Taking the First Step – Data Preservation

Introduction

In any computer forensic or electronic discovery matter, preservation of key electronically-stored information (ESI) ranks as one of the most important steps in the process. Proper preservation of such ESI involves the creation of an accurate rendition of such data including its metadata. Such metadata can include its location and/or the associated data file attributes such as its creation date or last modification data.

In order to properly collect ESI, it is imperative to understand what type of collection needs to be performed. Furthermore, a proper understanding of how the preserved data will ultimately be utilized must be developed since this will dictate the destination file formats that can be employed for the collection. Naturally, this impacts the choice of the collection tool that can be used.

It should be noted that there are other factors involved in making a proper collection. One such factor is the use of write-blocking mechanisms (when possible) to prevent modification of the source data. Without such mechanisms, the process of performing the collection can actually modify such aspects as the source data's metadata, e.g., time and date stamps. Likewise, sufficient documentation and impartiality of the data collector plays an important role.

Preservation Type

Before preservation of any electronic evidence can begin, it is necessary to identify exactly what needs to be preserved. This can be more difficult than first imagined. Along with identifying such information as the custodians of interest or servers involved, it is also necessary to determine if all available data needs to be preserved or just specific active data.

Determining this factor specifies the type of collection to occur: physical or logical. A physical collection involves the imaging of a real-world item, i.e., something that can be touched. The classic example of this is the hard disk drive contained in desktop and laptop computers. On the other hand, a logical collection preserves the virtual constructs that one sees when using a computer. For example, a single physical hard disk can be setup to hold two logical volumes. In Microsoft Windows, these two logical volumes might be called the "C drive" and the "D drive".

A physical collection methodology captures all available information contained on storage media. Such information can include hidden partitions or items in unallocated space. Information in unallocated space can include remnants of items that were once actively used on that system but were subsequently deleted. Such information is invaluable when conducting investigations where the presence of data that should not be

there, e.g., contraband or non-job-related intellectual property, is meaningful. It is for this reason that physical collections are typically employed.

Although not as comprehensive, a logical collection methodology can also be employed. This can be done at two levels. One is at the volume or logical drive level. The other is at the file and folder level. Although very often similar in size, a volume-level logical collection is not the same as a physical device collection. While both will capture ESI in matching unallocated space, a physical collection will always contain more. Thus, a physical collection is typically employed. However, there are several situations where a volume-level logical collection is preferred. One such case is when a volume or drive stretches across several physical hard drives. Another is on encrypted volumes. In both cases, the data when viewed on a volume-level is easier to comprehend than in a scrambled representation when seen on a physical level.

A logical collection on the file and folder level is often employed when physical or volume-level logical collections are impractical or unnecessary. These situations commonly arise when trying to collect data stored on enterprise file servers. In such cases, the business cost of bringing down such a key server can be high while the value of getting information in unallocated space is low. Furthermore, the server may contain mass amounts of data making it impractical to collect it all. Likewise, the target of the collection may only be a single folder or file(s) on a server. This can occur in discovery matters where only one project in a company's portfolio is under dispute.

This methodology can be used in situations where the active files themselves contain sufficient information for the matter at hand. One such case is a discovery matter where deleted information is not relevant. Another involves e-mail where the material information such as the date a message is sent is encapsulated in the file holding the e-mail message. Typically smaller than either physical or logical volume collections, file/folder-level logical collections can often be completed more quickly and more cost-effectively.

However, it must be stressed that file/folder-level collections represent a highly targeted approach. Thus, one must be certain that the specified collection target(s) represent(s) all of the information that must be preserved. For example, a file/folder-level data collection does not retain such ancillary information as what applications were installed on a computer. Such information could be highly useful in the processing of files for discovery. For example, many applications use the same extension for their data files, e.g., DAT. By imaging a computer's entire hard disk, the potential array of applications that stored some ESI into a generic file format could be easily identified. In this case, the end result could be that the ultimate cost encountered for the discovery matter when factoring in the cost of rendering the collected ESI into a usable format is actually less by using a physical collection approach rather than an overly-targeted file/folder-level logical one.

The key piece in determining what type of collection – a physical, a volume-level logical, or a file/folder-level logical – is needed depends on what ESI needs to be preserved. If all possible data needs to be preserved including things like potentially hidden data or ESI in unallocated space, e.g., deleted file fragments, then a physical collection should be employed. A volume-level logical collection should be employed when it preserves a

better representation of the ESI, e.g., a logical volume on a RAID array of several disks. File/folder-level collections should be employed where the collection of such files is sufficient for the ultimate goal of the collection and the cost to acquire the ESI is an issue.

Collection Tools & Deliverables

By determining the type of collection to be performed, the actual result of the collection – or deliverable – is further narrowed. Likewise, the file types of the deliverables can be dictated by the tool used to acquire the data. (See Table 1.)

Although there are other possible file types, physical collections are commonly stored in either dd or EnCase format. The program dd is originally a Unix/Linux tool which can be used to either copy or image almost any input including files and hard disks. Many forensic tools can be set to output a collection in a similar format. The dd format is also referred to as the raw format. The Encase format is the acquisition format of choice for Guidance Software's EnCase forensic tool. Other tools can also store the acquisitions in Encase format, notably Access Data's Forensic Tookit (FTK) Imager. For physical collections, EnCase format files are also commonly referred to as E01 files.

With logical collections, the range of file types of the deliverables is even greater than what is seen with physical collections. This stems from the fact that a variety of different tools can be employed to perform the acquisitions, and each might employ its own collection file format. The available collection tools vary depending on the platform, e.g., Windows, Macintosh OS, or Linux/Unix. Finally, collection tools can differ greatly in such aspects as forensic soundness, performance, and universal-acceptance. For example, unlike EnCase, Ghost does not provide a verifying MD5 hash of its acquisition. In short, the collection tool employed with its resulting deliverable file format often depends on the situation.

Table 1: Summary of Common Tools Available for Collections on Windows Systems.

Tool	Acquires Physical Devices?	Acquires Volumes/Logical Drives?	Acquires Individual Folders/Files?	Collection File Format (Physical/Logical/File)	Collection Format Easily Readable? (Physical/Logical/File)
Encase	Y	Y	Y	E01 E01 L01	Y Y N
FTK Imager	Y	Y	Y	E01/dd (raw)/SMART E01/dd (raw)/SMART AD1	Y Y N
Ghost	Y (special settings required)	Y	Y	GHO GHO GHO	N N N
WinZip	N	N	Y (special settings required)	N/A N/A ZIP	N/A N/A Y
Robocopy	N	N	Y	N/A N/A format of original files	N/A N/A Y

These various collection tools store the acquired ESI in differing file formats – which can vary not only between tools, but also on what is being collected. For example, both

EnCase and FTK Imager can store volume/logical drive level acquisitions in E01 files. However, when using them to perform a file/folder level acquisition, EnCase employs a proprietary L01 files while FTK Imager uses proprietary AD1 files. Each can only be read by programs from their respective publishers. Thus, only EnCase understands L01 files; FTK understands AD1 files; and only Ghost can read GHO files. This may not be an issue if the end goal is merely preservation. However, if forensic analysis is needed, then the former can be readily analyzed by the EnCase while the latter requires – at a minimum – an additional conversion step from a Ghost archive to native files before the data can be loaded into EnCase. This results in additional cost and complexity in cases where analysis is needed. More common tools like WinZip can also be employed. WinZip collects folders and files by encapsulating them into an archive file and can preserve the directory structure of the source data while doing so. WinZip archives also have the added benefit of being readable by a variety of tools. At other times, the only practical tool-at-hand may be Robocopy which preserves file metadata but not folder metadata. As might be imagined, it is important to understand not only what is being collected but also how the collection is performed.

Summary

There are indeed a variety of different ways of preserving data. Likewise, there are a variety of tools available to perform these collections. The choice of which collection to use is guided both by the nature of the data to be preserved as well as the ultimate use of the preserved data.

Which tool is the most appropriate depends on what ESI needs to be preserved. An improperly specified collection can result in insufficient or overly sufficient data being collected – resulting in the loss of a case or imposed sanctions for the failure to produce or, at best, inflated costs. In short, the key to good collections is proper planning and obtaining key information on a matter upfront.