New Jersey Law Journal

VOL. CLXXVII – NO. 11 – INDEX 958 SEPTEMBER 13, 2004 ESTABLISHED 1878

LAW OFFICE TECHNOLOGY

By SILLS CUMMIS EPSTEIN & GROSS P.C.

Basic Computer Knowledge Is Crucial

To respond to electronic discovery or to formulate it, counsel needs to understand computer systems

To respond to electronic discovery or to formulate it, counsel needs to understand computer systems, which may consist of large computers serving many users or personal computers working individually or linked through a network, or both. Counsel should have a basic knowledge of the servers, data storage devices, desktop computers and other hardware that make up the network; the operating system(s) that run the computers; the applications software, such as word processing and spreadsheet programs; and the back-up procedures and media.

The best source for this information is the company's management information system (MIS) or information technology (IT) department – the technicians who have day-to-day responsibility for designing and administering the company's computer system. Counsel for smaller companies without MIS or IT departments may have to rely on their computer system consultant, their vendors, or outside counsel to gain an understanding of their computer system. The technicians and other advisors should be consulted early on and throughout the process of electronic discovery to prevent serious problems later.

Types of Networks

The term "network" refers to the hardware and software that connects computers and allows them to share data. The most common computer networks are "local-area networks" (LANs), which are basically computers located close together, and "wide-area networks" (WANs), which are computers farther apart, connected by telephone, cable lines or radio waves. Networks provide many services, including file, e-mail, Web and database. These services can be provided by computers designated as servers (client/server model) or by any workstation within a network (peer to peer model). In a peer-to-peer network, any workstation can act as a file server. For electronic discovery purposes, this means that counsel should expect to find on any such workstations the scope of documents one would find on a file server.

A client-server network normally has file servers to which users can save documents they create or receive. File servers are centrally located repositories of data containing one or more hard drives. Typically, each user has some private space on those hard drives (home directory) to which to save files. The space is private in that other users can not access documents in a user's home directory. Sometimes the network is set up such that by default, a user's documents are saved to the user's home directory. In other networks, the user may have the option of saving documents "locally," that is, to the hard drive in the user's desktop or laptop computer, often called the "c:/ drive."

Lawyers experienced in electronic discovery no doubt have learned that users don't use home directories just for storing documents. Not infrequently, they archive gigabytes of their old e-mail there as well. Though a company may have only the last sixty days of e-mail on its servers, and though it may have only one year of e-mail back-up tapes, a user may have archived the last four years of e-mail in the user's home directory.

Further complicating electronic discovery, most corporate networks have other file servers, or areas on the file servers that house their home directories, where users can post documents for others to see and work on. These are sometimes called "shared drives" or "group shares." One of the

Sills Cummis litigates e-discovery issues in state and federal courts in New Jersey and other jurisdictions regionally and nationally. Its book on electronic discovery from which this chapter is excerpted, E-Discovery: A Guide for Corporate Counsel, *was published July 1, 2004. © 2004 Sills Cummis Epstein & Gross P.C. www.sillscummis.com.*

challenges to compliance in electronic discovery is filtering the documents of the relevant users out of the group shares, which often contain hundreds of thousands of documents and are normally organized, if at all, by subject, not user. Tiffing[1] the entire group share so that it may be searched can be prohibitively expensive, and using the file server's operating system to search for the relevant users is often too slow. Normally, tailored Perl scripts or other network-forensic tools can be deployed to cull out documents of the relevant users.

In complying with electronic discovery requests, lawyers should also be cognizant that users within a network can save documents to many forms of removable media and external devices such as CD-ROMs, floppy disks, USB thumb drives, jaz drives, PDAs and external hard drives.

Counsel must also consider Internet-based file servers. In some industries, the company's business model, or issues of data storage, require that corporate documents be posted on file servers housed not by the company but by a third party, as on xdrive.com. Documents stored in such fashion are subject to document requests just like documents on servers housed by the company.

Finally, with remote access to computers having blurred the line between home and work, users often store documents on their home computers, sometimes exclusively. Thus, compliance with a document request — whether voluntary or court-ordered — may require that data be harvested from home computers.

To sum up, in the mushrooming field of electronic discovery, the lawyer's responsibility now extends to multiple sources of documents:

· Laptops/desktops
· Home directories
· Group shares
· Removable media
· Home computers
· Internet-based file servers
· PDAs

The inconvenience and cost of gathering, reviewing and producing data from all these sources can be enormous, and the burden can be amplified where the server-based data has been captured in different forms on periodic back-up tapes. One way to reduce the potential burden is to negotiate the scope of compliance with the adversary, whether opposing counsel or a regulator. For example, if you agree to produce only server and desktop/laptop-based documents, you need not worry about the documents resident on Blackberries, Palm Pilots or floppy disks. In civil litigation, where what is good for the goose is good for the gander, an adversary may be content to limit discovery in this manner.

Types of Software

Software is divided into two broad categories: systems software and applications software. The operating system is the master program that runs the computer — it allows the other software programs to function. Windows, Netware, UNIX, Linux, Macintosh OS X and DOS are examples of operating systems. Systems software interacts with the computer at a very basic level. Applications software, also known as "end-user programs," carries out the tasks desired by the users. Word processing, spreadsheets and database management programs are examples of applications software. Virtually all requests for electronic discovery require the responding party to specify operating system and applications software so the requesting party can consider the format in which files and data will be produced.

Types of Data

The three basic types of data are: (1) active data; (2) back-up data; and (3) residual data. Residual data is not visible or accessible to end-users, but it may still exist on the system and be recoverable. Most prominent of the forms of residual data for electronic discovery purposes is the data cached to desktop and laptop hard drives after a document or e-mail is deleted, or after a document or e-mail is drafted, opened or viewed, even if it is not intentionally saved to a hard drive. Discoverable electronic data can also be present in the memories or buffers of the company's printers, fax machines and copy machines. The destruction or even the inadvertent loss of such data after notice of a claim could result in sanctions.

Active Data

Active data is the information stored either on a network server or locally on a hard drive and currently available and readily accessible to end-users through a desktop or other computer connection. It includes information needed for daily tasks, such as word-processing documents, calendars, memo pads, task lists, addresses, e-mail and databases — that is, the data to which you have access when you turn your computer on in the morning. Such data is the primary, most obvious source of active data, but it comprises only a part of it. Also included in active data is hidden information called "metadata" or "embedded data," regarding, among other things, when documents were created, accessed and modified and how they were changed (See below).

Readily available data. Extracting e-mail from e-mail servers is fairly straightforward, but collecting other active data is not. Consider the desktop environment. If users placed all their relevant files in their My Documents folder (and placed nothing else there), harvesting those folders would be simple. But users can and often do place documents in unpredictable places on their hard drives. Therefore, all hard drive directories must be searched. In that search, the examiner may encounter thousands of irrelevant program and system files. Accordingly, a protocol must be drafted (harvesting protocol) to sort each computer's files by file extension and then to extract only the files with those extensions, e.g., .doc, .xls, .ppt, .wpd and .pdf. This takes some sophistication because scores of file extensions are potentially relevant. Some are unique to the client's environment, and the list is ever expanding with the emergence of new applications.

A number of other wrinkles must also be handled. What if the user compressed all the user's .xls files with Winzip, a popular compression program? Then all the .zip files need to be exploded and analyzed pursuant to the harvesting protocol. The problem is that more than fifty types of compressed file formats exist, each with its own extension.

Or what if the user wrote word processing documents with software that saved them as .txt files (stripped-down word processing files with little formatting functionality)? The problem with harvesting all the .txt files is that thousands of system files also have .txt file extensions. In the end, a very careful protocol informed by all these considerations must be structured and then scrupulously followed. These same issues exist when harvesting data from file servers and group shares.

Similar issues arise when dealing with e-mail archived to local hard drives. The local hard drive must be profiled to determine how many e-mail programs the user has used. It may not be just the corporate e-mail application. Often, users transact company business over Internet mail accounts such as AOL, Yahoo! or hotmail and have archives of such mail on their c:/ drives. AOL creates a separate archive for every version of AOL installed, so all such folders must be checked. In addition, certain e-mail programs create hidden files that — unbeknownst to the user — contain e-mail. For example, when a user is synchronizing mail between a corporate Exchange server and a local Outlook client, an ".ost" file will be saved to the user's hard drive. Although the user cannot access that file, the file can forensically be converted to a .pst file, and large caches of past e-mail, calendar items and tasks can be found therein. Accordingly, be careful about representing that you've produced all the e-mail from a corporate environment. There may be more somewhere.

Metadata. Metadata is sometimes described as "data about data." It includes information recording, among other things, when a document was created, last accessed, last modified, and last printed, and who created the document (sometimes defined by the name of the person to whom the application suite that created the document is registered or the name of the person logged onto the computer on which the document is created). Metadata can exist not only at the file system level but also within documents themselves. Windows-based file systems maintain file allocation tables (FATs) (called master file tables [MFTs] in certain Windows systems) that enable the computer to know where saved documents are stored so they can be retrieved. FATs and MFTs also contain metadata that chronicles when a document was created, last modified and last accessed.

The documents themselves — such as Word documents and WordPerfect documents and Excel and Access files — also contain similar metadata, but the metadata that exists within the files may be different from that contained in the FAT or MFT. For example, if a Word document is created and modified in Windows-based network A on January 1, 2004, and e-mailed to a user in network B on January 4, 2004, and the network B user copies the document and saves it to network B on January 4, 2004, the metadata at the FAT/MFT level in user B's network will show a "created date" of January 4, 2004. But the act of saving the document will not change the embedded Word metadata, which will still show a "created date" of January 1, 2004. Also, Word draws all its information regarding created, last accessed and last modified dates from the file system; but Word documents contain other metadata such as last printed, last ten "authors," and last-saved-by that do not depend on the file system, and thus do not change as the document is moved from one file system to another.[2] Different file systems, such as those used by Linux and Solaris, present other metadata issues.

Requests for electronic discovery may require the preservation and production of metadata. A lawyer attempting to comply with such a request may wish to clarify which level of metadata is being requested: FAT/MFT, application-level metadata or both. Compliance with a request to preserve and produce "metadata" must be handled with care because many methods of copying electronic data can change the data regarding last access and even the creation date of the documents being copied. For example, if, after a preservation request, documents are opened in their native applications for lawyers to review for relevance or privilege, the last accessed date will change (both at the FAT/MFT and application levels) to the date the document is opened. Even if documents are copied from a desktop or server in an appropriate fashion, the act of burning copies of those documents to CD can cause new creation dates to appear, and last-accessed dates to completely disappear, on the CD-resident copies. When such data is requested, technical tools and methodologies for preserving metadata should be put in place at the beginning of counsel's electronic document collection. This may require the involvement of a computer forensics expert.

E-mails received from outside a company's network also have metadata in the form of hidden "Internet headers." These headers can normally be revealed by an operation within the e-mail program on a user's machine, such as Outlook or Outlook Express. The headers contain esoteric information not normally relevant to litigation, including the name of the e-mail server from which the e-mail was sent, the name of the e-mail server receiving the mail, the Internet protocol addresses of those servers, and time stamps for when the e-mail passed through those servers. This information is likely to be useful only in a dispute whether a particular e-mail was sent or received. A native-format preservation of e-mails automatically secures this metadata, and, unlike the preservation process for documents, it poses no risk that the metadata will be altered. On the other hand, if only the visible text of an e-mail is preserved, as through a tiffing process, the embedded Internet headers will be lost.

Embedded data. Substantive information created by the user and hidden within the file itself (i.e., not displayed in the default view of the document) is commonly known as "embedded data." Such data includes, for example, the substance of previous edits, formatting commands, links to other files, hidden rows or columns in spreadsheets, or 'electronic stickies,' which are notes or reminders that authors and reviewers leave for each other. Because embedded data can embody lawyer-client communications, work product and other confidential communications, it can be the subject of difficult negotiations in determining what data must be produced. Although electronic discovery productions are sometimes made in a form that strips both metadata and embedded data — such as tiffing or PDFing the documents — the metadata and embedded data still exist in the original, "native" documents. In appropriate cases, courts may order the production of files in native form so that such metadata and embedded data can be examined.[3]

Back-up Data

Anyone who has worked with computers has likely heard the cardinal rule, "Back up your files regularly." The institutionalized adherence to this rule has resulted in a proliferation of sources in which files and e-mails can be located, and it has drastically increased the complexity, uncertainty and cost of responding to discovery requests.

Typically, back-up data is copied by network administrators from a network drive, such as an e-mail server, a file server or a group share, to removable or remote media, such as a disk or a tape, to provide data redundancy in the event of a system failure. Less frequently encountered are a user's own back-ups from hard drives to media such as jaz drives, CD-ROMs, DVDs or other external hard drives. Network back-ups are usually generated on a regular schedule.

Each server environment will likely have a different back-up policy, and the administrator for that particular environment will be aware of the policy. The CIO may have no clue about such specifics. Typically, on smaller servers, the data on the server is backed up in full each day for several weeks, collectively called a "rotation." Each back-up is made to a separate tape. When the rotation is over, the back-up process continues, except that tapes from the prior rotation are re-used, beginning with the earliest tape. In a three-week rotation, tape one is overwritten on day twenty-two, which is the first day of the second rotation.

Larger servers may have too much data for the system to be backed up in full each night. Accordingly, administrators can make one full back-up (usually on a weekend day) and six "incremental" back-ups in which only the files that have changed since the previous day are backed up. To restore system data in case of a server failure, the administrator must restore the most recent full back-up and all subsequent incremental back-ups and add them together.

Companies often retain full end-of-month, end-of-six-month, and/or end-of-year back-up tapes for several years. Regulated businesses do so pursuant to statute or rule. Recently, companies have begun to minimize the number of back-up tapes they maintain to avoid the cost and inconvenience of litigation-related restoration and analysis.

Counsel relying on either written or informal rotation policies should be mindful that back-up tapes slated for discard may actually be retained or that, unbeknownst to the administrator, the back-up process on a particular day failed, leading either to a useless tape or to a tape containing older data that was not overwritten. In this regard, practice and policy may be quite different.

One of the greatest challenges for counsel is the volume of data on back-up tapes and the enormous levels of duplication among tapes. The closer that back-up tapes are in time, the greater the level of duplication. For example, if the user is a pack rat and has thousands of e-mails in the mailbox, a three-week rotation of "full" back-ups could produce a 95 percent duplication rate from one day to the next. The cost of restoring and de-duplicating all three weeks of that e-mail would be enormous.

Accordingly, counsel should negotiate with opposing counsel to make the production responsive but cost-effective. For example, if the subject events preceded the rotation period, then perhaps only the earliest back-up tape needs to be restored. Because e-mail attachments can increase the size of production by many multiples, producing just e-mails first can make sense, with selected attachments in a second round.

A company's back-up or preservation protocols will not ensure that every file is backed up at least once. The system will have gaps because a file can be received, reviewed and deleted before it is captured by a regularly scheduled network back-up. For instance, if an e-mail was received in the early morning and deleted immediately, and the company backs up e-mails at the end of the day, such an e-mail would not have been backed up. Thus, a company's back-up system may be under-inclusive as well as duplicative.

Several variables impact the accessibility of back-up data. One such variable is the method of storage. Back-ups are frequently stored on tapes that contain large quantities of data arranged in linear fashion and therefore not easily searched. Types of such tapes include the older reel-to-reel tapes that afford high capacity and are often used in conjunction with large mainframes; digital audio tapes (DAT); QIC tapes; and the newer format Travan tapes often used on smaller computers and networks. Optical drives and archival optical disks are also used as back-up media and present fewer problems in electronic discovery because the data on optical drives and disks does not have to be accessed in linear fashion and does not have to be restored to be searched.

Another variable that impacts the accessibility of back-up data is whether the data has been "compressed." To fit as much data as possible on a storage device, back-up programs normally compress the data, fitting more data into a smaller space but requiring the data to be decompressed and restored to a host drive for accessing. This is not difficult, but it can be time-consuming.

One of the most dangerous pitfalls in electronic discovery is the failure to remove relevant back-up tapes from rotation after a duty to preserve arises.

As a result, relevant information can be overwritten, possibly subjecting the company to sanctions. The relevant tapes should be taken out of rotation and placed in a secure location, with the administrator substituting clean, blank tapes for the ongoing rotation.

In sum, to navigate the universe of back-up data effectively in connection with electronic discovery, one must have a comprehensive understanding of a company's back-up procedures and protocols. This includes the intervals at which back-ups are generated, the media used and the retention period.

Residual Data

It is a common misconception that deleting a document removes it from the computer, or more accurately, the hard drive. When data is "deleted," the space on the storage device on which the data was stored is made available by the file system to store other data. Although the deleted data is not readily retrievable through normal end-user operations, it is not actually erased until it is overwritten by new data. Whatever portion of the deleted data is not overwritten remains recoverable for as long as the hard drive functions.

Only through the use of forensic utilities can this residual data be recovered. The odds that an item of deleted data will be overwritten increase with the volume and frequency of use of the computer. This is why it rarely makes sense to forensically image computer servers, as opposed to desktops and laptops. Servers often run at near-capacity, and even where substantial capacity remains, it is quickly consumed by a high level of use. In addition, compression techniques that fit more data on servers also result in overwriting. For all these reasons, remnants of deleted and unsaved data are quickly overwritten and rendered unrecoverable.

Even if data is not intentionally saved by the user to a c:/ drive, the computer's operating system can save documents or e-mails that the user merely opens, views, or drafts without saving. This "cached" material can be recovered only forensically. But as with deleted material, these forensically recovered fragments can be — and have not infrequently been — the "smoking guns" that resolve a litigation.

The effort of forensic utilities to recover this deleted or unsaved material may be frustrated by using "wipe" programs that overwrite the "unallocated free space" and "slack space" (the portions of the drive where deleted and unsaved data reside) with zeros, ones or other junk data. These programs, especially the free ones, are not foolproof, and sometimes they do not fully accomplish their overwriting agenda. Even if they do, forensic utilities can usually establish the date of their use. If a company has used a wipe utility after the company's duty to preserve arose, issues regarding sanctions may be raised.

An unresolved issue at the cutting-edge of electronic discovery is whether a litigant in a routine civil litigation has any duty to preserve the unallocated free space and slack space on hard drives so that forensic analysis can later retrieve deleted and unsaved items. This would impose a great burden on companies and would require full forensic images of each such desktop and laptop computer. The current wisdom is that such a burden is not required, but the law in this area is evolving.

Footnotes:

[1] Tiffing and PDFing are technologies that create an image file, which is a snapshot of the surface of a document and does not include any hidden data. Tiffs and PDFs can be opened only in specialized viewers such as Microsoft Imaging and Adobe Acrobat.

[2] The metadata displayed in the Word Properties box (revealed by selecting File on the standard Word toolbar and then selecting Properties) includes both the file system metadata (under the General tab) and Word metadata (under the Statistics tab). Other Word metadata — such as last-saved-time and last 10 authors — are not even reported in this view. They are discoverable only through the use of forensic tools such as Metadata Assistant.

[3] Metadata is sometimes defined to include embedded data, or aspects of it such as prior edits, because such data can be considered "meta" to what the end-user sees on the screen or what the end-user can easily access with a click of the mouse. ■