School of Information and Library Science at UNC-CH

The mission of the School of Information and Library Science (SILS) at the University of North Carolina, Chapel Hill (UNC-CH) is to advance the profession and practice of librarianship and information science, to prepare students for careers in the field of information and library science, and to make significant contributions to the study of information. Its professional degree programs are accredited by the American Library Association.

SILS offers two masters degrees (Master of Science in Library Science and Master of Science in Information Science), a Bachelor of Science in Information Science, and doctoral degrees. There are currently approximately 400 students at SILS: 300 MSLS or MSIS, 50 PhD, 50 undergraduate majors, and 30 undergraduate minors. There are 26 full-time teaching faculty members and about 25 adjunct and visiting faculty members associated with SILS each academic year. In April 2010, the SILS faculty approved a Graduate Certificate in Digital Curation.

SILS has active programs of digital library and digital curation education and research. From 2006 to 2009, the National Science Foundation funded two grants to SILS at UNC and Virginia Tech to develop a model digital library curriculum.[21] The project – with the SILS team being led by Jeffrey Pomerantz and Barbara Wildemuth – generated a framework, templates and modules that have been tested and implemented through course work in the MSIS/MSLS program at SILS and the computer science program at Virginia Tech. Since 2006, SILS faculty Helen Tibbo and Christopher (Cal) Lee have also administered five major digital curation education and professional engagement projects funded by the Institute for Museum and Library Services (IMLS): (1) Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum (DigCCurr),[22] (2) DigCCurr II,[23] (3) Educating Stewards of

Public Information for the 21st Century (ESOPI-21),[24] (4) Closing the Digital Curation Gap (CDCG),[25] and (5) Educating Stewards of the Public Information Infrastructure (ESOPI^2).

Prior to the DALL project, Lee had developed and run professional workshops on "Applying Digital Forensics Techniques to Materials Acquired on Physical Media" and had incorporated targeted digital forensics exercises in the DigCCurr II Professional Institute and the SILS course called Understanding Information Technology for Managing Digital Collections (INLS 465). However, the ideas had been conveyed through conceptual explanations, numerous walk-throughs based on screenshots, and a few limited hands-on tasks (viewing Hex dumps and generating/comparing checksums) that could be carried out using simple, free tools.

4. Course Development and Implementation

The DALL project included development of exercises and assignments for INLS 525 (Electronic Records Management) and development of a new one-credit special topics course called INLS 490-141: Acquiring Information from Digital Storage Media. Lee and another SILS faculty member, Richard Marciano, co-taught INLS 525; and Lee and Kam Woods co-taught INLS 490. Both courses were in the Spring 2011 term. The courses made use of commercial tools funded through DALL -- EnCase and FTK – as well as a variety of free and open-source tools.

Electronic Records Management is a course that Lee has been teaching at SILS since Spring 2006. It has evolved over the years, to include a variety of assignments and exercises. The DALL project allowed Lee to add elements to the course that provided direct, hands-on experience with the latest commercial digital forensics tools.

Acquisition of Information from Digital Media (INLS 490-141) was a course designed specifically to address the objectives of the DALL project. The course description indicates:

> *Students will learn about hardware, software and methods used to extract digital data that have been stored on removable media (e.g. hard drives, floppy disks, USB memory sticks). This course addresses common storage devices and interfaces; write-blocking equipment and its role in acquisition of data; levels of representation; basic filesystem structures; role and importance of hash values and hex views of bitstreams; and software used to conduct data acquisition. Students will have the opportunity to use a range of state-of-the-art digital forensics hardware and (commercial and open-source) software and explore ways that they can be applied by information professionals in a variety of collecting contexts.*

This course was designed specifically to provide students with both a conceptual background in the analysis of digital media and filesystems, and the practical skills required to perform low-risk, high-quality media acquisition, analysis, and reporting procedures.

5. Laboratory Setup

Both courses were conducted in a multi-use computing laboratory at SILS. This laboratory consists of approximately 30 workstations with Intel Core 2 processors, 8GB RAM, 500GB hard disks, running the Windows 7 operating system. The computers were configured with Deep Freeze, a kernel-level driver which maintains system integrity and returns the operating system to a "default" state (the state in which it is frozen) after reboot. This provides users with the advantages of Administrator-level access to the computer (within any given session) without the risk of damage to the system.

Figure 1: SILS Multi-Use Computer Laboratory

A vital aspect of acquiring data from removable media is ensuring that software on the target computer does not write data to the acquired device, which then alters the bitstream on that device. The standard mechanisms for copying files from a device cannot ensure such bitstream integrity, because they tend to write data in both directions. The widely-accepted solution to this problem is called a write blocker, which turns the connection between two devices into a "one-way street." As part of the DALL project, SILS purchased twelve Tableau USB forensic write blockers to be used in instruction. Buying these external devices prevented the need to permanently install write blockers on the lab machines and also allows students to use their own laptop computers for some tasks. We have used the write blockers in class sessions to demonstrate secure acquisition of bitstreams from USB devices such as external hard drives and USB flash media. Each write blocker came with a card that conveys the steps to be followed when using the device, making it relatively easy to walk students through the process of reading data from a drive in read-only mode. Students were assigned to groups that each shared a write blocker for the in-class exercises.

Figure 2: Tableau Forensic Write-Blocker

We also made extensive use of 15 USB thumb drives, each of which has a capacity of 16 GB. By replicating the exact same bitstream onto each thumb drive, we could then ask the students to create disk images of each, knowing that they would get the same results. We could also disseminate data and software to all of the students relatively quickly through use of the thumb drives. Finally, we could set up the thumb drives as bootable Linux-based environments, so students could run a variety of open-source tools directly from the drives, without having to install Linux or the other tools on the computer lab computers.

Prior to the start of the semester, the computers in the multi-use laboratory were "unfrozen" and updated with permanent installations of Forensic ToolKit (FTK) version 2.3 and EnCase version 6. We later updated the applications to versions 3 and 7, respectively.

In preparation of course materials, we have made use of a dedicated Forensic Recovery of Evidence Device (FRED) station, developed by Digital Intelligence. The FRED includes a variety of useful features, including an integrated set of write blockers, a ventilated imaging shelf, two high-capacity hard drives, and several built-in enclosures that can be used for "hot swapping" drives. Our preparation of disk images has also been supported by Runtime Software Disk Explorer (NTFS and Linux analytic software for Windows).

Individuals are increasingly making use of mobile devices to store and transport files, and the latest Deluxe version of EnCase includes a smart phone module. We have begun exploring the use of such software in SILS instruction, in order to demonstrate the similarities and differences between acquisition of data from disk and drives and acquisition of data from phones and SIM cards. This has included preliminary testing and investigation of the following: Dekart SIM Reader 2 (USB SIM reader + acquisition software) and Dekart SIM Explorer (SIM analysis).

6. Lessons Learned and Advice for Educators

During the course of this work, we developed a number of insights related both to the use of digital forensics technologies in classroom environments, and to more general questions of handling complex technical setups in shared-use environments.

We faced a variety of challenges associated with the two main commercial forensics applications – EnCase and FTK – not being specifically designed for administration in a classroom setting. Their default configurations and documentation are based on the assumption that one is working in a forensic lab environment in which either one or a small set of dedicated computers are being used for forensic investigations. Their typical workflow also requires large blocks of time for processes to run. We established a variety of conventions for working around these issues.

6.1. Data persistence and network shares

As in other courses, students frequently need to save data related to laboratory activities; this may be an ongoing project to be used in a later class, work that they intend to complete on their own time, or simply examples of techniques they have learned in class for future references.

This can be problematic in shared-use environments where the computing environments themselves are not persistent *and* where the students are working with large quantities of data. We discovered that students working with datasets more than a few hundred megabytes could not reliably save these datasets to their shared network storage (as provided by UNC) in a reasonable amount of time due to network bandwidth constraints or, on occasion, a lack of space.

As indicated above, SILS uses a program called Deep Freeze (Faronics Corporation), which freshly restores each lab computer to a set configuration every time the computer is restarted. This means that all of the computers are effectively reimaged every night. In order to install any new software, the computers must be "unfrozen" and reconfigured to include that new software each time the computer restarts. Both EnCase and FTK store information within "cases," which are written to a database that is stored on the local computer's hard drive. Because of the Deep Freeze setup, students could not build cases and then save them on the computer to be reused later. We worked around this in some cases by providing them with pre-built case files on USB thumb drives that they could load directly. In other cases, students had to build their own cases, which required additional export and import actions, all of which could cause the students problems if they did not take the proper steps.

A related issue was the installation of other free software for use in specific classroom exercises. In order to use a basic hex editor or disk image mounting application, for example, the students first needed to download and install the software on their lab computers. This would take time away from the classroom lessons and would occasionally fail due to some categories of students (undergraduates and UNC employees who were not enrolled in the SILS masters program) not having the appropriate administrative privileges on the computers. We could add the applications to the full set of default software on each computer, but that was often not warranted if the software was simply being used for one or a handful of classroom exercises.

As described above, they involved importing and exporting of cases and in-class distribution of prepared data on USB thumb drives. We are also investigating the possibility of moving the database components of the software onto a dedicated server, which will not be subject to the Deep Freeze data wiping issues.

6.2. Export and import of case data

EnCase and FTK have database back-ends that store case information. It can take several hours to load data into a new case, so administration of exercises cannot depend on creating new cases each time. Instead the instructor or students must export and import of cases between sessions. The software can be unforgiving to small differences between the environment in which the case file was created and the environment in which it is being used. We had to walk students through manual changes to directory paths for the case files, which had been hard coded into cases by FTK.

6.3. Shared authentication resources (dongles and network keys)

As commercial digital forensics software vendors, AccessData and Guidance Software both protect their products with hardware dongles loaded with unique product keys. The procedures these companies follow for key use, distribution, and sales for educational purposes can vary considerably. For our laboratory, AccessData provided unique physical dongles for every workstation; these were secured inside each machine and remained in the laboratory for the duration of the project so that students could use the software outside of class time.

Guidance Software provided a single dongle for our educational license of EnCase 6, preloaded with a multi-client key capable of supporting up to 30 clients over the network. This required setup of a dedicated machine outside of the laboratory (in our IT facility) that would remain powered on and serve out licenses via a keyserver. When we began this process, it became clear that EnCase (which has relatively few educational customers) did not provide appropriate documentation, either with the product or online) to accomplish this. We spent a significant amount of time determining the steps required to reconfigure the network firewall, the ports over which the server would communicate with the clients, and the steps required to set up each individual client.

6.4. Time required to index disk images for analysis

EnCase and FTK provide a variety of sophisticated processing options for digital materials. However, even using the default processing settings, analysis can be an extremely slow process, as the software will generally produce a searchable index of not only of the files within the filesystem, but also all of the strings found in any of several hundred formats, both binary (such as legacy Microsoft Office documents, or the EXIF data within JPGs) and text-based (XML and HTML files, text documents). Although *searching* the index is extremely fast, building it is not.

As an example, FTK 2.3 required approximately 10 hours to process one of the 4GB (10GB uncompressed) Advanced File Format disk images provided to students for the semester project in INLS 525 (Electronic Records Management). This presented a practical issue, as students could not remain in the lab during this time – and could not guarantee that another use would not simply reboot the system. We provided an alternative in the form of a dedicated system outside the lab that students could schedule for use.

6.5. Practical limits for data use within the classroom

We found that it was impractical to work with datasets larger than several hundred megabytes, as the time required simply to process the data prior to exploring the analytics features of the commercial software packages was otherwise prohibitive (e.g. more than 25% of the time allotted to the class session). Note that in the intervening period, both AccessData and Guidance Software have made some improvements to their software to reduce processing times.

6.6. Alternatives to full versions of commercial forensics programs

For shorter exercises and introductory concepts (for example, hexadecimal filesystem views and cryptographic file hashes) we found that it was often simpler and faster to use feature-limited free versions of particular software packages (or freeware utilities designed for a small number of tasks). In particular, we found that we could use AccessData's FTK Imager (a free utility to do simple disk imaging and filesystem verification) to demonstrate a wide variety of concepts without requiring the students to step through the somewhat complex case file creation process required by FTK 2.3 prior to analysis.

6.7. Value of realistic and open access disk images

Realistic data corpora (data items that have been synthesized in controlled environments and incorporate most or all of the qualities of data drawn from real-world sources) are important sources from which to develop class support materials. We have made use of two primary set of disk images for courses at SILS: (1) a set of images from CDs published by the US government, which are real data but not subject to data sensitivity issues, and (2) a set of hard drive images from the M57 data set, generated by a team at the Naval Postgraduate School (under the leadership of Simson Garfinkel)[26] to be realistic but contain no sensitive data and relatively few intellectual property constraints.

6.8. The "progress bar" problem

Users of contemporary operating systems and applications have a tendency to accept that when a progress bar appears on the screen, the computer is doing something that the user cannot control (except, perhaps, to cancel), and therefore do not need to monitor or investigate. Relatively few of the activities that students undertake in classes or laboratories (with a small number of exceptions) are so computationally intensive that this becomes a serious issue. However, digital forensics software may take hours, days, or even weeks to process large data objects, depending on what one is asking the software to do.

We have found it important to explain the connection between a lengthy process and what is happening to make the process take so long. This becomes a learning experience about how and why certain actions are computationally intensive and how to plan workflows around delays (e.g. parallel activities).

6.9. Managing class, homework, and project expectations

Digital forensics professionals typically spend several years acquiring the skills and expertise required to do their work. Training in digital forensics provides an intimate familiarity with the time required to process digital materials and the feasibility of specific approaches to information extraction and analysis.

Students encountering digital forensics tools for the first time may simultaneously express an excitement about the powerful analytic techniques that are available and skepticism or unhappiness about the amount of time required to extract usable data and perform various analyses. In some cases, students may never

have been tasked with running data processing tasks or experiments that require many hours or days to complete. Preparing students by explaining *why* these procedures take time can help to improve the outcome of a course.

Planning the incorporation of digital forensics techniques must balance two fundamental issues: (1) providing students with sufficient background to allow independent exploration of the tools and data materials (particularly for semester projects) and (2) providing sufficient time for students to complete complex, open-ended assignments.

Alternatively, instructors may choose to provide packaged, discrete exercises for which the outcome is known. This is similar to a cooking show, in which the audience is asked to pretend that a long process has just taken place (e.g. 45 minutes of baking in 10 seconds). Such an arrangement can make efficient use of class time, but it does not expose students to the full chain of activities involved in a data acquisition and processing workflow.

If instructors are to require students to work through processes that do take several hours to complete, then it is important to convey such expectations early in the term. Running and monitoring processes in a dedicated computer lab is very different from what students often encounter in their coursework for other library and information science courses: reading, writing and group project interactions, which can occur at a variety of locations.

6.10. Teaching skills vs. teaching interfaces

EnCase and FTK are complex, powerful pieces of software with interfaces that can be daunting to novice users. Both AccessData and Guidance Software provide training - at a rate of several thousand dollars per course - to digital forensics professionals, both to familiarize them with changes in these interfaces across product revisions and to teach specific skills related to analyzing particular types of devices, operating systems, and file formats. Training is a major source of revenue for these companies, and the audience (predominantly law enforcement and corporate security workers) benefits from the generally high quality and accessibility of the training materials, and from certificates of completion that can be presented to employers.

Students training to become digital forensics or information security experts in professional or vocational programs will often receive similar instruction. However, incorporating this level of detail into Information and Library Science programs may be impractical - both due to time constraints and the fact that ILS students may only encounter one such class in the entirety of their professional education.

Rather than focusing on “teaching the application,” we believe that an effective introduction to digital forensics techniques in this context is one which provides students with a grounding in fundamental concepts and terminology and then encourages them to explore applications in the course of self-directed projects. Preparing students for this undertaking early in the course is critical.

7. Next Steps

There are numerous elements of the DALL project that were designed to build capacity and ensure the sustainability of digital acquisition education activities. These include: (1) training of faculty and IT

support staff, (2) incorporating laboratory activities into coursework that SILS will offer on a regular basis in the future, (3) selecting hardware and software from vendors who have shown a consistently strong standing in the market, (4) taking advantage of beneficial multi-year licensing arrangements, and (5) building collaborations with partners on campus and beyond. We believe we have successfully advanced all five of the above elements.

The DALL work has also had an impact on continuing professional education activities. Lee and Helen Tibbo offered a one-day workshop called "An Introduction to Digital Curation for Public Records Professionals," in which Lee administered an exercise that has drawn from DALL activities. Lee is also developing a day-long workshop on "Digital Forensics for Archivists" to be offered by the Society of American Archivists as part of their new Digital Archives Specialization (DAS) curriculum, and he plans to offer a one-day short course at the IS&T Archiving conference in Copenhagen on June 12, 2012; both offerings will draw from contributions of the DALL project. Lee will also be incorporating DALL lessons and materials into the 2012 DigCCurr Professional Institute, to be administered in Chapel Hill in May 2012.

Another important avenue to pursue is the use of open-source digital forensics software in LIS courses. We have found the use of commercial software to be a fruitful and informative experience, but are also aware that not all institutions will be able to afford licenses to EnCase and FTK. We have attempted to create course materials that are relatively agnostic to the software used. See the Appendices for examples of such materials. One of the challenges of using open-source forensics software is that much of it is unapproachable to professionals who do not have extensive experience with a Unix-style command line and relatively advanced system configuration. SILS at UNC and the Maryland Institute of Technology in the Humanities (MITH) have recently begun a project funded by the Andrew W. Mellon Foundation, called BitCurator,[27] which aims to package and disseminate open-source digital forensics tools for use by professionals in collecting institutions. We hope that the BitCurator project will be help to further the resources available to instructors who are teaching digital forensics concepts.

Appendices

Selected Course Materials from INLS 525 – Electronic Records Management, Spring 2011

- **Lab Exercises:**
  - Examining and Considering Files on Your Computer [Week 3]
  - File Integrity through Hashing [Week 5]
  - Two Views of the Same ISO File [Week 7]
- **Data Recovery and Ingest Assignment:**
  - Group Assignment – Data Recovery and Ingest
  - Individual Paper on Data Recovery and Ingest Assignment

Unidentified Files Exercise (3 Versions):

- Curation of Unidentified Files - DigCCurr Professional Institute Version
- Unidentified Files Exercise - In-Class Version
- Curation of Unidentified Files – State Archives Version

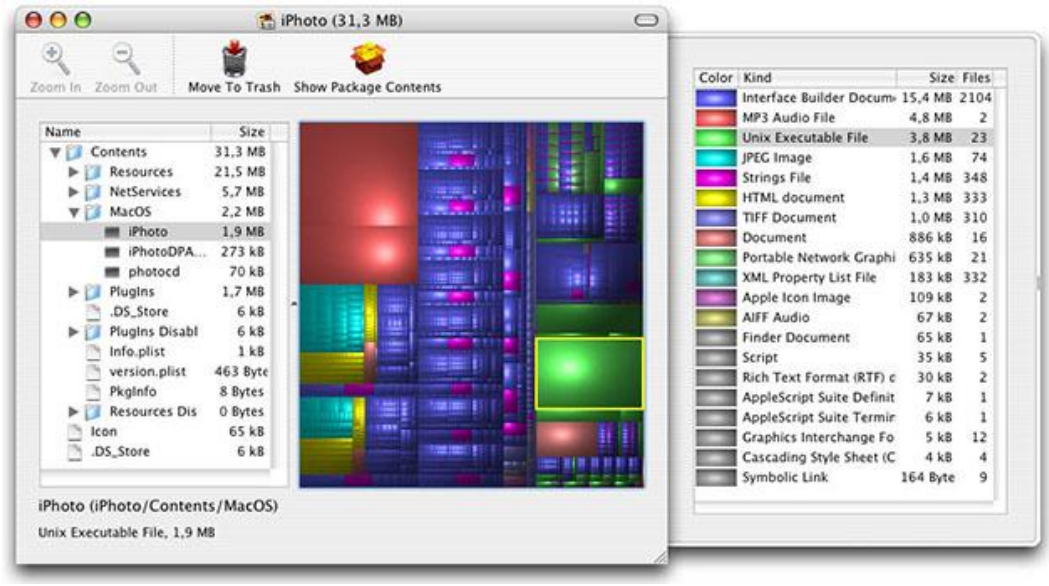
Special Topics Course: Acquiring Information from Digital Storage Media (INLS 490-141), Spring 2011 – Abridged Syllabus

Examining and Considering Files on Your Computer

Download and run one of the following (best if you can do this on your own computer):

- WinDirStat (Windows) - http://windirstat.info/
- Disk Inventory X (Mac) - http://www.derlien.com/

Note: these are not forensic tools, and you don't need to reveal anything about the contents of files.



Document what you discover about the contents of your drive(s) – e.g. screenshots, notes, or whatever mode you prefer. Bring you observations to class. Ideally, also bring the computer on which you ran this analysis.

What were the main things you discovered from this exercise?

- Things that reinforced your expectations
- Things that surprised you

Recordkeeping Implications?

RIM should "ensure:

- that adequate records are created to document business functions and meet administrative, legal, and other operational needs;
- that **recordkeeping** requirements are analyzed and included when information systems are first developed;
- that professionally sanctioned techniques are applied throughout the records life cycle;
- that records are retained and disposed of based on analysis of their functions and value; and
- that records of continuing value are preserved and accessible"[1]

Pretend someone whose records are important to you has just allowed you to use WinDirStat / Disk Inventory XP to analyze the contents of a computer that he/she uses.

Which of the objectives to the left could be supported by the kind of information you've just discussed? How?

What further information would you want, which is missing from the WinDirStat / Disk Inventory XP views?

[1] Stephens, David O. "Introduction: Status and Trends." *Records Management: Making the Transition from Paper to Electronic*. Lenexa, KS: ARMA, 2007. p.1 (emphasis in original)

File Integrity through Hashing

Step 0 – Preparation (Using Lab Computer)

- Download the following from [DESIGNATED COURSE SPACE] for Class Exercises to your desktop:
    - md5summer
    - HVIEW2000
    - 525-files.zip
- Extract the contents of 525-files.zip to your desktop

Step 1 - Generate hashes

- Open md5summer
- Select your desktop as the root folder
- Select "Create Sums"
- In the "Create list of files to sum" dialog, navigate to the 525-files folder on your desktop, then select all of the contents of that folder and select "Add"
- Click "Ok"
- Confirm that the md5 hashes are generated
- "Save as" to your desktop as 525files.md5
- Leave the md5summer application running

Step 2 – Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "525-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Step 3 – Change file names

- Change the file names of one or more of the files – make a note of which ones you change
- NOTE: Be sure **NOT** to open the files in an application. You should only change the names.

Step 4 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "525-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Step 6 – Opening files with hex editor

- Pick one or more items from the "files" folder and open them in HVIEW (you can drag the file onto the HVIEW icon on your desktop)
- Look at the hex view of each file and see what it reveals about the bitstream content of the file

Step 7 – Change content with hex editor

- For one or more of the files, change a bit within the file in HVIEW and then save the changed file
- Note which files you've changed and the SPECIFIC PLACE in the files where you changed bits

Step 8 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "525-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Step 9 – Change content back to earlier state in hex editor

- Open one or more of the files that you changed in step 7, and change the bits back to their previous state, and then save the changed file(s)

Step 8 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "525-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Two Views of the Same ISO File

Required for this exercise:

- ISO file - http://www.ils.unc.edu/callee/inls525-files/34.iso
- MagicDisc - http://www.magiciso.com/tutorials/miso-magicdisc-history.htm
- FTK Imager – installed on lab computer

Step 1 – Download 1 and 2 above and install MagicDisc

Step 2 – Mount the ISO file using MagicDisc

Step 3 – Find the drive using Windows Explorer and investigate its contents

Step 4 – Open FTK Imager and add 34.iso as an evidence item

Step 5 – Explore what you see in the drive using FTK Imager

Data Recovery and Ingest – Group Assignment

With a group of other students, you will receive a set of files and will then carry out a set of recovery and ingest tasks with those files. See instructions below for group assignments and basic instructions for accessing data. You will need to download these files in order to work with them.

Scenario:

You work for the Electronic Resource Acquisition Consortium (ERAC), which is a collaboration of libraries and archives across the U.S. that pay annual membership fees. As part of their membership, an institution can send digital media to the ERAC processing center, where the staff recover data from the media and provide both contents and associated metadata back to the institution. This ERAC service is new, and you have recently been hired to work on it. The ERAC has already defined a general high-level workflow, but many of the individual steps, methods and conventions are still in development. This assignment assumes that you are responsible for the first cases of this new process, so your group will need to determine how best to implement each of the tasks.

The workflow of the ERAC involves three major steps:

1. Physical Treatment and Acquisition Unit (PTAU) – This involves all measures required to generate a bit-by-bit copy (i.e. disk image) from each storage medium that an institution sends to the processing center.
2. Initial Assessment and Triage Unit (IATU) – This involves a general determination of the medium's content; assessment of potential use value and use cases of the content; documentation of system dependencies and preservation considerations; generation of metadata (including file tree of drive contents, hashes of all files, and any other basic metadata deemed important for further treatment of the content) and capture of the metadata outside of the disk images themselves; recommendations for future treatment of the disk image's contents, including arrangement, description, selection/appraisal and potential preservation actions. You work for the IATU, and the primary audience of your work is the staff of the FPDU (see below).
3. Final Processing and Distribution Unit (FPDU) – This involves any processing, transformation, annotation and packaging of the content that is deemed necessary. The FPDU then sends the resulting data back to the submitting institution. The presumption is that the materials will be sent to the submitting institution in a form that they can then manage within their own repository environments.

Each group will be instructed to download a set of disk images. They fall into two sets:

- Set 1 – These are CDs and floppy disks sent from a record center that houses records and publications of federal government agencies. The record center is now shutting down and these disks were included in a box of materials that were deemed to be of potential long-term retention value. Several state archives and state libraries that are members of ERAC have jointly asked the processing center to process the materials. They will then inspect the products to determine whether they should be acquired by one or more of the ERAC member institutions for inclusion in their own collections. The institutions' decisions will be significantly informed by any documentation you provide as a result of your work, but they will make the ultimate appraisal decisions based on their own collection missions and available resources.
- Set 2 – This is a single disk image obtained from an archive that is devoted to documenting research and innovation. This disk was included as part of the "personal papers" that an individual, Charlie Brown, donated to the archive along with various other materials. The donor worked for a patent research firm (see below).

Group Products:

Your group should produce and submit a report that documents the tasks that you have performed on the materials in both Set 1 and Set 2. Within the text of your report, you can reference any further associated data that you would like to provide. This should include, but is not limited to: file tree of drive contents, hashes of all files, and any other metadata that you would like to include.

As indicated above, the tasks for to perform and then describe in your report are:

- General determination of the medium's content
- Assessment of potential use value and use cases of the content
- Documentation of system dependencies and preservation considerations
- Generation of metadata (including file tree of drive contents, hashes of all files, and any other basic metadata deemed important for further treatment of the content) and capture of the metadata outside of the disk images themselves
- Recommendations for future treatment of the disk image's contents, including arrangement, description, selection/appraisal and potential preservation actions.

You are pioneering this process for the ERAC, which does not already have an established workflow in place for it. This has three important implications for what to include in your report:

- Detailed description of how you perform tasks will help the staff to perform these activities in the future. This includes what software you used, how you configured it, and limitations/challenges associated with options you chose.
- You should generate documentation that reflects both what you have been able to determine and important things you have **not** been able to determine (e.g. creator unknown). If you hit an impasse in parts of this work, it is important to explain what you tried, what challenges/questions you faced, and how you tried to address them, rather than simply skipping the task.
- Because the ERAC has not yet determine what (if any) specific preservation actions it will promise to make on collections that it processes, so you should feel free to elaborate a variety of options for what might be done with the materials. You do **not** need to develop one definitive preservation plan for the materials.

Profile of M57

M57.biz was a company that researched patent information for clients. The company had four employees.

Name	Position	Email Address	Email Password
Charlie Brown	Patent Researcher	charlie@m57.biz	brown01
Terry Johnson	IT Administrator	terry@m57.biz	Not provided
Pat McGoo	President	pat@m57.biz	Not provided
Jo Smith	Patent Researcher	jo@m57.biz	Not provided

Employees worked on site, and conducted most business exchanges over email. All of the employees worked in Windows environments, but they used a variety of applications to conduct their work. For example, some used Outlook and others used Thunderbird for email.

Group Assignments and Data Access

WARNING: The files are large, so download them from a computer with a fast connection. You may want to store them on a drive that you can readily access, rather than downloading them each time you use them. To mount the ISO disk images in a Windows environment (including the lab computers), you can download MagicDisc: http://www.magiciso.com/tutorials/miso-magicdisc-history.htm. Macs have a native mounting tool.

Group	Members	Files Assigned
1	Names of students in each group	http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/5.iso http://www.ils.unc.edu/callee/inls525-files/25.iso http://www.ils.unc.edu/callee/inls525-files/26.iso
2		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/18.iso http://www.ils.unc.edu/callee/inls525-files/20.iso http://www.ils.unc.edu/callee/inls525-files/29.iso
3		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/14.iso http://www.ils.unc.edu/callee/inls525-files/21.iso http://www.ils.unc.edu/callee/inls525-files/22.img
4		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/11.iso http://www.ils.unc.edu/callee/inls525-files/15.iso http://www.ils.unc.edu/callee/inls525-files/31.iso
5		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/13.iso http://www.ils.unc.edu/callee/inls525-files/24.iso http://www.ils.unc.edu/callee/inls525-files/36.iso
6		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/6.iso http://www.ils.unc.edu/callee/inls525-files/9.iso http://www.ils.unc.edu/callee/inls525-files/17.iso
7		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/3.iso http://www.ils.unc.edu/callee/inls525-files/7.iso http://www.ils.unc.edu/callee/inls525-files/27.iso
8		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/1.iso http://www.ils.unc.edu/callee/inls525-files/28.iso http://www.ils.unc.edu/callee/inls525-files/30.iso
9		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/2.iso http://www.ils.unc.edu/callee/inls525-files/4.iso http://www.ils.unc.edu/callee/inls525-files/33.img
10		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/8.iso http://www.ils.unc.edu/callee/inls525-files/12.iso http://www.ils.unc.edu/callee/inls525-files/35.img
11		http://www.ils.unc.edu/callee/inls525-files/hard-drive.aff http://www.ils.unc.edu/callee/inls525-files/16.iso http://www.ils.unc.edu/callee/inls525-files/34.iso

Individual Paper on Data Recovery and Ingest Assignment

Based on your experience on the group project, each student will write a briefing paper (4-6 pages, single spaced). The audience for the document is your supervisor. Topics to address in your paper:

- The current state of the files you investigated, i.e. the output of your efforts.
- Major accomplishments from the process (e.g. skills or capabilities that you've developed that will be helpful for future tasks, value that you added to the files after receiving them in their original state).
- Major challenges you confronted.
- Recommendations for what ERAC should do to prepare for and implement similar tasks in the future. This can include, but need not be limited to:
    - Further resources (technical or human) that would be helpful
    - Potential strategies for obtaining those resources
    - Procedures, conventions or policies that should be developed
    - Technical architecture or configuration issues that will be important to keep in mind.

Things you can assume about your audience (supervisor):

- She gave you the responsibility of carrying out the data recovery and ingest project, so she knows the details of the scenario you were given.
- She is well-versed in the principles of archives and records management, and she has a solid high-level understanding of the requirements for a good recordkeeping system (e.g. respect for provenance, importance of contextual information, assurance of file integrity).
- However, she has never worked directly with any of the software you used for this assignment.
- This was the first accessioning case for the ERAC service, so she has also not yet had a chance to think through many of the logistical details required or the potential workflow for this within your organization.
- In other words, she does not need to be told what the requirements were in your earlier group assignment, because she was the person who assigned you to do them. But you should not assume that she is familiar with any of the further details in the assignment that falls underneath those tasks.

Notes on style of the paper:

- Try to be succinct in your presentation. Use bullet points, tables or other visual representations that make your main points quick and easy to grasp.
- Begin the paper with an executive summary, which should be about half a page long and summarize the main points from the document.
- One important deviation from the "briefing paper" genre is that you should include citations to readings or other resources whenever you feel they would help to support your points. Use either footnotes or endnotes, so that the citations don't break up the flow of your text.

Curation of Unidentified Files - DigCCurr Professional Institute[2]

You work as a librarian for Disco Tech, a university in the Midwestern US. You are working with, Dr. Matilda Bigschat, a prominent senior member of the electrical engineering faculty, who would like to submit her materials to Disco Tech's Research Extending Institutional Repository (RTIR).

When you met with Dr. Bigschat, she told you about her research, which included numerous studies on the efforts of government agencies to respond to technological challenges. She then provided you with a box full of physical media (CDs, floppies, and an external hard drive). In some cases, she could tell you specifically what was on them and why it was important to retain. In other cases, she could only recall that the contents of the media were important to her research, based on where she had filed them in her office.

You examined one of the CDs. There are no labels on the CD itself, but the case containing it has a sticker that reads "Files." You loaded the CD into a drive and used FTK Imager to create a disk image of the CD. You then used FTK Imager to export the files, which you can now find in files.zip in Blackboard. For purposes of tasks 1-4, we'll work from files.zip.

Task 0 – Preparation

- Download the following:
  - Data files:
    - files.zip [From DCE, save to your desktop]
    - files.iso [From DCE, save to your desktop]
  - Tool to generate file hashes:
    - md5summer (Windows) [From DCE, save to your desktop]
  - Hex viewer and editor:
    - HVIEW2000 (Windows) [From DCE, save to your desktop]
    - Hex Fiend (Macintosh) [From http://ridiculousfish.com/hexfiend/]
- Create a folder on your desktop called bigschat-files, then extract the contents of files.zip to that folder
- If you'd like to complete the tasks in the ***optional Task Set 3***:
  - Tool for mounting an ISO file so that Windows treats it like just another drive on your computer:
    - MagicDisc (Windows) [From http://www.magiciso.com/tutorials/miso-magicdisc-history.htm]
  - Tool for generating and doing basic inspection of disk images:
    - FTK Imager [From DCE, save to your desktop]

Task Set 1 – Seeing what you have – looking in bigschat-files

Task 1: Inspection at the file level

[2] The DigCCurr Professional Institute was funded by a grant from the Institute for Museum and Library Services (IMLS) - Grant Award #RE-05-08-0060-08.

- Based simply on the properties and names of the files and directories (don't open any files yet), try to make some inferences about what these files are and how they might be related.

Task 2: Looking at names of specific files

- Find a file called Circular.596
- Do you see any other files that have similar names? If so, make note of which ones they are.

Task 3: Inspection at the bitstream level

- Generate a hex view of the file called Circular.596 – using HView or Hex Fiend
- What can you infer about this file from looking at the hex representation?
- If you identified other files in task 2, also view those in hex view. Do you notice anything similar?

Task 4: Investigating the .ISO (disk image) file

- Open FTK Imager
- Go to File | Add Evidence Item…
- Select "Image File"
- Browse to files.iso and then select "Finish"
- Navigate the file tree and discuss what you observe

Task Set 2 – Investigation based on hash values

NOTE: These tasks are based on using md5summer in Windows. If not using a Windows computer, following along with someone else for this set. You can also generate individual file hashes from the command line in Mac/Unix with the md5 command or use the online utility: at http://www.webutils.pl/MD5_Calculator.

Task 5 - Generate hashes

- Open md5summer
- Select bigschat-files as the root folder
- Select "Create Sums"
- In the "Create list of files to sum" dialog, navigate to the bigschat-files folder on your desktop, then select all of the contents of that folder and select "Add"
- Click "Ok"
- Confirm that the md5 hashes are generated
- "Save as" to your desktop as bigshat-files.md5
- Leave the md5summer application running

Task 6: Finding duplicate files

- Look at the MD5 value of Circular.596
- If you identified other files in task 2, are the values the same or different?
- If you notice something interesting about the MD5 values, what can you infer about what happened to the files?

Task 7 – Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "bigschat-files" directory as your root folder
- At the "Open md5sum file" dialog, select bigschat-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Task 8 – Change file names

- Change the file names of one or more of the files – make a note of which ones you change
- NOTE: Be sure NOT to open the files in an application. You should only change the names.

Task 9 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "bigschat-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors). If you're getting red dots, why might that be?

Task 10 – Opening files with hex editor

- Pick one or more items from the "files" folder and open them in HVIEW or Hex Fiend (you can drag the file onto the HVIEW icon on your desktop)
- Look at the hex view of each file and see what it reveals about the bitstream content of the file

Task 11 – Change content with hex editor

- For one or more of the files, change a bit within the file in HVIEW and then save the changed file
- Note which files you've changed and the SPECIFIC PLACE in the files where you changed bits

Task 12 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "525-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Task 13 – Change content back to earlier state in hex editor

- Open one or more of the files that you changed in step 7, and change the bits back to their previous state, and then save the changed file(s)

Task 14 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "525-files" directory as your root folder
- At the "Open md5sum file" dialog, select 525-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Task Set 3 – Two Views of the Same .iso File [Optional – Time Allowing]

Task 1 – Install MagicDisc

Task 2 – Mount the ISO file using MagicDisc

Task 3 – Find the drive using Windows Explorer and investigate its contents

Task 4 – Open FTK Imager and add files.iso as an evidence item

Task 5 – Explore what you see in the drive using FTK Imager

Unidentified Files Exercise – In-Class Version

Course: Acquiring Information from Digital Storage Media, Spring 2011

You work as a librarian for Disco Tech, a university in the Midwestern US. You are working with, Dr. Matilda Bigschat, a prominent senior member of the electrical engineering faculty, who would like to submit her materials to Disco Tech's Research Extending Institutional Repository (RTIR).

When you met with Dr. Bigschat, she told you about her research, which included numerous studies on the efforts of government agencies to respond to technological challenges. She then provided you with a box full of physical media (CDs, floppies, and an external hard drive). In some cases, she could tell you specifically what was on them and why it was important to retain. In other cases, she could only recall that the contents of the media were important to her research, based on where she had filed them in her office.

You examined one of the CDs. There are no labels on the CD itself, but the case containing it has a sticker that reads "Files." You loaded the CD into a drive and used FTK Imager to create a disk image of the CD. You then used FTK Imager to export the files, which you can now find in files.zip in Blackboard. For purposes of tasks 1-4, we'll work from files.zip.

Task 1: Inspection at the file level

- Based simply on the properties and names of the files and directories (don't open any files yet), try to make some inferences about what these files are and how they might be related.

Task 2: Looking at names of specific files

- Find a file called Circular.596
- Do you see any other files that have similar names? If so, make note of which ones they are.

Task 3: Inspection at the bitstream level

- Generate a hex view of the file called Circular.596 – using HView
- What can you infer about this file from looking at the hex representation?
- If you identified other files in task 3, also view those in hex view. Do you notice anything similar?

Task 4: Finding duplicate files

- Visit: http://www.webutils.pl/MD5_Calculator and calculate the MD5 value of Circular.596
- Save the result in a text file.
- If you identified other files in task 3, also generate MD5 values for them. Are the values the same or different?
- If you notice something interesting about the MD5 values, what can you infer about what happened to the files?

Task 5: Investigating the .ISO (disk image) file

- Open FTK Imager
- Go to File | Add Evidence Item…
- Select “Image File”
- Browse to files.iso and then select “Finish”
- Navigate the file tree and discuss what you observe

Curation of Identified Files – State Archives Scenario

Used in Digital Curation for Public Records Professionals Workshop , July 16, 2011

You work at the Archives of State X. In response to deep budget cuts, the executive branch has recently undergone a major reorganization and reduction of personnel. This has included the elimination of the Department of Configuration Management (DCM) from within the state's Office of Information Technology. On her last day of her employment, the assistant to the Director of DCM (who also served as the records officer) has sent you several boxes full of records from the now-defunct unit. She tells you that all of the boxes contain at least some records that have been scheduled for permanent retention by the state archives.

One of the boxes is full of physical media (CDs, floppies, and an external hard drive). In some cases, she could tell you specifically what was on them and why it was important to retain the records. In other cases, she could only recall that the contents of the media were part of the DCM's efforts to compare configuration management efforts across different states in the U.S.

You examined one of the CDs. There are no labels on the CD itself, but the case containing it has a sticker that reads "Files." You loaded the CD into a drive and used FTK Imager to create a disk image of the CD. You then used FTK Imager to export the files, which you can now find in files.zip. For purposes of tasks 1-4, you should work from files.zip.

Task 0 – Preparation

- Obtain the following:
    - Data files:
        - files.zip [http://www.ils.unc.edu/callee/files.zip - save to your desktop]
        - files.iso [http://www.ils.unc.edu/callee/files.iso - save to your desktop]
    - Tool to generate file hashes:
        - md5summer (Windows) [http://download.cnet.com/MD5summer/3000-2248_4-10050856.html- save to your desktop]
    - Hex viewer and editor:
        - Cygnus Hex Editor (Windows) [http://download.cnet.com/Cygnus-Hex-Editor-Free-Edition/3000-2352_4-10448945.html- save to your desktop]
        - Hex Fiend (Macintosh) [From http://ridiculousfish.com/hexfiend/]
- Create a folder on your desktop called dcm-files, then extract the contents of files.zip to that folder
- If you'd like to complete the tasks in the ***optional Task Set 3***:
    - Tool for mounting an ISO file so that Windows treats it like just another drive on your computer:
        - MagicDisc (Windows) [From http://www.magiciso.com/tutorials/miso-magicdisc-history.htm]
    - Tool for generating and doing basic inspection of disk images:
        - FTK Imager [http://accessdata.com/support/adownloads#FTKImager - save to your desktop]

Task Set 1 – Seeing what you have – looking in bigschat-files

Task 1: Inspection at the file level

- Based simply on the properties and names of the files and directories (don't open any files yet), try to make some inferences about what these files are and how they might be related.

Task 2: Looking at names of specific files

- Find a file called Circular.596
- Do you see any other files that have similar names? If so, make note of which ones they are.

Task 3: Inspection at the bitstream level

- Generate a hex view of the file called Circular.596 – using Cygnus Hex Editor or Hex Fiend
- What can you infer about this file from looking at the hex representation?
- If you identified other files in task 2, also view those in hex view. Do you notice anything similar?

Task 4: Investigating the .ISO (disk image) file

- Open FTK Imager
- Go to File | Add Evidence Item…
- Select "Image File"
- Browse to files.iso and then select "Finish"
- Navigate the file tree and discuss what you observe

Task Set 2 – Investigation based on hash values

NOTE: These tasks are based on using md5summer in Windows. You can also generate individual file hashes from the command line in Mac/Unix with the md5 command or use the online utility: at http://www.webutils.pl/MD5_Calculator.

Task 5 - Generate hashes

- Open md5summer
- Select dcm-files as the root folder
- Select "Create Sums"
- In the "Create list of files to sum" dialog, navigate to the dcm-files folder on your desktop, then select all of the contents of that folder and select "Add"
- Click "Ok"
- Confirm that the md5 hashes are generated
- "Save as" to your desktop as bigshat-files.md5
- Leave the md5summer application running

Task 6: Finding duplicate files

- Look at the MD5 value of Circular.596
- If you identified other files in task 2, are the values the same or different?
- If you notice something interesting about the MD5 values, what can you infer about what happened to the files?

Task 7 – Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "dcm-files" directory as your root folder
- At the "Open md5sum file" dialog, select bigschat-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Task 8 – Change file names

- Change the file names of one or more of the files – make a note of which ones you change
- NOTE: Be sure NOT to open the files in an application. You should only change the names.

Task 9 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "dcm-files" directory as your root folder
- At the "Open md5sum file" dialog, select dcm-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors). If you're getting red dots, why might that be?

Task 10 – Opening files with hex editor

- Pick one or more items from the "files" folder and open them in Cygnus Hex Editor or Hex Fiend (you can drag the file onto the Cygnus Hex Editor icon on your desktop)
- Look at the hex view of each file and see what it reveals about the bitstream content of the file

Task 11 – Change content with hex editor

- For one or more of the files, change a bit within the file in Cygnus Hex Editor and then save the changed file
- Note which files you've changed and the **specific place** in the files where you changed bits

Task 12 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "dcm-files" directory as your root folder
- At the "Open md5sum file" dialog, select dcm-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Task 13 – Change content back to earlier state in hex editor

- Open one or more of the files that you changed in step 7, and change the bits back to their previous state, and then save the changed file(s)

Task 14 – Re-Verify hash values

- Go back to md5summer (launch it again if you closed the application)
- Select "Verify sums"
- Select the "dcm-files" directory as your root folder
- At the "Open md5sum file" dialog, select dcm-files.md5 from your desktop and select "Open"
- Note whether there are green dots (verified) by the files or red dots (errors)

Task Set 3 – Two Views of the Same .iso File [Optional – Time Allowing]

Task 1 – Install MagicDisc

Task 2 – Mount the ISO file using MagicDisc

Task 3 – Find the drive using Windows Explorer and investigate its contents

Task 4 – Open FTK Imager and add files.iso as an evidence item

Task 5 – Explore what you see in the drive using FTK Imager

INLS 490-141: Acquiring Information from Digital Storage Media – Abridged Syllabus

One-Credit Special Topics Course
Administered: Spring 2011

COURSE DESCRIPTION

Students will learn about hardware, software and methods used to extract digital data that have been stored on removable media (e.g. hard drives, floppy disks, USB memory sticks). This course addresses common storage devices and interfaces; write-blocking equipment and its role in acquisition of data; levels of representation; basic filesystem structures; role and importance of hash values and hex views of bitstreams; and software used to conduct data acquisition. Students will have the opportunity to use a range of state-of-the-art digital forensics hardware and (commercial and open-source) software and explore ways that they can be applied by information professionals in a variety of collecting contexts.

COURSE OBJECTIVES

- Understand the major levels of representation in digital collections
- Investigate implications of different levels of representation for digital curation activities
- Working familiarity with hardware and software for the acquisition of data from storage media
- Understand implications of digital forensics tools and methods for long-term curation of digital data

COURSE REQUIREMENTS

- Complete required readings and participate in class discussions
- Paper about incorporating digital forensics methods or tools into a digital curation process
- Final exam, administered in class

PART 1 - Conceptual and Practical Foundations

Session 1 - Introduction, Overview and Motivation

- Motivation and Scope
- Applying digital forensics to digital curation
- Overview of digital forensics
- Industry, field, tools and methods

Reading:

Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. "Digital Forensics and Born-Digital Content in Cultural Heritage Collections." Washington, DC: Council on Library and Information Resources, 2010. http://www.clir.org/pubs/reports/pub149/pub149.pdf

Session 2 - Nature of digital objects, layers, components, representation, relationships (working up from the hardware level)

- Explanation of multiple levels of representation

Reading:

Farmer, Dan, and Wietse Venema. "The Spirit of Forensic Discovery (3-15) and "The Persistence of Deleted File Information" (145-160) in *Forensic Discovery*. Upper Saddle River, NJ: Addison-Wesley, 2005.

Session 3 - Three essential forms of representation: disk images, hashes, and hex dumps

- Definition and role of disk images in acquisition of data
- Hashes and Hex views of data
- Role of hash libraries

Reading:

Petzold, Charles. *Code: The Hidden Language of Computer Hardware and Software*. Redmond, WA: Microsoft Press, 1999. [Bytes and Hex (180-189)]

Session 4 - Hardware basics

- Explanation of various ports and drive types
- Showcasing of different hardware options

Reading:

White, Ron and Timothy Edward Downs. *How Computers Work. 9th Edition*. How It Works Series. Indianapolis, IN: Que, 2007. [Data Storage (158-75, 182-3, 186-7)]

PART 2 - Procedures for Acquiring and Transforming Data

Session 5 (Feb 21) - Generating disk images

- Role and use of write blockers
- Software to generate disk images (e.g. dd, EnCase, FTK Imager)

Reading:

Woods, Kam, and Geoffrey Brown. "From Imaging to Access - Effective Preservation of Legacy Removable Media." In *Archiving 2009: Preservation Strategies and Imaging Technologies for Cultural Heritage Institutions and Memory Organizations: Final Program and Proceedings*, 213-18. Springfield, VA: Society for Imaging Science and Technology, 2009. http://www.digpres.com/publications/woodsbrownarch09.pdf

Session 6 - Bulk data extraction, annotation and manipulation using open source software

- Use of fiwalk, bulk extractor

Readings:

Garfinkel, Simson L. "Digital forensics research: The next 10 years." Digital Investigation 7 (2010): S64-73. http://dfrws.org/2010/proceedings/2010-308.pdf

Garfinkel, Simson L. "Forensic feature extraction and cross-drive analysis." Digital Investigation 3S (2006): S71-81. http://simson.net/clips/academic/2006.DFRWS.pdf [Specifically: Sections 1-3, p.S71-75]

Session 7 - Standard digital forensics workflow using commercial tools - Part 1 (FTK)

- Creating a case and loading disk image data
- Basic software features
- Common tasks supported

Reading:

Handouts to be disseminated (copyright AccessData)

Session 8 - Standard digital forensics workflow using commercial tools - Part 2 (EnCase)

- Creating a case and loading disk image data
- Basic software features
- Common tasks supported

Reading:

Handouts to be disseminated (copyright Guidance Software)

Session 9 - Curation of forensic data for long-term access

- Extracting data from commercial tools
- Main options for packaging the data into files - .eo1, AFF
- Data redaction and controlling access

Reading:

Garfinkel, Simson L. "AFF: A New Format for Storing Hard Drive Images." *Communications of the ACM* 49, no. 2 (2006): 85-87. http://simson.net/clips/academic/2006.CACM.AFF.pdf

Session 10 - Synthesis and conclusions

Final exam in class

EndNotes

[1] Kirschenbaum, Matthew G. *Mechanisms: New Media and the Forensic Imagination*. Cambridge, MA: MIT Press, 2008.

[2] Lavagnino, John. "The Analytical Bibliography of Electronic Texts." Paper presented at the Joint Annual Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities, Bergen, Norway, 1996.

[3] Woodyard, Deborah. "Data Recovery and Providing Access to Digital Manuscripts." Paper presented at the Information Online 2001 Conference, Syndney, Australia, January 16-18, 2001.
[4] Ross, Seamus, and Ann Gow. "Digital Archaeology: Rescuing Neglected and Damaged Data Resources." London: British Library, 1999.

[5] John, Jeremy Leighton. "Adapting Existing Technologies for Digitally Archiving Personal Lives: Digital Forensics, Ancestral Computing, and Evolutionary Perspectives and Tools." Paper presented at iPRES 2008: The Fifth International Conference on Preservation of Digital Objects, London, UK, September 29-30, 2008.

[6] Glisson, William Bradley. "Use of Computer Forensics in the Digital Curation of Removable Media." In *Proceedings of DigCCurr2009: Digital Curation: Practice, Promise, and Prospects*, edited by Helen R. Tibbo, Carolyn Hank, Christopher A. Lee, and Rachael Clemens, 110-1. Chapel Hill, NC: University of North Carolina, School of Information and Library Science, 2009.

[7] Garfinkel, Simson, and David Cox. "Finding and Archiving the Internet Footprint." Paper presented at the First Digital Lives Research Conference: Personal Digital Archives for the 21st Century, London, UK, February 9-11, 2009.

[8] Elford, Douglas, Nicholas Del Pozo, Snezana Mihajlovic, David Pearson, Gerard Clifton, and Colin Webb.. "Media Matters: Developing Processes for Preserving Digital Objects on Physical Carriers at the National Library of Australia." Paper presented at the 74th IFLA General Conference and Council, Québec, Canada, August 10-14, 2008.

[9] Underwood, William E., and Sandra L. Laib. "PERPOS: An Electronic Records Repository and Archival Processing System." Paper presented at the International Symposium on Digital Curation (DigCCurr 2007), Chapel Hill, NC, April 18-20, 2007.

[10] Underwood, William, Marlit Hayslett, Sheila Isbell, Sandra Laib, Scott Sherrill, and Matthew Underwood. "Advanced Decision Support for Archival Processing of Presidential Electronic Records: Final Scientific and Technical Report." Technical Report ITTL/CSITD 09-05. October 2009.

[11] Kirschenbaum, Matthew G., Richard Ovenden, and Gabriela Redwine. "Digital Forensics and Born-Digital Content in Cultural Heritage Collections." Washington, DC: Council on Library and Information Resources, 2010. http://www.clir.org/pubs/reports/pub149/pub149.pdf

[12] Thomas, Susan, Renhart Gittens, Janette Martin, and Fran Baker. "Workbook on Digital Private Papers." 2007. Paradigm Project. http://www.paradigm.ac.uk/workbook/introduction/index.html.

[13] Thomas, Susan, and Janette Martin. "Using the Papers of Contemporary British Politicians as a Testbed for the Preservation of Digital Personal Archives." *Journal of the Society of Archivists* 27, no. 1 (2006): 29-56.

[14] Williams, Pete, Katrina Dean, Ian Rowlands, and Jeremy Leighton John. "Digital Lives: Report of Interviews with the Creators of Personal Digital Collections." *Ariadne* 55 (2008).

[15] John, Jeremy Leighton, Ian Rowlands, Peter Williams, and Katrina Dean. "Digital Lives: Personal Digital Archives for the 21st Century >> An Initial Synthesis." Version 0.2. March 3, 2010. http://britishlibrary.typepad.co.uk/files/digital-lives-synthesis02-1.pdf.

[16] Kirschenbaum, Matthew G., Erika Farr, Kari M. Kraus, Naomi L. Nelson, Catherine Stollar Peters, Gabriela Redwine, and Doug Reside." Approaches to Managing and Collecting Born-Digital Literary Materials for Scholarly Use." College Park, MD: University of Maryland, 2009.

[17] Born-Digital Collections: An Inter-Institutional Model for Stewardship (AIMS), http://www2.lib.virginia.edu/aims/.

[18] Lee, Christopher A. "Addressing the Messiness of Electronic Records Acquisition: Discussion of Methods and Proposed Professional Directions." Society of American Archivists Annual Meeting, Electronic Records Section. Austin, TX, August 14, 2009.

[19] Lee, Christopher A. "Bringing Values to the Bitstream: A Framework for Digitally-Aware Professional Ethics of Curation." Society of American Archivists (SAA) Research Forum, Austin, TX, August 11, 2009.

[20] Lee, Christopher A. "Value-Enabled Digital Curation: Toward Ethically Responsible Acquisition." Fifth International Digital Curation Conference, 2-4 December 2009. London, UK. [Poster]

[21] Digital Library Curriculum Project. http://curric.dlib.vt.edu/.

[22] Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum (DigCCurr). http://www.ils.unc.edu/digccurr/.

[23] DigCCurr II : Extending an International Digital Curation Curriculum to Doctoral Students and Practitioners. http://www.ils.unc.edu/digccurr/aboutII.html.

[24] Educating Stewards of Public Information for the 21st Century (ESOPI-21). http://ils.unc.edu/esopi21/index.html.

[25] Closing the Digital Curation Gap (CDCG0. http://ils.unc.edu/gap/index.html.

[26] Supported by NSF Grant DUE-0919593.

[27] BitCurator project, http://bitcurator.net.