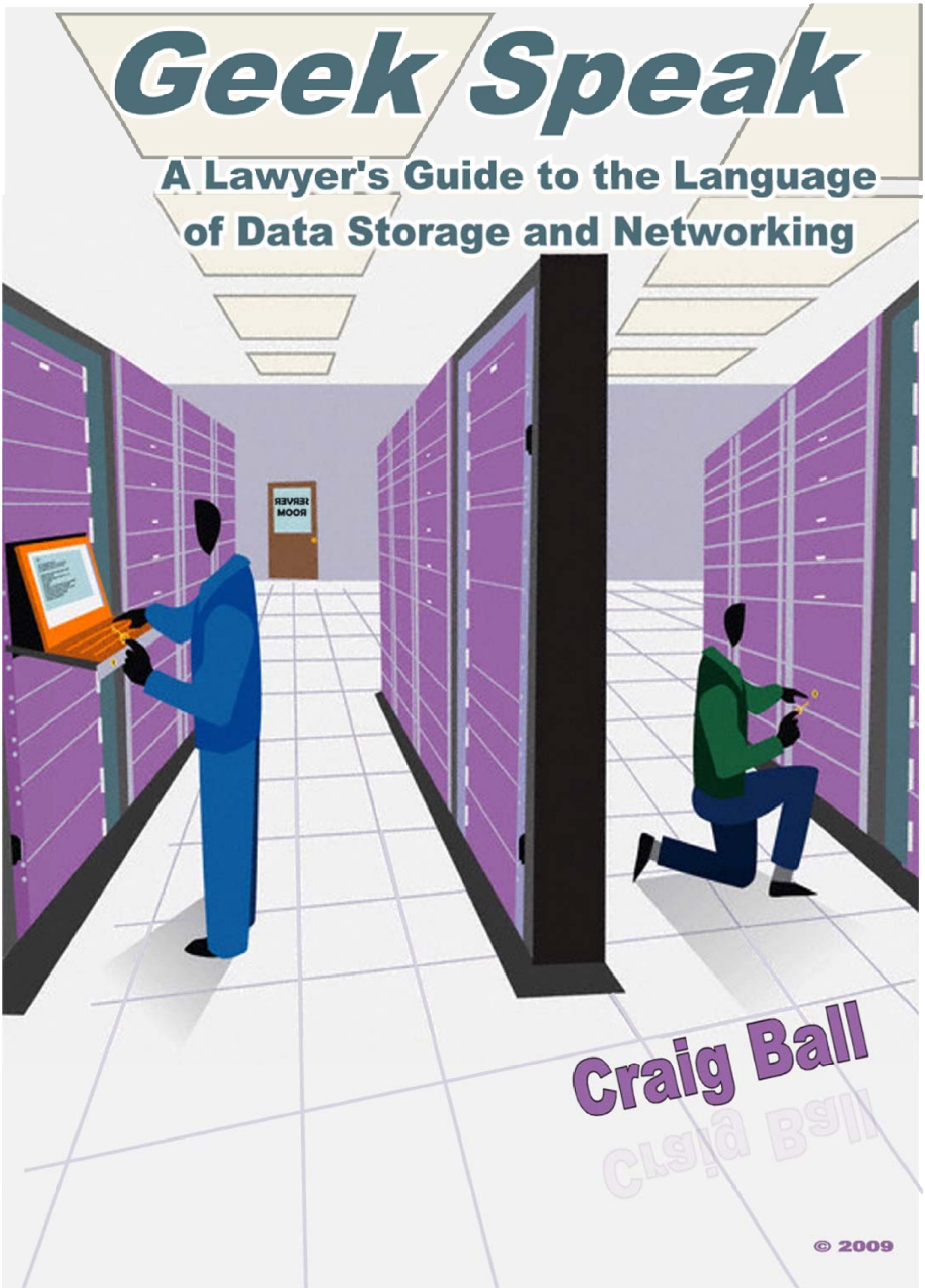
Geek Speak

A Lawyer's Guide to the Language of Data Storage and Networking



Craig Ball

© 2009

Speak Geek
A Lawyer’s Guide to the Language of Data Storage and Networking
Craig Ball

In 1624, when John Donne mused, “No man is an island,” he could scarcely have imagined how connected we’ve become. The bell not only tolls for thee, it beeps and vibrates, too. No iPhone is an iLand.

Networks are the ties that bind our global village and make the world flat. Without networks, our laptops, iPods and Blackberries are just pricey pocket calculators. Networks also transit and store much of the electronic evidence sought in electronic discovery. This article looks at network architecture and data storage devices in the form of an occasionally irreverent glossary offered to help lawyers be at ease discussing the technology of electronic discovery.[1]

Dealing with electronically stored information (ESI) is like living with a teenager—always running in, changing its clothes and heading out again, tracking metadata all over the carpet! But litigants and lawyers aren’t relieved of the duty to find and collect potentially relevant ESI just because it’s flitting about and messy. They’re still obliged to track down the data and make sure it’s safe from harm and will stay put (or come home) until needed in discovery. Rooting out responsive data begins with knowing where to look and the right questions to ask, so it helps to have a working knowledge of the terminology of data storage and networking.

Storage and Network and Memory, Oh My!

Though the terms “storage” and “network” are surely familiar, the technologies they describe take many forms, prompting some confusion. Many mistakenly refer to data *storage* devices like hard drives as “memory.” Hard drives are *storage*; that is, any non-volatile and semi-permanent electronic, optical, mechanical or magnetic device into which data can be entered and subsequently retrieved on demand. Storage is also a location on a network that enables access to storage devices. *Memory* is a term that should be reserved to devices, particularly *Random Access Memory* or *RAM*, where data resides temporarily during processing but is typically lost or overwritten when an application closes or power is interrupted.[2]

A “network” can be any number of computers or devices connected for the purpose of sharing information or capabilities. The largest and most widely used network is, of course, the Internet; but, businesses and homes deploy Wide or Local Area Networks (WANs or LANs) to share

[1] For a more comprehensive (and sober) glossary of e-discovery terms, download The Sedona Conference Glossary for E-Discovery and Digital Information Management (2nd Ed.) from http://www.thesedonaconference.org/dltForm?did=TSCGlossary_12_07.pdf

[2] The line between storage and memory is getting harder to find. Non-volatile flash *memory* is widely used as a means of data *storage* in cameras, thumb drives and solid state drives. Flash memory has almost entirely supplanted photographic film, and solid state drives will soon replace hard drives in laptops and MP3 player. Moreover, it’s unclear how *long* information must be “stored” to be called electronically *stored* information. One case has lawyers worried that the interval may be measured in mere nanoseconds. *Columbia Pictures, Inc. v. Bunnell*, 245 F.R.D. 443 (C.D. Cal. 2007) (defendants ordered to produce contents of RAM).

databases, mail systems, applications, printers and Internet service. There can be a lot of overlap. WANs may be composed of multiple LANs and connect to the Internet.

B

Backup

Although sharing information and resources is the raison d'être for networking generally, an imperative for business networking is the ability to backup many user's data from a single location. Without networking and the mapping of users' storage areas to networked storage devices, users must periodically backup their own data—a responsibility consuming many hours and fostering tragic outcomes.

With networking, each user can be allotted space on a common storage server and the network configured to route that user's activities to the assigned storage location when the user logs on. The user's machine may be configured to assign a specified drive letter (e.g., M:) or folder name to the user's networked storage location. Because the network storage *device* is shared among many users, its allotments are called ***network shares***. But these user-assigned storage areas are typically not "shared" with (i.e., *accessible* to) multiple users. Still other allocations may be open to all or just particular users granted access privileges.

With many users' critical data consolidated in a single locale, albeit in discrete "shares," it falls to the ***information technology (IT)*** staff to insure that all that data gets thoroughly and reliably duplicated at regular intervals to protect against its loss as a consequence of system failure or other disaster. Ideally, the duplicate data is physically or electronically transported to a distant secure location unlikely to be affected by the disaster and is then used to get the downed machines back up again; hence the duplicates are called ***backups*** and their use is termed ***disaster recovery***.

Because it's cheap, durable and portable, magnetic tape is the most common medium used for backup, although remote duplication (***mirroring***) to other network storage devices is fast becoming a viable alternative as hard drive costs plummet. To save time and space, backup regimens seldom copy commercial software programs that can be reinstalled from other media. More time and space is saved—along with network bandwidth--by only occasionally making ***full backups*** of all user created data, opting instead to create more frequent ***differential backups*** of files created or changed since the last full backup and ***incremental backups*** of just what's been created or changed since the last incremental backup. When disaster strikes, the full, differential and/or incremental sets are pieced together like Humpty-Dumpty, a process called ***tape restoration***.

Businesses only need disaster recovery data for a brief interval because no business wants to restore its systems with stale data. Accordingly, the only backup tapes essential for recovery are the last complete, uncorrupted set before the river rose. As a cost savings practice, older tapes may be reused by overwriting them with the latest data, a practice called ***tape rotation***.

In practice, companies may keep backup tapes well beyond their utility for disaster recovery--often years longer and occasionally past the companies' ability to access tapes created with obsolete software or hardware. These ***legacy tapes*** are business records—sometimes the last surviving copy—but afforded little in the way of records management. Even businesses that overwrite tapes every two weeks replace their tape sets from time to time as faster, bigger options hit the market. Consequently, old tapes get set aside and forgotten in offsite storage or a box in the corner until their existence is uncovered in discovery.

Backup tapes store data in significantly different ways than the computer systems they protect. Further, large complex enterprises demand large, complex backup systems protecting hundreds of servers. Such backup systems may occupy room-sized silos where robotic arms ceaselessly cycle through thousands of tapes, and databases are required just to track their convoluted contents. This is an arena where broad brush e-discovery efforts go horribly awry and where transparency, close analysis and well-honed choices are vital. Cooperation between opposing sides is essential, and Judges should tread carefully before issuing orders with untoward costs and consequences.[3]

C

Cache

Downloading data over a network is slower than accessing data on a local hard drive, so networked computers sometimes store or "cache" data obtained from the network to avoid the need to download the same data when later needed. Used as a noun, a cache is an area where oft-used information is stored to facilitate its faster access. Devices like hard drives and processors use caching to improve performance, as do certain software programs. For example, Windows computers running the Internet Explorer web browser use a file cache on the local hard drive called Temporary Internet Files which (with some exceptions) holds the HTML code and images of each web page viewed on the machine until the cache is full or emptied by the user. Users revisiting a cached website experience faster page loads because the browser can pull identical data from the cache instead of downloading it from the Web. Though this requires the system to compare the network and cached data to determine if the network data has changed, caching is still faster than needlessly downloading the data a second time.

From the standpoint of electronic discovery, information in the Temporary Internet Files cache may be relevant, especially where Internet usage is at issue or where data (like web mail) may not be available from more accessible locations.

Client

A client, as in ***client-server model***, is a program, computer or other device that connects via a network to another computer or device called the ***server***. Internet browsers are client applications that obtain web pages from web servers. Microsoft Outlook is an e-mail client that connects to e-mail servers like Microsoft's Exchange server. When the client is a personal computer and

[3] Judges and counsel may find value in Ball, *What Judges Should Know about Discovery from Backup Tapes* (2008); Available at http://www.craigball.com/What_Judges_Backup_Tapes-200806.pdf

performs much of the processing of the data, it's ungraciously called a ***fat client***. When the client device or application cedes most processing to the server, it's called a ***thin client*** (or even a ***dumb terminal*** when it has no processing or local storage capabilities at all).

Cloud Computing

Cloud Computing refers to reliance on web-based tools and resources to supplant local applications and storage. It encompasses ***Software as a Service*** (***SaaS***), where users "lease" programs via the Internet (Google Apps is a prime example), as well as the much-touted, yet elusive ***Web 2.0--***a catchall for all manner of web-enabled phenomena: ***social networking, blogs, wikis, Twitter, YouTube, Google mashups*** and arguably any web-centric venture that survived the great dot-com meltdown.

Gen Xers and Millennials embrace "cloud computing" as if they invented it, but Boomers knew cloud computing when it was called client-server or thin client. Then as now, it was screens and keyboards talking to Big Iron elsewhere, the latter doing the heavy lifting. With SaaS and Web 2.0, we've come full circle and are richer for the journey. As cloud computing takes hold, the bits and bytes of our lives will again move out and get their own places, this time in the ether, but we'll have their cell numbers and can call when we need them.

Cloud computing creates new opportunities in e-discovery because the candid, probative revelations once the exclusive province of e-mail now flood ***MySpace*** and ***Facebook***. But cloud computing creates new challenges for e-discovery because it's harder for employers to isolate and search custodial collections without physical dominion of the storage devices and their users' log in credentials. Additionally, repatriation of cloud content depends on the compatibility of cloud formats with local storage formats, including the ability to preserve and produce relevant metadata. Consider ***Gmail***. Though it's feasible to download Gmail messages into a local mail client application like Microsoft Outlook using Gmail's POP3 support feature, the functionality, searchability and some associated metadata will vary between cloud and local counterparts.

Collection

As a noun in e-discovery, collection refers to any discrete set of electronically stored information, particularly the set amassed after targeted retrieval and culling efforts have occurred. However, it's not uncommon to hear parties speak of their entire universe of ESI as the "collection." For this reason, it's important to define the parameters of any ESI collection to insure common expectations.

Container Files

Sometimes called ***compound files***, container files hold other files, often in compressed, encrypted or proprietary formats or nested—container-within-container--like Russian matryoshka dolls. Container files commonly encountered in e-discovery include compressed Zip and RAR archives, Outlook PST and OST mail files and Lotus Notes NSF mail files. Container files can severely distort document volume estimations as a function of data volume, e.g., a one gigabyte mail container can easily hold tens of thousands of messages and attachments.

Custodian

A custodian is a caretaker, and in the context of e-discovery, the term refers to a person who holds or is charged with overseeing and maintaining potentially relevant information, whether stored electronically, on paper or by other means. For litigation purposes, one is the custodian of his own e-mail, locally and server-stored documents, voice and electronic messaging, smart phone data and any other information to which he has a right of ownership, access or control, including information in the hands of third parties over whom he may exercise direction or control. Custodian also refers to the persons to whom legal hold notices are directed.

Identifying custodians becomes particularly important when ESI is resides in shared network repositories and no one person bears the duty to preserve, search or produce the data. When *everyone* is responsible, often *no one* steps up. Accordingly, efforts to identify potentially responsive ESI should always inquire into the existence of, or rights of access to, shared repositories.

D

Database

A database is a structured collection of records or information organized according to a framework called a ***data model*** or ***schema*** that typically facilitates search and recall of the records using ***query language***. Massive, costly and enormously complex, databases play vital roles in most large enterprises. For companies like Google, Amazon.com and e-Bay, databases serve as the nexus of virtually all operations. Yet, databases come in all sizes and forms, for tasks as varied as balancing checkbooks, organizing family photos and tracking stock portfolios. Even many common file formats are structured as databases, including Microsoft Outlook mails containers and Adobe Acrobat PDF files.

Databases are the most important resources shared across networks, and they also serve as repositories for much information of importance in e-discovery. Many transactions and documents that would once have been memorialized on paper now exist solely as disparate records stored within databases. Because databases assemble documents on-the-fly and are constantly being updated and purged, they can be particularly challenging sources from which to preserve, isolate and produce responsive data. E-discovery from databases requires detailed assessment of the contents, users, capabilities, applications and schema. Responsive contents may need to be extracted using queries constructed expressly for the purpose of isolating evidence and protecting privileged or confidential content, and the form of production is a key consideration, as many requesting parties lack the hardware and software to assimilate database contents in its native format.

Distributed Data

Distributed data might also be called "willy-nilly data," in that it describes all the potentially responsive ESI that's not on the server, but is strewn across laptops, handheld devices, external hard drives, flash drives, CDs, DVDs, home machines, online storage and webmail. Distributed data is costly to collect and sometimes difficult to process because it tends to be the most idiosyncratic ESI and that most prone to obstructive intervention by custodians. A common mistake in e-discovery is assuming that the responsive ESI is on the server without taking reasonable steps to preserve and assess (even by sampling) the contents of distributed data sources.

Domain

A domain is a group of networked computers (typically in the same physical facility) that share common peripherals, directories and storage areas. E-mail systems are customarily organized and backed up by domain.

Domino Server

A Domino server is a network-accessible computer holding users' centralized e-mail stores and employing the IBM Lotus Notes e-mail application. If an IT person mentions the company's Domino server (and you aren't discussing pizza delivery), be prepared for Lotus Notes e-mail and the unique e-discovery challenges and opportunities it entails.

E

ECM

Enterprise Content Management is an umbrella term describing a range of technologies designed to help companies identify, access and use the information stored in their documents, photographs, video, web content, databases and e-mail, especially siloed repositories and unstructured content that tends to be unavailable or difficult to access companywide. ECM applications tend to encompass document management and version control, integration of paper records, records management and retention, web content management and collaboration tools. The most familiar implementation of ECM is probably Microsoft's SharePoint Services (MOSS and WSS).

From an e-discovery perspective, the consequences of a substantial ECM implementation are manifold. ECM may operate at cross-purposes with—or at least complicate--legal hold obligations. Further, collaborative environments are heavily dependent on metadata to support functionality, making preservation and production of a broad range of metadata essential to meet the obligation to produce ESI in reasonably usable forms. Within some ECM environments, documents exist in untraditional and proprietary formats necessitating new and creative approaches to selecting forms of production that preserve look, feel and function of multimedia and informational content. On the positive side, a successful ECM system should facilitate cost-effective identification and search of responsive ESI (though cynics might suggest that savings will be offset by having to deal with all the potentially responsive ESI that ECM makes impossible to ignore).

Enterprise

Enterprise is variously the flagship Federation starship commanded by Captain James T. Kirk, a low cost rental car company favored by skinflint insurance carriers or, in e-discovery, the term of choice when "company" or "business" are insufficiently pretentious.

Ethernet

A set of network cabling and communication protocols for bus topology[4] local area networks. That is, an agreed-upon set of instructions, akin to a language, that permits devices to exchange information. If that's not helpful, think of it as the *other* way computers talk to each other when they're not speaking Internet (TCP/IP).

Exchange Server

An Exchange server is a network accessible computer holding users' centralized e-mail stores and running the Microsoft Exchange e-mail and calendaring application. Typically, users access Exchange servers with Microsoft Outlook mail clients. Microsoft Exchange accounts for 65% of market share among all organizations, with significantly larger shares among businesses with fewer than 49 employees and those in the health care and telecommunications sectors. Consequently, Exchange Server e-mail crops up in the overwhelming majority of cases and understanding its architecture is an essential e-discovery skill.[5] ***See also*** the discussion of Microsoft Outlook, ***infra.***

Extensible Markup Language (XML)

Extensible Markup Language or XML provides a basic syntax that can be used to share information between different kinds of computers, applications and organizations without first converting it. XML employs coded identifiers paired with text and other information. These identifiers can define the appearance of content (much like the Reveal Codes screen of WordPerfect documents) or serve to tag content to distinguish whether 09011957 is a birth date (09/01/1957), a phone number (0-901-1957) or a Bates number. Plus, markup languages allow machines to talk to each other in ways humans understand.

Like multilingual speakers agreeing to converse in a common language, as long as two systems employ the same XML tags and structure (typically shared as an XML Schema Definition or .XSD file), they can quickly and intelligibly share information. Parties and vendors exchanging data can fashion a common schema tailored to their data or employ a published schema suited to the task, such as that under development by the Electronic Discovery Reference Model. [6]

Extranet

An extranet is a private network made available via the Internet to a select group of users, typically customers or suppliers. When used to support transactions, extranets are often called ***virtual deal rooms***. Extranets are increasingly used as a collaborative tool in e-discovery and as a host repository for ESI. Access may be secured by use of a VPN connection or by a conventional link employing user ID and password alone.

[4] See "Topology," *infra*, for further discussion of network topologies.

[5] For a more detailed discussion of Exchange Servers and e-discovery, see Ball, *Meeting the Challenge of E-Mail in Civil Discovery* (2009) at p.25 et seq., available at http://www.craigball.com/em2008.pdf

[6] http://edrm.net

F

File Server

File servers, the heart of any client-server network, are computers typically equipped with fast, redundant storage devices that store and deliver each user's files and other data. Very small networks may not use dedicated file servers but instead allow workstations to share data amongst themselves in a peer-to-peer configuration.

FTP

File Transfer Protocol or FTP is a set of standards and instructions that permit transfer of files between networked computers, most often via the Internet. You'll encounter FTP in e-discovery both as a potential repository to be explored for "orphaned" responsive data not available from other accessible sources and as a mechanism to transfer large volumes of data to and from clients and e-discover service providers.

G

Gateway

A gateway is a combination of hardware and software that allows two networks to communicate. A gateway is essentially a protocol translator that enables, e.g., the wireless network in your home to communicate with the Internet. In this role, the gateway is also called a ***router***.

H

Hub

A hub allows multiple computers to share a network connection, not unlike a power strip allows multiple electrical devices to share AC power from an outlet. Hubs support simple peer-to-peer networking between computers.

I

IM

Instant Messaging or IM is a form of real-time textual communication between two or more persons where such messages are carried by the Internet or a cell phone network. It is the instantaneous receipt and response of IM and its evanescence that distinguishes IM from e-mail. Though relevant, non-privileged IM messages are as subject to preservation and production duties as any other evidence, IM messages typically reside only on the local device sending or receiving the message, not on network servers, and not in active data unless the user has enabled message logging. Accordingly, litigants obliged to preserve IM traffic must either compel message logging and periodic collection of the logs or implement a packet capture mechanism to scan for IM traffic and snare and copy messages as they enter and leave the company's Internet gateway. Neither method is wholly satisfactory.

When a company obliged to preserve IM traffic fails to do so, the data loss may be mitigated by collection from other parties to the dialog or by forensic examination of the machines or devices employed, although recovery of message traffic is by no means assured.

Internet

You're not *really* going to make me define Internet, are you? Where have you been the last 15 years?! Okay, if you insist.

Turning to none other than the august personage of former Alaska Senator (cum felon) Ted Stevens in a speech delivered on June 28, 2006 as chairman of the Senate Committee on Commerce, Science and Transportation:

> [T]he Internet is not something that you just dump something on. It's not a big truck. It's a series of tubes. And if you don't understand, those tubes can be filled and if they are filled, when you put your message in, it gets in line and it's going to be delayed by anyone that puts into that tube enormous amounts of material, enormous amounts of material.

So, the Internet is a series of tubes, not a big truck, and it's best to keep a plumber's helper at hand while Web surfing.

Intranet

An intranet is a private web site, typically reserved to the exclusive use of an organization's employees or members. Intranets tend to be hosted internally on a local access network, but may be Internet-enabled so as to permit secure connections by authorized users via the Internet.

IP Address

An Internet Protocol or IP address is a unique series of four numbers joined by periods and sometimes called a Dotted Quad. It is the numerical designation of the host system that connects you to the Internet and is cross-referenced to the domain name such that either the name or the number can be employed to correctly designate your host system. An IP address can also serve as a unique identifier for computers and other Web-enabled devices on a network employing the standard TCP/IP protocol that serves as the basic computer-to-computer language of the Internet. For example, the IP address of the computer used to write this article is 192.168.0.189.

IP addresses can be useful in e-discovery when constructing a company's data map. Using IP addresses, machines claimed to exist can be correlated against those actually connected to a network. An IP address can also tie ESI to a particular device and, thus, a particular user.

ISP

An Internet Service Provider or ISP is a business or other entity that supplies Internet access via dial-up, cable modem, DSL or ISDN lines or dedicated high speed connections. ISPs routinely host their customers' e-mail accounts and thus may be a source of ESI by subpoena or constitute a third party custodian who should be put on notice of legal hold obligations.

J

Journaling

Journaling is a means of archiving electronic messages, principally e-mail, but potentially IM and VM, too. A journaling mail server copies all messages or, per established rules, certain incoming and outgoing messages to a mailbox or storage location serving as the journaling repository. Journaling serves to preempt ultimate reliance on individual users for litigation preservation and regulatory compliance. Properly implemented, it should be entirely transparent to users and secured in a manner that eliminates the ability to alter the journaled collection.

Accordingly, journaling is a valuable safety net for companies obliged to preserve e-mail because of litigation or regulatory obligations, and counsel should inquire to determine if journaling was enabled, as journaled e-mail traffic can mitigate custodial preservation errors and misconduct. Journaling also helps protect the company against rogue employees seeking to conceal wrongdoing by destroying their e-mail stores before leaving.

Exchange Server supports three types of journaling:

- Message-only journaling, which does not account for blind carbon copy recipients, recipients from transport forwarding rules, or recipients from distribution group expansions;
- Bcc journaling, which is identical to Message-only journaling except that it captures Bcc addressee data; and
- Envelope Journaling which captures all data about the message, including information about those who received it.

Envelope journaling is the mechanism best suited to e-discovery preservation and regulatory compliance. Unlike messages preserved after delivery, journaled messages won't include metadata reflecting the addressee's handling of the message, such as foldering or indications that the message was read.

Journaling should be distinguished from e-mail archiving, which may implement only selective, rules-based retention and customarily entails removal of archived items from the server for offline or near-line storage to minimize strain on IT resources and/or implement electronic records management. However, Exchange journaling also has the ability to implement rules-based storage, so each can conceivably be implemented to play the role of the other.

L

LAN

A Local Area Network or LAN is an interconnected group of computers typically situated in a single location and connected by cable or wirelessly. LANs tend to be used in offices and homes to share Internet connections, files and printers, though they may also be configured to exchange e-mail internally.

Lotus Notes

Lotus Notes is an IBM client application supporting e-mail, calendaring, web browsing and a host of collaborative features. Notes works in conjunction with an IBM Lotus Domino server, although it can also be configured to retrieve e-mail from Microsoft Exchange servers. Though Lotus Notes reportedly has just a 10% overall market share, it enjoys a much higher percentage base among manufacturers with at least 5,000 employees, and IBM claims it has sold 140 million Notes licenses worldwide. Still, the relative infrequency with which E-discovery service providers encounter Lotus Notes means that not all providers are equipped or experienced to process Notes content.

Unlike Microsoft Exchange, which is a purpose-built application designed for messaging and calendaring, Lotus Notes is more like a toolkit for building whatever capabilities you need to deal with documents—mail documents, calendaring documents and any other type of document used in business. Notes wasn't designed for e-mail—e-mail just happened to be one of the things it was tasked to do.

Notes is database-driven and distinguished by its replication and security. Lotus Notes is all about copies. Notes content, stored in ***Notes Storage facility*** or ***NSF*** files, is constantly being replicated (synchronized) here and there across the network. This guards against data loss and enables data access when the network is unavailable, but it also means there can be many versions of Notes data stashed in various places within an enterprise. Thus, discoverable Notes mail may not be gone, but lurks within a laptop that hasn't connected to the network since the last business trip.

M

Mail Client

A mail client is any software application used to prepare, send, receive and read e-mail. E-mail clients can be rudimentary or, more common today, feature-laden productivity tools like Microsoft Outlook or Lotus Notes, which offer a sophisticated and highly-customizable interface. The configuration of a user's mail client may determine whether messages are stored locally, on the mail server or in both places. Additionally, the mail client records and manages key metadata detailing a user's handling of e-mail, including the user's folder structure and various flags indicating whether*, inter alia*, the user opened a particular message, tied it to a calendar entry or flagged it for action.

Microsoft Outlook

Microsoft Outlook is an e-mail client and calendaring tool coupled with several other productivity features to comprise a personal information manager (PIM) toolset. Outlook serves as both a standalone mail client compatible with all mail protocols in common use, but in business, it's usually deployed in conjunction with ***Microsoft Exchange Server*** or, lately, ***Microsoft Office SharePoint Server*** (MOSS).

Despite the confusing similarity of their names, Outlook is a much different and substantially more sophisticated application than Outlook Express (now called Windows Mail). One of many important differences is that where Outlook Express stores messages in plain text, Outlook encrypts and compresses messages. The most significant challenge Outlook poses in discovery is the fact that all of its message data and folder structure, along with all other information managed by the program (except the user's Contact data), is stored within a single, often massive, database file with the file extension .pst. The Outlook PST file format is proprietary and its structure is poorly documented, limiting your options when trying to view or process its contents to Outlook itself or one of a handful of PST file reader programs available for purchase and download via the Internet.

While awareness of the Outlook PST file has grown, even many lawyers steeped in e-discovery fail to consider a user's Outlook .ost file. The OST or offline synchronization file is commonly encountered on laptops configured for Exchange Server environments. Designed to afford access to cached messages when the user has no active network connection., e.g., while on airplanes, local OST files often hold messages purged from the server—at least until re-synchronization. It's not unusual for an OST file to hold e-mail unavailable from any other comparably-accessible source.

By default, when a user opens an attachment to a message from within Outlook (as opposed to saving the attachment to disk and then opening it), Outlook stores a copy of the attachment in a "temporary" folder. But don't be misled by the word "temporary." In fact, the folder isn't going anywhere, and its contents—sometimes voluminous--tend to long outlast the messages that transported the attachments. Thus, litigants should be cautious about representing that Outlook e-mail is "gone" if the attachments are not.

The Outlook "viewed attachment folder" will have a varying name for every user and on every machine, but it will always begin with the letters "OLK" followed by several randomly generated numbers and uppercase letters (e.g., OLK943B, OLK7AE, OLK167, etc.).

Mirroring

Mirroring refers to the creation of an exact copy of a dataset. Mirroring may be used locally for data integrity and protection or across a network as a form of backup, duplicating the entire contents of a server to some distant, identical system. Disk mirroring, also called RAID 1, entails simultaneously writing identical data to two different hard drives, affording redundancy should either drive fail.

N

Nearline Storage

Nearline storage refers to voluminous data that, while not in such demand as to require instantaneous access via the network, must nonetheless be available from time-to-time without human intervention. Nearline data tends to be stored on high capacity media (like magnetic tape) that can be robotically loaded on demand, occasioning only a brief delay between a request and delivery of data.

NAS

Networked Attached Storage or NAS is a dedicated file server designed expressly for data storage. Because a NAS isn't called upon to do general computing tasks, it can employ a file system built exclusively for its limited role. When inquiring about devices, be careful not to reference only computers and servers, as a too-literal interpretation might allow someone to overlook a NAS.

Node

Anything connected to a network can be termed a "node;" however, anyone who uses the word node in this way must be termed a "nerd."

O

Offline Data

Offline data denotes ESI housed on media that is not connected to the network and requires human intervention, e.g., mounting or restoration, to access the contents. Backup tapes sent offsite for storage, legacy systems in the warehouse and even a CD-R in your desk drawer are examples.

The e-discovery challenge of offline data is that it must be proven not reasonably accessible to be excluded from search and production. Even then, producing parties must identify offline data with sufficient specificity to allow the requesting party to determine if the producing party is right about the data's inaccessibility. But there's the catch: how does a producing party do that without examining the contents?

To economically manage offline data, insure that its contents are indexed and the media clearly labeled *when the data goes offline* so as to obviate the costly and time-consuming need to bring it online, albeit briefly, to identify its contents. This isn't going to help with legacy data, but it's a no-brainer going forward.

P

Partition

A partition is a division of the storage area of a hard drive such that a single physical drive can be seen by the computer as multiple drives. If you think of an unpartitioned hard drive as a big metal cabinet, a partition is the division of that cabinet into file drawers. Though it's most common to encounter drives created with a single partition encompassing the entire storage area of the drive, in Windows, a hard drive can currently have up to four primary partitions or three ***primary partitions*** and one so-called ***extended partition*** that can be subdivided into as many as 24 extended partitions. Only one of the four partitions can be designated as an active partition, signaling the partition that holds the operating system the machine should boot on start up.

Partitioned hard drives can hold multiple operating systems such that a snippet of code called a ***boot loader*** can point the system to a partition other than the active partition to initiate a different operating system. Thus, a machine with a single drive can be configured to boot in Windows Vista, Linux or Windows XP via a start up menu. From the standpoint of e-discovery, a thorough search for ESI should include accounting for the full storage capacity of a hard disk, in case responsive data lurks on another partition. If you think this sounds farfetched, take a look at *Phoenix Four, Inc. v. Strategic Res. Corp.*[7]

Path

The complete local or network address to a particular folder, file or device, expressed hierarchically from a root location of a server or disk volume. If I were a file, the path to me might be expressed as ***Earth:\\North America\USA\Texas\Austin\78735\3723 Lost Creek Blvd\Lab\Craig Ball***. Traversing a path to a file is sometimes called "drilling down."

Peer-to-Peer Network

In a ***peer-to-peer*** or ***P2P*** network, each connected computer serves as both client and server for the purpose of sharing resources, but most often for sharing files (notably copyrighted music and video, as well as adult content and pirated software).

Peripheral

Just about any device you connect to a computer by cabling or networking (other than another computer or server) is called a peripheral. It most commonly refers to printers and scanners.

[7] No. 05 Civ. 4837, 2006 WL 1409413 (S.D.N.Y. May 23, 2006).

Protocol

An agreed-upon set of instructions, akin to a language, that permit devices to exchange information. Networks notably employ Ethernet or TCP/IP protocols to intelligibly transmit and receive data. As language can be thought of as a "protocol" for written or oral communications, a network protocol is a framework to sensibly interpret the ones and zeroes of digital communications.

R

RAID

A ***Redundant Array of Independent (or Inexpensive) Disks*** or ***RAID*** is a way of combining multiple hard drives to achieve greater performance, greater reliability or a mix of the two. The various types of RAID configurations are numbered. The three most commonly used configurations are RAID 0, RAID 1 and RAID 5.

A RAID 0 divides (or *stripes*, in storage parlance) data between two hard drives to combine the capacity into a single large volume and to increase the speed at which data is read and written. But because the data zigzags across two drives, a failure of either drive means the loss of all data.

A RAID 1 opts for complete redundancy, mirroring all contents between two drives such that a failure of either drive results in no loss of data--the trade off being that you can use only half of the combined capacity of the two drives and get no performance boost.

A RAID 5 uses three or more disks, garnering some of the speed boost seen in RAID 0 and the ability to fully recover all data should any one drive fail.

Because any one drive in a RAID 5 array can fail without data loss, RAID storage allows for the removal and replacement of drives from the array without the need to down the server. Thus, RAID storage—particularly RAID 5 configurations with more than 3 disks—are ubiquitous in mission critical servers. RAID 5 arrays are typically seen by the server as a single logical disk with a capacity of about two-thirds of the combined capacity of all disks in the array.

Despite its reliability, a RAID is not a substitute for a backup. A fire, flood or disgruntled employee won't destroy just one or two drives in the array, and all data will be unrecoverable absent a backup.

Root

Root refers to top level of a file system's directory structure, typically C:\ in a Windows system. In hacking, it also refers to a level of unrestricted access to a system, where "getting root" means taking unauthorized control of the system, often using hacker tools called ***root kits***.

Router

A router (sometimes called a ***switch***) is a device that directs the flow of the data packets by which information is transferred across a network. Unlike a hub, which merely relays all packets to all connections, a router actually assigns unique addresses to connections and steers packets to and from those addresses.

S

SaaS

Software as a Service or ***SaaS*** is software distribution mechanism where, instead of purchasing applications and installing them, programs are accessed on the Internet or downloaded on-the-fly as needed. The advantage of SaaS is that there is no need to purchase upgrades or install patches because the software's always up-to-date. The down side is that you do not own the software and must continue to pay for its use, as well as security concerns. In e-discovery, complications derive from the loss of physical dominion of the devices storing the data, as discussed previously under Cloud Computing. A notable example of SaaS is the Google Apps package of applications, which virtualizes a user's e-mail, contacts and calendar, along with document, spreadsheet and presentation authoring tools. The provider of SaaS is called an ***Application Service Provider*** or ***ASP***.

SAN

A ***Storage Area Network*** or ***SAN*** is a mass storage configuration that allows *network*-attached devices to be shared among servers at very high speeds yet appear as if they are *physically* attached to each server. SANs are tied to two important trends in networking: ***storage replication*** (where data is remotely mirrored for disaster recovery) and ***virtualization*** (where physical devices are subdivided into multiple virtual devices that appear to be distinct, physical machines like servers but actually exist as emulations using software). SANs allow large aggregations of physical storages devices to be logically re-allocated to various servers and tasks. Instead of adding a 120GB hard drive to a server, a 120GB "slice" of a multi-terabyte array can be assigned to appear and function as a physically-connected 120GB drive.

Server

A server is a device or application that delivers information to networked devices. When applied to hardware, server usually denotes a computer optimized and tasked to perform certain functions for other machines on the network. Servers tend to be isolated in locked and refrigerated server rooms, protected by backup systems and equipped with fail-safe or redundant components mounted in accessible racks, all to minimize downtime and increase security. Though a single server can perform a variety of tasks, businesses tend to dedicate servers to particular functions, such as storing user data, running applications like databases, delivering web content, managing printing, routing Internet traffic, handling e-mail stores, etc.

Share

Also called a ***Network Share***, see the discussion of shares in **Backup,** above.

Single Instance Storage

Networks and e-mail systems are replete with multiple iterations of identical documents. When an entire department receives an e-mail with the same attachment, or when thousands of employees keep a copy of the same memo, storage is wasted. Single instance storage performs ***de-duplication*** and replaces the individual copies with a *pointer* to an identical master copy. SIS aids backup by facilitating the use of fewer tapes and reducing the time required to complete the task. When dealing with a SIS volume in e-discovery, be careful to collect the de-duplicated document and not just its SIS pointer.

T

TCP/IP

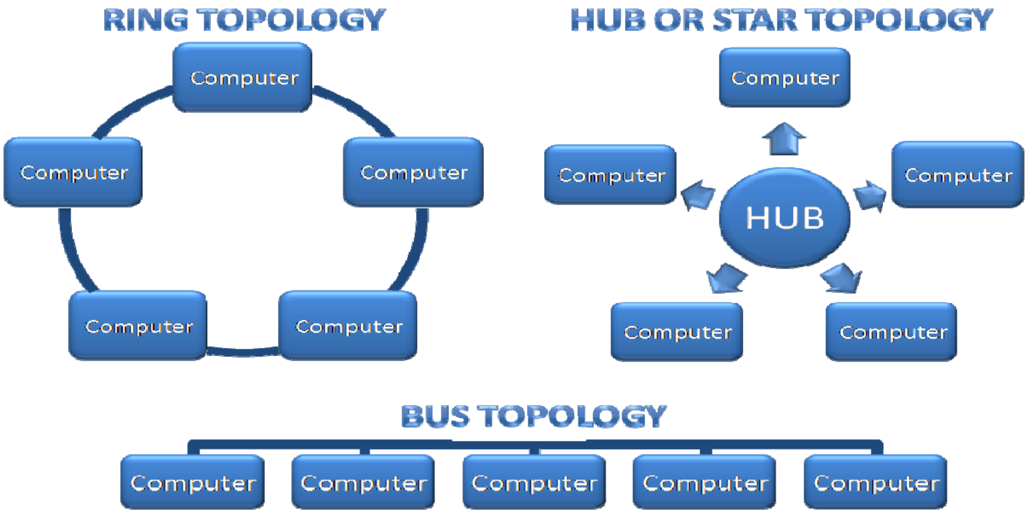
Transmission Control Protocol/Internet Protocol or ***TCP/IP*** is the universal computer-to-computer language of the Internet, but can also be implemented to support an intranet.

Thin Client

See *Client*

Topology

A geometric description of a network's structure based upon the way devices interconnect. Compare communication routes of the Ring, Hub or Star and Bus topologies depicted below.



V

Virtual Machine

Virtual machine or ***VM*** refers to the use of software to emulate or mimic the presence and function of hardware. Using VM software, a complete hardware and software computing environment, including operating systems, applications, data and emulated peripherals, can be stored in a single file. When that file is loaded to a VM player, it looks and works just like a real machine, but runs in a window, like any other piece of software.

Virtual machines have found enthusiastic acceptance in the IT world as a means to deploy, protect and backup virtualized servers, as well as a method to extract more value from hardware because one "real" machine can run many virtual machines without a notable drop in performance.

Because VMs can replicate almost any computing platform or environment, it promises to be a viable form of production for complex ESI. Virtualization enables opposing sides to enjoy comparable levels of functionality in native production even when one side lacks the hardware and software resources of the other. Not only does the evidence look the same for both sides, but it *works* the same way and can be easily shielded from inadvertent alteration and intentional manipulation.

Volume

A volume is a logical division of a hard drive that can hold a single operating system. Where a partition was akin to the physical drawer in a file cabinet, a volume speaks to the division of that drawer into compartments to hold file systems and files.

VPN

A ***Virtual Private Network*** or ***VPN*** is a private (i.e., secure) network that employs public pathways (i.e., the Internet). By employing authentication protocols and encryption of data as it traverses public pathways, the network traffic over a VPN is protected from interception and thus said to "tunnel" through public areas.

W

Workgroup

A workgroup is a subset of users in a local area network environment who are assigned privileges enabling them to collaborate by sharing files and peripherals. Microsoft Windows uses the term workgroup to identify the participants in a peer-to-peer network.