eDiscovery Terminology Glossary (5/10/13)

Compiled by Robert D. Brownstone of Fenwick & West LLP

For other Online Glossaries, click HERE

Active Data: Data currently displayed on a computer screen, and/or files on a computer that can be accessed without having to use a restoration process. [Richard A. Lazar, The Guide to Electronic Discovery, at 37 (Fios, Inc. 2002) ("Fios Guide")]**.** The information readily available and accessible to users, including word processing files, spreadsheets, databases' data, e-mail messages, electronic calendars and contact managers**.** [**Feldman, "*The Essentials of Computer Discovery*," Computer Forensics Inc. (1/1/01)**] On a PC a list is viewable through Windows Explorer [*Id.*]

Archive [Data]: 1. Any kind of information maintained for historical reference [**Whatis?com's Backup and recovery glossary**], including, for example, a copy of data on a computer drive, or on a portion of a drive. [**Fios' eDiscovery Glossary**] 2. A backed up set or subset of an organization's enterprise-wide data [**Webopedia**] 3**.** The practice of copying of some or all of an organization's data into a storage format [**Whatis?**] e – in the old days, mostly on tape, but now often on disk [**Enterprise Networking Planet**] 4. Data copied onto backup media, from which it can be restored, most typically after a disaster [**ComputerLanguage**]; 5. Live data in "near-line storage" available for a relatively easy connection to an organization's network. [**SearchStorage**] 6. E-mails stored somewhere other than in people's active e-mailboxes [**SearchStorage**]. 7. E-mails stored in an automated, company-wide repository of employees' older e-mails [**SearchStorage**]. 8. E-mail messages that an individual manually copies *ad hoc* to a personal storage file, known as a ".pst" file in the Microsoft Outlook environment [**Microsoft Office Online**]

Backup: To create a copy of data as a precaution against the loss or damage of the original data. Most users backup some of their files, and many computer networks utilize automatic backup software to make regular copies of some or all of the data on the network. Some backup systems use tape as a storage medium. [**Kroll Ontrack's Glossary**]

Backup Tape: Disaster Recovery Tape is a portable medium used to store data that is not presently in use by an organization to free up space but still allow for disaster recovery. [**Kroll Ontrack's Glossary**]

Back-up Tape Recycling: Backup Tape Recycling describes the process whereby an organization's backup tapes are overwritten with new backup data, usually on a fixed schedule (*e.g.*, the use of nightly backup tapes for each day of the week with the daily backup tape for a particular day being overwritten on the same day the following week; weekly and monthly backups being stored offsite for a specified period of time before being placed back in the rotation, etc.). When a litigation hold is triggered, an attorney should assess the need to stop the potential overwriting of backup data due to recycling. [**Applied Discovery's "Bone up on Backup"**]

Bit-by-bit copy or capture: Bit stream backup (also referred to as mirror image backup) involves the backup of all areas of a computer hard disk drive or another type of storage media. Such a backup exactly replicates all sectors on a given storage device. Thus, all files and **ambient data storage areas** are copied. Bit stream backups -- sometimes also referred to as "evidence grade" backups -- differ substantially from traditional computer file backups and network server backups. [**NTI's Computer Forensics Definitions**]

Blowback or "Blow back": Printing electronic files to paper for review or production in hardcopy form [NOT recommended by FWPS]. [**Albert Barsocchini, "Data Collection Standards" (LTN 1/15/04)**]**.** The to-be-printed electronic files may have previously been scanned from paper into electronic form . . . and/or originated in native electronic form. In either event, somewhere along the way the files may have been converted into .tif (or .pdf) and/or endorsed with Bates numbers, privilege stamps, confidentiality redaction overlays, etc. [**RBrownstone**]

Boolean Search: The term "Boolean"' refers to a system of logical thought developed by an early computer pioneer, George Boole. In Boolean searching, an "and" operator between two words or symbols results in a search for documents containing both of the words or symbols. An "or" operator between two words or symbols creates a search for documents containing either of the target words/symbols. A "not" operator between two words or symbols creates a search result containing the first word/symbol but excluding the second. [**Applied Discovery's Glossary**]

Chain of Custody: All information on a file's travels from its original creation version to its final production version**.** A detailed account of the location of each document/file from the beginning of a project until the end. A sound chain of custody verifies that you have not altered information either in the copying process or during analysis. If you cannot show the chain of custody, you may have a difficult time disproving that outside influences might have tampered with the data**.** A chain of custody failure — *i.e.*, the mishandling of electronic evidence (even fully recovered files) — can cause a litigation defeat. [**Top Ten Things To Do When Collecting Electronic Evidence," Computer Forensics Inc. (11/10/03)**; **Fenwick & West**]

Custodian De-Duplication: Culls a document if multiple copies of that document reside within the same custodian's data set. For example, if Mr. A and Mr. B each have a copy of a specific document, and Mr. C has two copies, the system will maintain one copy each for Mr. A, Mr. B, and Mr. C. Contrast with **case de-duplication** and **production de-duplication**. [**Kroll Ontrack's Glossary**]

De-Duplication: "De-Duping"' is the process of comparing electronic records based on their characteristics and removing duplicate records from the data set. [**Kroll Ontrack's Glossary**]

DLT Tape: Digital linear tape is a form of magnetic tape and drive system used for computer data storage and archiving. DLT is one of several technologies developed in recent years to increase the data-transfer rates and storage capacities of computer tape drives. [**SearchStorage.com**]

Electronic Evidence: *Any* computer-generated data that is relevant to a case. Included are email, text documents, spreadsheets, images, database files, deleted email and files and back-ups. The data may be on floppy disk, zip disk, hard drive, tape, CD or DVD. [**Norcross Group FAQ's**]

E-mail: A simple text message – a piece of text sent to a recipient. In the beginning and even today, e-mail messages tend to be short pieces of text, although the ability to add attachments now makes many e-mail messages quite long. Even with attachments, however, e-mail messages continue to be text messages. [**How Stuff Works**]

External drive: *See* **Portable Drive**.

Fingerprinting or **Hash** or **MD5 Hash:** *See* **MD5 Hash**.

File Transfer Protocol: *See* **FTP**.

Firewire drive: *See* **Portable Drive**.

FTP File Transfer Protocol, the protocol for exchanging files over the Internet. [**Webopedia**] FTP works in the same way as HTTP for transferring Web pages from a server to a user's browser and SMTP for transferring electronic mail across the Internet -- in that, like these technologies, FTP uses the Internet's TCP/IP protocols to enable data transfer. [**Webopedia**]

Hash or **MD5 Hash** or **Fingerprinting:** *See* **MD5 Hash.**

Inaccessible Data ("Not Reasonably Accessible Data"): In contrast with **active data**, this data has to undergo a **restoration** process to be displayed on a computer screen. Two prevalent subsets are "**backup tapes**" and "erased, fragmented or damaged data." According to **Fed. R. Civ. P. 26** as interpreted by prevailing federal case law, this category of electronically stored information ("ESI") is the one as to which a court can consider cost-shifting. [***Zubulake I* (S.D.N.Y. 5/12/03)**]

Index: (n.) In database design, a list of keywords, each of which identifies a unique record. An index makes it faster to find specific records and to sort records by the field used to identify each of them. (v.) To create an index for a database, or to find records using an index. [**Webopedia.com**]

Load File: A data file that sets out links between the records in a database and the document image files to which each record pertains. This is a critical deliverable of any scanning and coding job. Without a correctly structured load file, documents will not properly link to their respective database records. [**Commonwealth Legal's "Litigation Support Glossary**"]

MD5 Hash or **Hash** or **Fingerprinting:** MD5 is an **algorithm** that is used to verify **data integrity** through the generation of a unique 128-bit digital fingerprint of a file of any length (even of an entire hard disk drive). [**SearchSecurity.com**] There are 10^{38} **(a/k/a 10 followed by thirty seven more zeros)** different possible hash values. [**NTI**] Thus, it is highly unlikely that any two files would have the same hash value. Furthermore, it is "computationally infeasible" at this time to manufacture a file that generates a particular hash value. Thus, one can readily identify a known file through its MD5 hash value.

Metadata: Data that describes how, when and by whom a particular set of data was created, edited, formatted, and processed. Access to metadata provides important evidence, such as blind copy (bcc) recipients, the date a file or e-mail message was created and/or modified, and other similar information. Such information is lost when an electronic document is converted to paper form for production. [**Applied Discovery's Glossary**]

Multi-Page TIFF: A **.tif file** comprised of all of the pages contained in the underlying electronic file or hardcopy document prior to its conversion to or scanning into .tif format. As distinguished from the situation where each page of an underlying multi-page document becomes a separate .tif file. [**RBrownstone**]

Native File/Format: A file saved in the format of the original application used to create it. Dealing with native files can minimize expensive per-page costs for the traditional TIFF processing and will maximize the relevant information available from the file. [**RenewData's Glossary**]

OCR: Optical Character Recognition is the conversion of a scanned document into searchable text and the rendering of its text susceptible to copying for pasting into a new file. Following the scanning of a given document, OCR software evaluates the scanned data for shapes it recognizes as letters or numerals. OCR technology relies upon the quality of the printed copy and the conversion accuracy of the software. Generally acknowledged to be only 80-85 percent accurate. [**RBrownstone**]; [**Applied Discovery's Glossary**]

Portable drive: An external disk drive that is plugged into a port on a computer, typically a USB or FireWire port. Typically used for backup, but also as secondary storage. Such units rival internal drives in capacity. [**TechWeb TechEncyclopedia**]

Predictive Coding: An industry-specific term generally used to describe a **Technology-Assisted Review** process involving the use of a Machine Learning **Algorithm** to distinguish **Relevant** from **Non-Relevant** documents, based on a **Subject Matter Expert**'s Coding of a **Training Set** of Documents. *See* **Supervised Learning** and **Active Learning**. [**Grossman-Cormack Glossary of Technology-Assisted Review**]. *See also* **Technology-Assisted-Review (TAR)** below.

PST: In Microsoft Outlook, the Personal Folders file (.pst) is a data file that stores all of a user's messages and other items on his/her computer. An Outlook user can create one or more .pst's to organize and back up items for safekeeping. Even when an e-mail system is being run on a Microsoft Exchange Server, Outlook data can be backed up to a .pst file stored either locally on a hard drive or on a network drive -- rather than on the e-mail server. Each .pst file contains all of one's Outlook folders, including the Inbox, Calendar, and Contacts. [**Microsoft Office Online**]

Relational Database: A collection of data items organized as a set of formally-described tables from which data can be accessed or reassembled in many different ways without having to reorganize the database tables. Invented by E. F. Codd at IBM in 1970. [**SearchDatabase.com**]

Relatively Inaccessible Data: *See* **Inaccessible Data ("Not Reasonably Accessible Data")**

Restore or Restoration: In data management, a process that involves copying backup files from secondary storage (tape or other backup media) to hard disk to return data to its original condition if files have become damaged, or to copy or move data to a new location. [**Whatis.com**]

Searchable TIFF*:* An imaged file accompanied, in a database, by OCR'd text that is searchable. A misnomer/oxymoron/fiction, *i.e.*, a medium created for exchange of electronic files in litigation. [**RBrownstone, NC JOLT at 44-46 (citing Nimsger, et al.)**]

Spoliation: Spoliation is the destruction or alteration of evidence during on-going litigation or during an investigation or when either might occur sometime in the future. Failure to preserve data that may become evidence is also spoliation. [**Norcross Group FAQ's**]

Technology-Assisted-Review (TAR): **Prioritizing** or **Coding** a **Collection** of **Document**s using a computerized system that harnesses human judgments of one or more **Subject Matter Expert(s)** on a smaller set of documents and then extrapolates those judgments to the remaining **Document Collection**. Some TAR methods use Machine Learning **Algorithm**s to distinguish **Relevant** from **Non-Relevant** documents, based on **Training Example**s **Code**d as Relevant or Non-Relevant by the Subject Matter Experts(s), while other TAR methods derive systematic **Rule**s that emulate the expert(s)' decision-making process. TAR processes generally incorporate **Statistical Model**s and/or **Sampling** techniques to guide the process and to measure overall system effectiveness. [**Grossman-Cormack Glossary**]. *See also* **Predictive Coding** above.

TIFF: Tagged Image File Format is a graphic file format used for storing still-image bitmaps. TIFFs are stored in tagged fields, and programs use the tags to accept or ignore fields, depending on the application. [**Fios' eDiscovery Glossary**]