Electronic Discovery:
Glossary of 123 Commonly Used Terms

The following is a glossary of 123 commonly used terms to help you navigate the world of Electronic Discovery. If you have any questions about these terms or need assistance with a discovery issue, please contact Laura Marquez-Garrett or Darin Sands, co-chairs of Lane Powell's Electronic Discovery, Technology and Strategy Practice Group, at garrettl@lanepowell.com or sandsd@lanepowell.com.

Active Data: This term is often used to describe information currently displayed on a computer screen. The more technical usage refers to information stored on local storage media or a device that is visible to the operating system and/or application software with which it was created. Active data is accessible to users immediately and without modification or restoration.

Application (commonly used in place of "program" or "software"): A program or group of programs designed to enable end-users to manage computer resources and/or utilize end user programs.

Architecture: This can refer either to hardware or software, or a combination of both, that makes up a computer system or network. "Open architecture" refers to components that are easier to interconnect and operate, while "closed architecture" refers to those that are not.

Archival Data: Information that is maintained for long-term storage and record keeping purposes, but is not immediately accessible. Archived data can be stored in a number of ways. For example, it can be written onto removable media, like a DVD or backup tape, or maintained on a system hard drive.

Attachment: This term most commonly refers to a file attached to an email message. More generally, it refers to a file or record that is attached or associated with another, often for purposes of retention, transfer, processing, review, production and/or routine records management. Multiple attachments can be associated with a single file or record (referred to as the "parent" or "master" record).

Author (or **Originator**): The person or office that created or issued an item.

Backup Data: Active electronically stored information (ESI) copied onto a second medium (like a CD, DVD or backup tape) in its exact form, often intended as a source for recovery should the first medium fail. Usually, backup data is stored separately from active data, and differs from archival data (though may be a copy of archival data) in method and structure of its storage.

Backup-tape Recycling: The process of overwriting backup tapes with new data, typically on a fixed schedule (referred to as a "rotation"). Rotation schedules vary depending on the type and purpose of the backup tape. *Practice Note:* Once a duty to preserve arises, parties must suspend all routine deletion practices likely to result in the alteration or loss of potentially relevant evidence. This includes backup-tape overwriting, which commonly is overlooked.

Bates Number: Sequential numbering used to track documents, images or production sets (as with productions made in native format), which often includes a suffix or prefix to help identify the producing party, case name or similar information.

Cache (pronounced "cash"): A special high-speed storage mechanism that usually is utilized for frequently used data. Website contents, for example, often reside in cached storage locations on a hard drive.

Chain of Custody: Process of documenting and tracking possession, movement, handling and location of evidence. Chain of custody is tracked from the time evidence is obtained, until presentation in court or other submission. A clear chain of custody is important when issues of admissibility and authenticity arise, as it can establish that the evidence was not altered or tampered with in any way.

Claw-back Agreement (or **Quick Peek Agreement**): An agreement that protects against waiver of privilege and/or work production protections when inadvertant production occurs. *See* Federal Rule of Evidence 502. One distinction between this type of agreement and others, such as a traditional protective order, is that the return of inadvertently disclosed records is automatic and does not require a showing of reasonable steps to prevent disclosure.

Cloning: Cloning is a term generally used when referring to making a copy of the drive, as an example, to put into another machine without having to install everything from scratch. Another reason for cloning is mainly for backup purposes. Typically, cloning programs are not configured properly to get all areas of the drive. There is also a problem with later authentication, meaning there is no way to tell if anything was deleted or added to the clone after the day it was made. *See* Mirroring.

Cloud Computing: Internet-based computing wherein services — such as storage and applications — are delivered through the Internet, as opposed to using local servers or devices.

Coding: Process of examining and evaluating documents through the use of predetermined codes, and recording the results.

Compression: The reduction in the size of data to save storage space and reduce bandwidth necessary for access and transmission. "Lossless" compression preserves the integrity of the data (e.g., ZIP and RLE), while "lossy" compression does not (e.g., JPEG and MPEG).

Corrupted File: A corrupted file is one that has been damaged and cannot be read by a computer in part or in whole. Common causes include viruses, hardware or software failures, and degradation due to the passage of time.

CPU (**Central Processing Unit**) (aka "Microprocessor"): The primary silicone chip that runs the operating system and application software, controls essential operations and performs a computer's essential mathematical functions. The CPU is considered to be the "brain" of a computer.

Custodian: *See* Record Custodian.

Data: Generally refers to information stored on a computer, which can be created automatically (as with log files) or by users (as with information entered into a spreadsheet).

Data Extraction: Process of parsing data from electronic documents into separate fields, such as "Date Created," "Date Modified," "Author," etc. In a database, this allows for searches across data or by sorting respective fields.

Data Filtering: Use of specified parameters to identify specific data.

Data Mapping: Method used to capture information relating to how ESI is stored, both virtually and physically. A basic data map will include name and location information, while a more complex data map may include several, if not all, of the following: software and formatting information; description of backup procedures in place; interconnectivity and utilization of each type of ESI within the organization; accessibility, policies and protocols for retention and management; and record custodian information.

Data Set: Named or defined collection of data.

Database (**DB**): Refers to a collection of information, organized so that specific data or groups of data can be identified and searched quickly.

Decompression: To expand or restore compressed data to its original size and format.

Deduplication ("**Deduping**"): The process of comparing electronic records and removing or marking duplicates within a data set. Deduplication can be done across custodians or across the corpus of the data. *Practice Note:* Before any deduping is performed, an agreement as to methodology and what constitutes a duplicate record should be reached. Failure to obtain agreement can result in discovery disputes and unnecessary costs.

Defragment ("**Defrag**"): Defragmentation refers to the use of a computer utility to reorganize files in a contiguous manner.

Fragmentation occurs naturally when a hard drive or other storage medium frequently is used and will result in the storage of a file in noncontiguous clusters. The more places that need to be searched, the slower the data will be accessed. *Practice Note:* Defragmentation may be set up to run automatically and with little or no oversight by users, and will result in the overwriting of information residing in unallocated space. In some cases, preservation may require suspension of defragmentation and imaging to preserve such information.

Deleted Data: Refers to live data that has been deleted by a computer system or user activity. "Soft deletion" refers to data marked for deletion that may no longer be accessible to the user (such as the emptying of one's "recycle bin"), but has not yet been physically removed or overwritten. Soft deleted data may be recoverable. Further, deleted data in general may remain on storage media, in whole or in part, until it is overwritten or "wiped" and, even after being wiped, it may be possible to recover information relating to the deleted data.

Deletion: The process of removing or erasing data from active files and other data storage structures, although some or all of it may be recoverable with special data recovery tools.

Directory: An organizational unit or container used to organize folders and files into a hierarchical or tree-like structure. Some user interfaces use the term "folder" instead.

Discovery: General term used to describe the process of identifying, locating, obtaining, reviewing, evaluating, and/or producing information and other evidence for use in the legal process.

Diskwipe: A utility used for the purpose of overwriting or erasing existing data.

Document (or **Document Family**): Pages or files produced either in hard copy or through a software application, which constitute a logical single communication of information. For example, fax cover letter, faxed letter and attachments. For document review purposes, the cover letter often is referred to as the "parent," and the letter and attachments as the "child."

Document Type (or **Doc Type**): A typical field used in coding, examples include "correspondence," "memo," "agreement," etc.

Download: The process of moving data from another location to one's own, typically over a network or the Internet.

Duty to Preserve: Duty arising under state and federal law, upon reasonable anticipation of litigation, to preserve documents, electronic records and data, and any other evidence or information potentially relevant to a dispute. The duty also arises in the context of audits, government investigations and similar matters. The scope of the duty and what is required under a specific set of circumstances is determined by considerations of reasonableness and proportionality. *Practice Tip:* While the duty to preserve unquestionably arises upon the filing of a formal complaint, it also often arises sooner, such as when an investigation takes place or after receipt of a credible complaint or demand letter.

Early Case Assessment (**ECA**): ECA can refer to any number of various tools used to analyze data sets. As opposed to a linear review, effective ECA utilizes search terms, filters and/or additional criteria to cull the data set and enable a more efficient and focused review.

Electronic Discovery (**E-Discovery**): General term used to describe the process of identifying, preserving, collecting, preparing, reviewing and producing ESI, as opposed to hard copy documents and other records.

Electronic Document Management: This refers to the process utilized in the management of documents, whether hard copy or electronic. In the case of hard copy documents, it includes those steps necessary to make them available electronically, such as images, archiving, etc.

Electronically Stored Information (**ESI**): Information that is stored electronically (regardless of the media or original formatting) as opposed to paper.

Email (**Electronic Mail**): Digital messages from an author to one or more recipients. Email operates across the Internet or computer networks.

Email Archiving: The process of preserving and storing email.

Embedded Metadata: Metadata embedded with content. *See* Metadata.

Encryption: A procedure that makes the contents of a file or message scrambled and unintelligible to anyone not authorized to read it.

Expanded Data: *See* Decompression.

File: A collection of data or information stored under a specific name, called a filename.

File Extension: A suffix to the name of a file, separated by a dot. Often an abbreviated version of the name of the program in which the file was created or saved, the suffix indicates the program that may be used to open the file.

File Server: Refers to a computer attached to a network, of which the main purpose is to provide a centralized location for shared storage of computer files (such as documents and images) that can be accessed by the workstations attached to the same network. File servers are the heart of any server network. They can contain data for other programs or direct access to documents themselves.

Filename: A unique identifier assigned to a specific file. Filenames can be descriptive (e.g., LtrToJohn) or cryptic (e.g., 5787720) and are followed by an extension.

Firewall: A system designed to prevent unauthorized access to a specific computer, server or private network.

Flash Drive: A small, data storage device used to store files or transport them from one computer to another, also commonly referred to as a USB or thumb drive.

Forensics: The application of scientifically proven methods to retrieve, examine and/or analyze data in a way that can be used as evidence.

Form of Production: The manner of producing documents and data, including file format (native format vs. TIFF or PDF) and method of production (electronic vs. paper).

Format (noun): The internal structure of a file that defines the way it is stored and the programs in which it can be used.

Format (verb): The act of preparing a storage medium ready for first use.

FTP (**File Transfer Protocol**): The protocol for transferring files over a network or the Internet.

Full-text Search: When a data file can be searched for specific words and/or numbers.

Fuzzy Search: Searches that use approximate, rather than exact, matches.

Hard Drive: Self-contained storage device, generally with a high capacity, that has a read-write mechanism and one or more hard disks.

Hash: A hash value (or hash) is an alpha-numeric string that is generated by an algorithm and uniquely identifies original data. It is useful to authenticate data (such as a file) for evidence admissibility in court, for determining duplicate documents and for identifying alterations to documents. Common hashes are MD5 or SHA. An example of a hash value: d41d8cd98f00b204e9800998ecf8427e.

Hash Coding: Method of coding that provides quick access to data items capable of being distinguished through use of a key term, like the name of a person. Each data item to be stored is associated with the key term, the hash function is applied to that term, and the resulting hash value may then be used as an index that permits users to select one of several "hash buckets" in a hash table. The table contains pointers to the original item.

Hidden Files or Data: Files or data not readily accessible or visible. For example, in the case of many operating systems, critical files are "hidden" to prevent inexperienced users from accidentally deleting or altering them.

Hyperlink: An element in an electronic document (usually appearing as an underlined word or image) that links to another place in the same document or to an entirely different document when clicked.

Image: Refers to an exact replica. Image may refer to a type of document, such as a .tif or .jpeg. To image a hard drive means to make an identical copy. Forensic imaging is a bit for bit copy and can be made at a logical or physical level, meaning a user just copies the C drive or the D drive or the unallocated space (which is where deleted data resides). The main advantage to forensic imaging is the checksums and verification by digital fingerprint contained in the image format, which show that the image has not been altered in any way since the day it was made. If the image has been altered, the CRC values (checksums) and the digital fingerprint (such as the MD5 hash) will change and not match, and the image will not verify.

Image Processing: Capturing an image, usually from data in its native format, so it can be entered into another computer system for processing and, often, manipulation.

Import: To bring information or data to one environment or application from another.

Inactive Record: Records that no longer are routinely referenced but must be retained, usually for audit or reporting purposes.

Index: In the context of electronic discovery, index refers to database fields used to categorize, organize and identify each document or record.

IRT (**Intelligent Review Technology**): *See* Predictive Coding.

IS/IT (**Information Systems or Information Technology**): Commonly refers to those individuals responsible for computers and computer systems.

Keywords: Words designated as being important for search purposes.

Litigation Hold (or **Legal Hold**): Communications issued upon notice of a duty to preserve, and instructing individuals and entities in the efforts required to ensure adequate preservation of potentially relevant evidence. *See* Duty to Preserve, and Preservation.

Media or Medium: An object on which data is stored. Examples include disks, backup tapes, servers and hard drives.

Metadata: Data (typically stored electronically) describing the characteristics of specific ESI that can be found in different places and different forms. *Practice Note:* The obligation to preserve includes metadata. Additionally, in cases with more than minimal discovery, parties should discuss metadata at the outset as it may determine the format in which documents must be produced.

Mirroring: Duplicating data or a disk in a manner that results in an exact copy. This is often done for backup purposes.

Native Format: Refers to an electronic document's original form, as defined by the application that was used to create the document. Sometimes documents are converted from their native format to an imaged format, such as TIFF or PDF. Once converted, the original metadata cannot be viewed. *Practice Tip:* Parties should discuss issues of production and formatting early in the litigation to avoid later disputes and the risk that production efforts will have to be redone.

Natural Language Search: A manner of searching that does not require formulas or special connectors (e.g. ,"origin" and "basketball"), but can be performed by using plain statements (e.g., "What is the origin of basketball?").

Near Duplicates: The process of identifying and culling documents that are nearly duplicate. Deduplication software can group near duplicate documents by percentage of similarity, so reviewers can quickly review and code documents for responsiveness or privilege. *See* Deduplication.

Network: Two or more computer systems that are linked together.

OCR (**Optical Character Recognition**): A process that reads text from paper, translating and converting the images so they can be manipulated by a computer. In the context of electronic review, a useful application of OCR is the ability to then search the text.

Off-line Storage: Storage of electronic records on a removable disc or other device for disaster-recovery purposes.

Operating System (**OS**): Provides the software platform that directs the overall activity of a computer, network or system. Common examples include UNIX, DOS, Windows, LINUX and Macintosh. The operating system is the foundation on which applications are built.

Overwrite: To copy or record new data over existing data, as with backup-tape recycling or when updating a file or directory. *Practice Note:* Overwritten data cannot be retrieved, making it important to suspend policies and procedures likely to result in the overwriting of potentially relevant data, once on notice of the duty to preserve.

PDA (**Personal Digital Assistant**): Handheld devices with computing, Internet, phone/fax and similar capabilities.

Peripheral: Refers to a device that attaches to a computer, such as a printer, modem or disk drive.

Predictive Coding: A method of culling relevant documents for production or review. Predictive coding uses algorithms to determine the relevance of documents based on linguistic and other properties and characteristics. It relies on the coding from a human sampling of documents called a "seed set." The seed set allows the computer to identify and evaluate the remaining documents. Also referred to as IRT or TAR.

Preservation: The process of managing, identifying, and retaining documents and other data for legal purposes. *Practice Note*: Reasonable efforts to preserve include the suspension of routine deletion policies, issuance of adequate preservation instructions and oversight as appropriate. Delegation is not a defense when evidence is lost, altered and/or destroyed after a parties' duty to preserve arose.

Private Network: A network that is connected to the Internet, but limits access to only those persons operating within the private network.

Privilege Data Set: Documents withheld from production despite being relevant and/or responsive on the grounds of legal privilege. Parties are generally required to produce a privilege log, identifying enough information about each document so that the opposing party can determine whether or not to challenge the withholding (e.g., senders and recipients, creation date, general description of subject matter and privileges asserted).

Production: The process of producing or making available for another party's review the documents and/or other ESI deemed responsive to one or more discovery requests.

Program: *See* Application and Software (synonymous with software).

Protocol: A common format for transmitting data between two devices. TCO/IP is one of the most common protocols for networks.

Quality Control (**QC**): Efforts undertaken to ensure the quality of a product or task.

Record Custodian: Individual responsible for the physical storage and protection of records. Custodian may also refer to various individuals with knowledge and/or possession of, or who created, sent, received and/or stored emails, documents and other data relevant to an ongoing or potential dispute.

Records Manager: The person responsible for implementation of a records management program.

Records Retention Period (or **Retention Period**): The length of time a given set or series of records must be maintained. The retention period is often expressed as a period of time (such as six years), an event or action (such as completion of an audit), or both (six years after completion of an audit).

Redaction: The intentional concealing of a portion of a document or image, done for the purpose of preventing its disclosure. *Practice Note:* Redactions and their basis should be clearly indicated and disclosed to avoid an appearance of bad faith.

Restore: Transferring data from a backup medium to an active system. Data is often restored for the purpose of recovering the data after a problem, failure or disaster, or where the data is relevant and has not been preserved or cannot be accessed elsewhere.

Review: Process used to read or otherwise analyze documents in order to determine content, relevance or applicability of some other objective or subjective standard.

Rewritable Technology: Storage devices that permit data to be written more than once, such as hard drives and floppy disks.

Sampling: Usually refers to any process of which a large collection of ESI or a database is tested to determine the existence and/or frequency of specific data or types of information.

Seed Set: The initial set of data/documents used in predictive coding. The seed set is "trained" by learning algorithms to cull data down to a potentially relevant set for reviewers to analyze for production or privilege. *See* Predictive Coding. *Practice Note:* Cooperation between counsel as to methodology and selection of the seed set(s) can avoid challenges down the road as to whether sufficient efforts were made to cull the data for relevant documents, as well as allegations that relevant evidence was withheld.

Server: A computer or device on a network that manages network resources. There are several different types of servers. Typically, a server will be dedicated, which means that they perform no task other than their server tasks (i.e., a file server stores files, a print server manages one or more printers, etc.).

Smart Card: A credit card size device that contains a battery, memory and microprocessor.

Social Network: A group of people who utilize social media, typically based on a specific theme or interest. Facebook is an example of a popular social network.

Social Media: Internet applications that permit users to publicly and interactively share and communicate information, whether of a general or personal nature.

Software: Programs used to direct the operations of a computer, which includes operating systems and software applications.

Spoliation: The destruction of evidence and information that may be relevant to ongoing or reasonably anticipated litigation, government audit or investigation. Courts differ as to the requisite level of intent required for imposition of sanctions, with fault (possession and failure to preserve) on one end and willfulness on the other.

SQL (Structured Query Language): A database computer language used to manage data.

Stand-alone Computer: A computer not connected to a network or other computers, except possibly through use of a modem.

Storage Device: This usually refers to mass storage devices, such as disks and tape drives, but can be used to describe any device capable of storing ESI.

System Administrator: The person in charge of keeping a network working.

System Files: Nonuser created files that permit computer systems to run.

TAR (Technology Assisted Review): *See* Predictive Coding.

Temporary (Temp) File: Files created by applications and stored temporarily on a computer. Temp files enable increased processor speeds. In the case of temporary Internet files, for example, a browser stores website data so that the next time the same website is accessed it can be loaded directly from the temporary Internet file. Stored data may also be viewed even in the absence of an Internet connection.

Thread: A series of related communications, usually on a particular topic.

TIFF (Tagged Image File Format): A widely supported and utilized graphic file format. TIFFs, used to store bit-map images, originated in the early 1980s. TIFFs are a static format, which means the data cannot be altered, as opposed to native documents that can be altered. Native documents might be converted to TIFF, for example, resulting in the loss of metadata but permitting commonly used functions such as Bates labeling and redaction.

Unallocated Space: The area of computer media, such as a hard drive, that does not contain normally accessible data. Unallocated space frequently results from deletion, wherein data resides but is not generally accessible, until being overwritten, wiped or retrieved through utilization of forensic techniques.

USB (Universal Serial Bus) Port: A socket on a computer or other device into which a USB cable or device can be inserted.

VPN (Virtual Private Network): Secure networks that utilize mechanisms, such as encryption, to ensure access by authorized users only and prevent data interception.

Zip Drive: A specific kind of removable disk storage device.

ZIP: Common file formatting allowing fast and simple storage for the purposes of archiving or transmitting.

For more information, please contact the co-chairs for Lane Powell's Electronic Discovery, Technology and Strategy Practice Group: Laura Marquez-Garrett (garrettl@lanepowell.com) or Darin M. Sands (sandsd@lanepowell.com).