Where Data Resides – Data Discovery from the Inside Out
by Julie A. Lewis

You just landed a major case with a Fortune 1000 company and are beginning discovery proceedings. Your understanding of discovery in the paper world is first rate, but what about all that electronic data that may be relevant to your case? You've been told that 93% of documents and communication occur digitally at your client. Where do you begin? You are an expert in legalities, but the "bits and bytes" of technology is something you never envisioned you'd have to confront when you went off to law school. Take a deep breath, and begin with these tips and tools of the trade. This article provides a high level overview of how data is similar to the physical file world and how it differs. Whether you are a lawyer not well versed in technology or fairly tech savvy, this article will provide you with a basic understanding of data and will allow you to communicate effectively with others without having to become a technologist yourself.

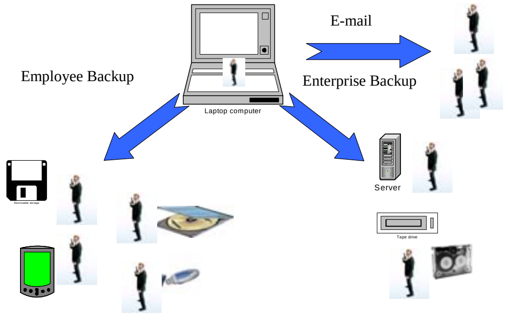
What Storage Media Exists in an Enterprise Environment?

Every client may have a different storage environment that may need approaching in unique ways. Do not assume that your client's storage environment is similar to your law firm's. As you begin your data discovery journey, it is important to proactively strategize where data resides, what electronic evidence is pertinent to the case and how much it will cost to produce the desired information. Although there are lots of applications, let's start with e-mail since it is the most familiar and commonly used mode of communication, and can be stored in many locations (see Exhibit A - The Viral Effect of Data). E-mails are typically hosted on a server. These servers contain multiple e-mail accounts. In the physical world, these servers are comparable to the post office boxes at a postal carrier in that they centralize and store correspondence in a common location with a separate box for each user. The difference is that mail can be copied and forwarded almost instantaneously.

Employee laptops or personal computers (PCs) can also contain copies of e-mails if configured to work off-line. Many users that fly or commute set-up their laptop computers this way to maintain productivity while not logged-in to their company's network. For e-mail to be stored on an employee's computer, the laptop must store data on the employee's hard drive. Sometimes you will also see PCs configured this way if an employee has concerns that the network may go down, but he or she still wants to maintain productivity. Depending on the type of e-mail your client uses, the file extension will vary. For example, Outlook is stored as a ".pst" file; Outlook Express is stored as a ".dbx" file; Groupwise is stored as a ".gwi" file; and Lotus Notes stores e-mail as an ".nsf" file.

Individuals backing up their data will typically use CD, DVD, ZIP disks or a USB/pen/thumb drive. The 3.5" floppy is quickly becoming extinct because of its 1.44 MB storage capacity. Beyond enterprises backing up their data, individuals may back-up their files for a variety of reasons which may include: 1) they don’t want to rely on IT staff for proper backup or retrieval of their data 2) they may want to work on a file from their home computer and avoid transporting their work computer 3) the document retention plan at the company has too short of a retention period and the employee wants to maintain files older than their company’s document retention system allows or 4) they are taking proprietary company data for later use. Individual data and e-mail may be stored on a portable data assistant (PDA). It is common practice to synchronize e-mails, contacts and calendar items with a PDA, so keep in mind that valuable data can also reside on a PDA. Employees may also e-mail files to their personal accounts (e.g. Yahoo, Hotmail and other Web-based e-mail). Also, keep in mind that the e-mail residing in an employee’s inbox may exist within many other user’s e-mail boxes within the company or outside of the company. Investigating where an individual has stored data requires creativity. Similar to your law firm, every individual has unique working styles. Unlike how individuals store backup data, enterprise backup procedures are typically documented and performed on a regular basis.

Exhibit A - The Viral Effect of Data



Enterprises backup data to tapes from servers. These backups can occur at different times, but typically occur at night because they can slow down a server. Some enterprises keep multiple copies of the same data at all times called mirrored copies. This data redundancy is typically found at companies that must have high availability at all times. Extra copies of the data may also be used for data mining and analysis, or just for failover capabilities. Most enterprises will do backups in the evening and will backup from the server used for data mining or failover if it exists. Enterprises may recycle their backup tapes periodically which is typical industry practice. Also, many companies will do incremental backups which means they will perform a full backup once a week or whatever period is set in their backup policy and perform incremental backups for any changes that have occurred since the last full backup. Therefore, when you ask for backup tapes, please keep in mind if the incremental backups are worth restoring. Restoring tapes can cost anywhere from $100-$1,000 per tape. A sampling technique may be appropriate for some of your cases to minimize costs and evaluate if what you’re looking for even exists on backup tapes.

Similar to tapes that you may have used or still use for your car, data is stored on backup tapes in sequential order. Therefore, you may have to look at a lot of irrelevant data before finding what you need. Effective cataloguing by organizations helps in effectively identifying data. Tape can be damaged or corrupted, so there may be times that the data stored on tapes may not be able to be restored for discovery purposes. Each company's organization of files, whether paper or computer, will vary based on how they operate and you'll find that cataloguing of these tapes will differ across organizations.

Some enterprises you encounter will have a robotic tape media handler to automate backup activities and change-out tapes. In more complex backup environments, there are backup servers. There are two types of backup servers to be aware of in some storage environments:

1. Master backup servers that schedule backup jobs and maintain catalogs of backed up data; and
2. Slave backup servers or media servers that control backup storage media.

If an enterprise has a master backup server, the catalogs produced by the master backup server will help simplify your data discovery. Also, be aware that data may be stored at a storage service provider (SSP) or Internet Service Provider (ISP). SSPs and ISPs are third parties that may host and store data for a company. Depending on the state where the SSP or ISP is located, there may be privacy issues that arise. Data mining may also occur offsite to record historical trends or for customer market studies.

Beyond traditional storage mediums, data may also reside in voicemail systems, fax machines, printers and other types of electronic equipment. Storage complexities may exist when you get into database environments, storage networking, mirroring and replication. We've provided a simple overview of storage environments. For complex data discovery planning, we strongly recommend bringing in a consultant to work with you on a data discovery plan. As prudent law firms representing your clients, it will always be a trade-off between cost and benefit in your data discovery efforts. Expect a negotiation session with opposing counsel whether you're representing the plaintiff or defendant. Okay, now that we've covered storage environments (the outside), let's go under the hood of storage (the inside) to the hard drive level. The hard drive is what resides within your computer and is where data is stored.

What is a File System?

A file system is simply an index of files. Similar to the Dewey Decimal Classification System used in a library, a file system is used to organize and keep track of files. Although the operating system provides its own file management system, enterprises sometimes buy separate file management systems from vendors such as VERITAS Software that provide more features, such as improved backup procedures and stricter file protection. In some cases large database companies such as Oracle will provide their own file system and not use the one that came with the operating system. Often this is a choice the database administrator makes upon installation.

When a file is deleted, the file system removes the pointer to the file, so it appears that the file is deleted. However, the file will remain intact until overwritten with new data. With storage cheap and hard drive capacity large, deleted data could remain intact and be discoverable for an indefinite period of time. Also, when a hard drive is reformatted, all the data is still recoverable. Formatting just removes the pointer to the file. Therefore, when an organization's IT department reformats hard drives and sells or donates older equipment, its proprietary intellectual property may be readily available for an advanced computer user to discover. Two MIT students did an analysis on more than 150 hard drives they purchased through eBay auctions and other sources. The students reported finding thousands of credit card numbers and extremely personal information on many individuals. In order to safely eliminate data, enterprises must wipe their drives rather than reformat. By wiping a drive with special software, data is eliminated and overwritten.

What is Metadata?

Metadata is simply data about data – yes, it's a fairly broad term. Metadata is similar to Aunt Bee from the *Andy Griffith Show* sitting on a porch on a hot summer day. Beyond making you a glass of lemonade, she will tell you in excruciating detail about how she made the glass of lemonade, who she saw at the grocery store and more information about her day than you'd ever want to know. When you ask for metadata, make sure you really want to go there. It may be helpful on your fact-finding mission. Other times, it may just be noise.

Some use metadata to refer to properties in documents (e.g. title, author, date created, etc.). You may be surprised to learn that in a Word document, metadata may contain up to the last 10 authors of a document and may show changes to the document even if the user was not using the "track changes" feature. Metadata can also be used to describe formulas in Excel documents. If the discussion is about Web pages, metadata may refer to metatags for the title, description or keywords for the Web site making it easier for search engines such as Google to find. For e-mail, metadata may be the routing information through various IP addresses over the World Wide Web. In the storage world, metadata is information about when "writes" to the storage media occur. Always ask what someone means by metadata if you are confused; it's a word thrown around a lot that can vary depending on the context.

What is Ambient Data?

Ambient data is typically used to describe data that is stored in non-traditional computer storage areas and typically includes deleted files, file slack, volume slack, Windows swap file, unallocated space, stored printer images and Internet artifacts. Finding this data in electronic discovery may or may not be relevant to your case. However, you should be cognizant in your data discovery planning what is available "under the hood".

When a file is *deleted,* it's similar to having a reference to a book erased in the library catalogue system. The book may still be there on the shelf, but the index does not reflect its presence. When a file is deleted, the pointer in the file allocation table is removed and the data remains intact until overwritten. This includes deleted e-mail that was emptied out of the deleted items folder.

File slack is the space between the logical end and the physical end of a file. Operating systems store data in clusters (the allocable units that a file system can allocate). The logical end of a file comes before the physical end of the cluster in which it is stored. The remaining bytes in the cluster are remnants of previous files or directories stored in that cluster. If a file is saved over a previously deleted file, and does not occupy the entire cluster, then residual data from the deleted file will not be overwritten and could be recovered with the proper computer forensics tools. While many can download a computer forensics tool and retrieve data, it is the trained forensic expert that can interpret and testify to how data was located, where it came from, who was involved, etc.

Volume slack exists when physical hard drive storage is partitioned. All storage may not be used. Volume slack is the wasted space that could otherwise be used for storage. Remnants of deleted or old files may exist in volume slack.

The *Windows swap file* is virtual RAM (Random Access Memory) used by the operating system when needed. Instant messaging sessions and other relevant electronic evidence may be found here. The swap file randomly grabs data history. It’s similar to a security camera in a store that randomly takes pictures. Instant messaging sessions and other data may be found here.

Unallocated space is unused disk capacity available for future storage. In a physical library, not all shelf space may be used yet. The unallocated space is the unused shelf space ready for books to be placed and stored.

Stored printer images may reside on a hard drive. Printing involves a spooling process that delays the transmission of data to a printer. This allows a computer user to continue to work on other items while the printing takes place in the background. Print spooling creates temporary files that contain both the data to be printed and sufficient information to complete the print job. Beyond this data that may be found on a computer’s hard drives, some companies have actually gone so far as to purchase Adobe Capture Servers that convert and store printer images in PDF files.

Internet Artifacts include Internet favorites stored as bookmarks, Internet cookies (records user preferences for targeted marketing), Internet history and temporary Internet files stored in cache (fancy word for storage). These artifacts can tell you quite a bit about an individual computer user, their trends and preferences. Most of a user’s Web browsing can be recreated through the use of Internet artifacts. Internet e-mails sent or received from Yahoo, Hotmail or MSN may be able to be viewed here. This can be a powerful tool to consider for data discovery in some cases.

In Conclusion

It was our intent to provide you with a high level overview of storage environments and what exists on a hard drive. True enlightenment is “knowing what you don’t know”. We have touched the tip of the iceberg, so you can make informed decisions on when you need to bring in a data discovery expert and what may be available to you in your future data discovery journeys.