An E-Discovery Primer:

Thoughts on the Technical, Practical, and Proportional

North Carolina
Superior Court
Judges Conference
(2017)

Allison Mullins
Juan Sosa
Turning Point Litigation
Mullins Duncan Harrell &
Russell PLLC
Greensboro, NC

Introduction

E-discovery disputes occupy unique real estate in the legal landscape. They exist at the intersection of the physical and the virtual. Primarily, there are two disconnects that fuel the confusion surrounding such disputes. The first is the relative sophistication of adverse parties and their counsel. The second is technical jargon. This manuscript aims to address the pain points associated with decision-making in e-discovery disputes by (1) discussing the role that relative sophistication of adverse parties and their counsel should play in a judge's decision-making process and (2) assigning practical context to the overly-technical jargon typically associated with e-discovery.

Relative Sophistication and the E-discovery Maturity Model

When it comes to e-discovery issues, individuals and entities necessarily have varying levels of sophistication, largely based on prior experience or the lack thereof. But much like the eggshell plaintiff in a tort case, litigation finds these parties as they are and e-discovery expectations must be sufficiently flexible to account for varying levels of proficiency. If the requirements of discovery generally and e-discovery specifically are established in a "one-size-fits-all" method, there is a significant risk for e-discovery to become a significant barrier to justice—requiring either too much or too little. Both the North Carolina Business Court Rules in Rule 10.3(a)[1] and the Federal Rules of Civil Procedure in Rule 26(b)(1)[2] and Rule 26(b)(2)(B)[3] recognize the need for the e-discovery solution to fit the individual case. The concept is commonly referred to as "proportionality." As recognized by these rules, it is not only the size of the case or the importance of the issues that feeds

[1] "Counsel should discuss the scope of discovery, taking into account the needs of the case, the amount in controversy, limitations on the parties' resources, the burden and expense of the expected discovery compared with its likely benefit, the importance of the issues at stake in the litigation, and the importance of the discovery for the adjudication of the merits of the case." BCR 10.3(a).

[2] "Unless otherwise limited by court order, the scope of discovery regarding any nonprivileged matter that is relevant to any party's claim or defense and proportional to the needs of the case, considering the importance of the issues at stake in the action, the amount in controversy, the parties' relative access to relevant information, the parties' resources, the importance of the discovery in resolving the issues, and whether the burden or expense of the proposed discovery outweighs its likely benefit. Information within this scope of discovery need not be admissible in evidence to be discoverable." Fed. R. Civ. P. 26(b)(1).

[3] "A party need not provide discovery of electronically stored information from sources that the party identifies as not reasonably accessible because of undue burden or cost…." Fed. R. Civ. P. 26(b)(2)(B).

into the proportionality analysis, it is also the positions and sophistication of the parties.

The competence aspect of institutional e-discovery knowledge is succinctly reflected in the "e-discovery maturity model." *See* Adam Hurwitz, *The E-Discovery Maturity Model*, (2010) available *at* http://www.edrm.net/papers/the-e-discovery-maturity-model/. This model assigns maturity levels (1 through 5) to parties involved in civil litigation. Parties at maturity level 1 possess little, if any, expertise of their own. Companies at maturity level 1 rely exclusively on outside counsel and vendors to "just get it done." They fumble through a chaotic process. They are not proactive or forward thinking regarding e-discovery issues because their involvement in civil litigation is sporadic. They only address e-discovery "problems" as they arise. In short, Level 1 parties treat e-discovery like a game of whack-a-mole.

On the other end of the spectrum, Level 5 parties possess fully automated processes, and their e-discovery workflows are fully mature. Level 5 parties devote substantial resources to automating their e-discovery processes because they are almost always involved in, or under the threat of, civil litigation. Their e-discovery expenses are so substantial that the automation of internal e-discovery processes presents a significant opportunity for cost-reduction. Some Level 5 companies include their e-discovery expenses in their annual information technology budgets. It is easy to see how Level 1 and Level 5 organizations would have conflicting views on how the e-discovery process should unfold if they were to meet as opponents in a given case.

It is important to note that, although the parties' relative maturity levels can inform the court of which e-discovery mechanisms may be reasonable to impose on a given party in a particular action, a party should not be penalized for its lack of maturity. Most often, Level 1 parties and their counsel have acquired that label solely due to a lack of resources or experience with e-discovery. Fortune 500 companies almost exclusively reside at Level 5 because they have the most resources and the most experience conducting civil litigation. Individuals and small to medium sized businesses inevitably cluster around Level 1 and Level 2 for similar reasons. They are only sporadically involved in civil litigation. As a result, they have no reason to include potential e-discovery expenses in their annual budgets. They simply do not have the same resources or capacity to conduct e-discovery as thoroughly, efficiently, and expansively as a Level 5 company. In fact, e-discovery expenses could financially cripple, or even sink, a Level 1 company if forced to conduct e-discovery like a Level 5 company.

Another noteworthy point is that parties and their counsel often reside on different levels of the e-discovery maturity spectrum. As the cliché goes, a chain is

only as strong as its weakest link. Accordingly, the disparate e-discovery maturity levels between clients and their counsel can be just as problematic as disparate maturity levels between adverse parties.

In general, the maturity level of a party and its chosen counsel is a good proxy to inform the court of the burdens a particular party will likely encounter when responding to an adverse party's requests. In other words, understanding the parties' relative e-discovery maturity levels is a factor, in addition to other considerations such as the nature of the case, that may significantly aide the court in gauging the reasonable expectations to be assigned to the scope of discovery.

Hypothetical – How Understanding Relative E-discovery Maturity Can Aid Decision-Making

Several parties (all individuals) are involved in a caveat proceeding in which the decedent and his widow did not use e-mail very often but whose children used e-mail extensively. The plaintiff is one of six children and the only child who was left out of the decedent's will. The decedent died unexpectedly, leaving significant funds to his widow and $100,000 to each of the decedent's children except the plaintiff. The will was amended to exclude the plaintiff one month before death. The estate is defending the interests of the widow and the five children who took under the will ("the heirs") against the plaintiff's challenge.

In hopes of proving that the heirs colluded to have the allegedly incompetent decedent remove the plaintiff from the will, the plaintiff's attorneys have requested that the defendants review all of the heirs' and decedent's e-mails for the two years leading up to the decedent's death. They also have requested that the estate produce all e-mails that refer to the decedent or his property in the load file format used by their e-discovery software so that all of the metadata associated with the files is preserved. The plaintiff's request would require the estate's attorneys to review 100,000 e-mails and 4,000 attachments, comprising 30 gigabytes of data. The estate estimates that it would take around 300 hours to review all of the data requested by the plaintiff. The estate claims that the heirs and the decedent exchanged a total of 10,000 e-mails during the two year time period, resulting in 3 gigabytes of data, which would take approximately 30 hours to review. Because the plaintiff has only spoken to his family members ten times over the past two years, his counsel will devote only a few hours collecting, reviewing, and producing the plaintiff's documents.

The estate has objected to the plaintiff's request for production as overly broad and unduly burdensome. The estate argues that it should be required to review only e-mails exchanged between the heirs and those exchanged with the decedent or his estate planners. The estate also argues that the only metadata associated with e-mails are the sender, the recipients, and the transmission dates.

Because all of that information appears on the face of a PDF print out of an e-mail, the estate's lawyers argue that they should be allowed to produce the messages in PDF format along with the native files of any relevant attachments. The estate also wishes to filter out all the e-mails in the heirs' accounts except for those between the heirs, the decedent, and the decedent's estate planners, thereby restricting its scope of review.

In a dispute before the court, there are several questions that the Judge may want to consider in order to make an accurate and just determination of the discovery dispute. Is the plaintiff asking for too much, or is his request for defendants to review the larger set of 100,000 e-mails reasonable? Should the estate be required to produce the e-mails in a format that is compatible with the e-discovery software used by the plaintiff's firm? Should the estate be required to produce the metadata associated with the e-mails? Does it appear that the estate is hiding something by trying to avoid producing the metadata? If the estate should be required to produce the e-mails in a compatible format, should the associated costs be taxed to the plaintiff?

Additional facts about the parties and their counsel may shed additional light on the court's analysis. None of the parties has ever been involved in civil litigation, so they all reside at Level 1 on the e-discovery maturity scale. The plaintiff's attorneys work at a large law firm that is at Level 5 on the e-discovery maturity scale. The plaintiff is an executive at a Fortune 500 company that the firm represents on other matters. The plaintiff's law firm has a practice group entirely devoted to e-discovery. The firm's e-discovery software was purchased at a sunk cost, is hosted on the firm's servers, and acts as a profit center for the firm. In fact, the firm has created a separate entity that serves as an e-discovery vendor for smaller firms.

The estate's attorneys work at a small firm and reside at Level 2 on the e-discovery maturity scale. The estate's law firm outsources its e-discovery obligations on a case-by-case basis, but only if a review job is simply too big for its small operation to handle. For the estate's firm, the only way to meet the plaintiff's request for review of all 100,000 e-mails and production of e-mails in a format compatible with plaintiff's e-discovery software is to process the data with its e-discovery vendor's web-based software.

The estate's e-discovery vendor charges a flat rate of $50 per month for each gigabyte of data hosted on its system. The vendor's e-discovery system retains multiple versions of uploaded files in order to be compatible with other e-discovery platforms. As a result, uploading one gigabyte of data results in four gigabytes of data that is hosted on the vendor's system. Accordingly, it would cost the estate $6,000 per month (30 GB x 4 GB file expansion x $50 per month) to process the

amount of data requested by the plaintiff. Because the estate would need to keep the data hosted at least until the close of the discovery period, which is expected to last 6 months, it would cost around $36,000 to process the data with the attorneys' e-discovery vendor. This does not include the cost of the 300 attorney hours it would take to review the e-mails, which, assuming the review could be accomplished by associates with an average rate of $200 per hour, could approach $60,000.

In contrast, the more limited review set of 10,000 e-mails would only cost $600 per month to process through the e-discovery vendor's software. The 30 hours of review time associated with the reduced set would likely cost less than $6,000. If the estate was allowed to simply produce the e-mails in PDF format, the defendants would avoid $600 per month in costs associated with processing the data with the e-discovery vendor's software as needed to create a load file that is compatible with the e-discovery software used by the plaintiff's firm. Because the estate would need to keep the data hosted at least until the close of the discovery period, which is expected to last 6 months, allowing the estate the produce the documents in PDF format would translate into a $3,600 savings. If the defense wins the argument, the review would likely cost less than $6,000 total to the estate.

As demonstrated by the hypothetical, in addition to considering the nature of the case and the importance of the issues, examining the relative sophistication of the parties and their counsel can provide further information about the practical effects (including costs) of e-discovery requests. This analysis can thus aid the court in setting reasonable expectations for both discovery and e-discovery and reaching determinations that are just and equitable for all concerned.

Scope of the "Practical Glossary"

The e-discovery process begins when a lawsuit becomes reasonably foreseeable. The process can be summarized as containing three basic steps: (1) identification, preservation, and collection of ESI, (2) review, and (3) production. This practical glossary intends to demystify some of the more common technical terms involved in these three phases of the e-discovery process. For ease of reference, this glossary has been organized into those three sub-parts.

It is not difficult to find the technical definitions of e-discovery terminology. A quick web search will yield numerous glossaries. Some of them are hundreds of pages long and bask in the minutia of what constitutes a bit versus what constitutes a byte. In contrast, this article aims to function as a practical glossary that explains some of the more ubiquitous terms and concepts associated with the collection, review, and production phases of e-discovery. This glossary will contain references to the technical definition of the term and, when warranted, will seek to translate

the definition into practical terms. In certain instances, the relative benefits and burdens that may be associated with a process are also included.

Lastly, it is helpful to note that adverse parties generally fall into one of two categories: expansion advocates or restriction advocates. Restriction advocates usually have a defensive posture. They invariably argue that their opponent's discovery request results in an undue burden caused by the inefficiencies involved in reviewing and producing the requested items. Expansion advocates are usually the aggressors in the discovery dispute. They argue that the requested information is vital to their claim, and that the request imposes only a minimal or reasonable burden. Accordingly, whether an e-discovery method is classified as a "benefit" or "burden" largely depends on the party's posture relative to the claim. This manuscript takes the middle road and seeks to address the benefits and burdens associated with employing a particular e-discovery method from a neutral position. The underlying presumption is that cheaper and faster e-discovery is preferred; that is, of course, so long as the method results in an acceptable margin of error for the particular case (for example, 90%-95% of responsive documents are reviewed and produced).

Glossary Attribution re: Technical Terminology Definitions

All of the technical definitions appearing in this glossary are attributed to The Sedona Conference Glossary: E-Discovery and Digital Information Management (Fourth Ed., 2014) *available at* http://thesedonaconference.org/download-pub/3757.

Glossary: Step 1, Identification, Preservation, and Collection of Electronically Stored Information (ESI)

Term/Concept: ***Archive* or *Archival Data* or *Backup***

Technical Definition:

> Information an organization maintains for long-term storage and record keeping purposes but which is not immediately accessible to the user of a computer system. Archival data may be written to removable media such as a CD, magneto-optical media, tape or other electronic storage device or may be maintained on system hard drives. Some systems allow users to retrieve archival data directly while other systems require the intervention of an IT professional.

Practical Definition:

> Archival data is a backup of information that typically includes only a small portion of a hard drive and only the information that the archiving individual selects for inclusion. Archives are typically

compressed into various different file types to save hard drive space. Accordingly, an "archive" is inherently different and less complete than an "image" or "bit stream backup" of a hard drive.

Benefit:

Archive files can be as small or as big as the person storing the files wishes to make them. This makes file transfers and data imports faster.

Burden:

Archive files are inherently less complete than hard drive images and bit stream backups. The user who compiles the archive has the ability to manipulate its contents. When a party agrees to produce an "archive" or "backup" of a hard drive, as opposed to an "image" or "bit stream backup," the contents may have been manipulated by the user who created the archive or backup.

Term/Concept: ***Compliance Search***

Technical Definition:

The identification of and search for relevant terms and/or parties in response to a discovery request.

Practical Definition:

The compliance search is the search performed by a party after receiving a production request. Its purpose is to identify files, documents, or e-mails that are responsive to the discovery request. It encompasses all of the ESI that was identified as potentially responsive and collected by the receiving party.

Term/Concept: ***Computer Forensics***

Technical Definition:

The use of specialized techniques for recovery, authentication, and analysis of electronic data when an investigation or litigation involves issues relating to reconstruction of computer usage, examination of residual data, authentication of data by technical analysis or explanation of technical features of data and computer usage. Computer forensics requires specialized expertise that goes beyond normal data collection and preservation techniques available to end-users or system support personnel and generally requires strict adherence to chain-of-custody protocols.

Practical Definition:

> Computer forensics is the process of harvesting and analyzing all of the data, including partially deleted data, hidden files, and logs of system activity, as they existed on a particular system on the date the system is received by the analyst.

Benefit:

> Computer forensics can be a useful technique in cases when there is a concern about deception or a party that was motivated to "cover his tracks," such as financial fraud cases. This method helps to ensure the capture of all metadata associated with the files on the system. Typically, the forensic analyst harvests hidden files and partially deleted files, which are unreadable until restored by the analyst.

Burden:

> Performing computer forensics requires the services of an outside technology vendor. The costs vary greatly depending on the scope of the job but can easily and quickly exceed $10,000.

Term/Concept: **Computer Forensics - *Bit Stream Backup* or *Forensic Copy* or *Drive Image* or *Mirror Image of Drive***

Technical Definition:

> A "bit stream backup" or "drive image" or "forensic copy" is a sector-by-sector/bit-by-bit copy of a hard drive; an exact copy of a hard drive, preserving all latent data in addition to the files and directory structures.

Term/Concept: **Computer Forensics - *Deleted Data***

Technical Definition:

> Information that is no longer readily accessible to a computer user due to the intentional or automatic deletion of the data. Deleted data may remain on storage media in whole or in part until overwritten or wiped. Even after the data itself has been wiped, directory entries, pointers or other information relating to the deleted data may remain on the computer. Soft deletions are data marked as deleted (and not generally available to the end-user after such marking) but not yet physically removed or overwritten. Soft-deleted data usually can be restored and accessed by a computer forensics specialist.

Practical Definition:

Deleted Data comes in two basic forms. The first type of deleted data is "wiped data" that has been deleted and overwritten using a special computer program. The second type of deleted data consists of files that, on a typical operating system, have been emptied from the recycle bin. They can be retrieved using forensics programs until the physical space occupied on the hard drive has been overwritten by another file or disk formatting software.

Term/Concept: **Computer Forensics - *Hidden Files or Data***

Technical Definition:

Files or data not readily visible to the user of a computer. Some operating system files are hidden to prevent inexperienced users from inadvertently deleting or changing these essential files.

Practical Definition:

Most operating systems allow users to hide files by changing the file's properties to "hidden" using a file explorer program. The files are not visible on the graphical user interface (e.g. desktop) but the file remains present on the computer's hard drive. In addition to hiding files, users can hide data within files. For example, a user can hide a tab within a Microsoft Excel file to prevent any future viewers from knowing its true and complete content.

Term/Concept: **Computer Forensics - *Latent Data* or *Residual Data***

Technical Definition:

Deleted files and other ESI that are inaccessible without specialized forensic tools and techniques. Until overwritten, these data and files reside on media such as a hard drive in unused space and other areas available for data storage.

Practical Definition:

Latent data or residual data is the data left behind after a user "soft-deletes" the file by emptying the recycle bin. It resides on the hard drive until the physical space it occupies is overwritten or formatted.

Term/Concept: ***Custodian or Record Owner***

Technical Definition:

A custodian is an individual responsible for the physical storage of records throughout their retention period. In the context of electronic records, custodianship may not be a direct part of the records management function in all organizations.

Practical Definition:

The term custodian indicates the person who owned/operated the device or account from which a particular piece of ESI was harvested.

Term/Concept: ***Early Data Assessment***

Technical Definition:

The process of separating possibly relevant ESI from non-relevant ESI using both computer techniques, such as date filtering or advanced analytics, and human assisted logical determinations at the beginning of a case. This process may be used to reduce the volume of data collected for processing and review.

Practical Definition:

Early data assessment is the first step in the e-discovery process. It occurs before a party has received any discovery requests. Typically, the client works with an attorney to determine relevant date ranges, custodians, and locations of responsive ESI. This helps the party and its attorney to estimate costs and timelines that will be associated with responding to e-discovery requests.

Term/Concept: ***Electronic Discovery (E-Discovery)***

Technical Definition:

The process of identifying, locating, preserving, collecting, preparing, reviewing, and producing ESI in the context of the legal process.

Term/Concept: ***Electronically Stored Information*** **(ESI)**

Technical Definition:

As referenced in the United States Federal Rules of Civil Procedure, information that is stored electronically, regardless of the media or whether it is in the original format in which it was created, as opposed to stored in hard copy (i.e., on paper).

Term/Concept: ***Information Governance***

Technical Definition:

> Information governance is the comprehensive, inter-disciplinary framework of policies, procedures, and controls used by mature organizations to maximize the value of an organization's information while minimizing associated risks by incorporating the requirements of: (1) e-discovery, (2) records & information management, and (3) privacy/security into the process of making decisions about information.

Practical Definition:

> Information governance is the practice of crafting and implementing policies related to the storage and handling of ESI, which is typically conducted by mature organizations. They consider the legal requirements of e-discovery and incorporate legally compliant information retention, destruction, and storage measures into their corporate practices to reduce e-discovery costs.

Term/Concept: ***Log File***

Technical Definition:

> A text file created by an electronic device or application to record activity of a server, website, computer or software program.

Glossary: Step 2, Review

Term/Concept: ***Algorithm***

Technical Definition:

> With regard to electronic discovery, a computer script that is designed to analyze data patterns using mathematical formulas, and is commonly used to group or find similar documents based on common mathematical scores.

Practical Definition:

> An algorithm is a computer program that can be used to search through the text in a batch of digital files for the purpose of identifying digital documents that concern a certain topic. For example, Google searches use algorithms to search for, identify, and return a list of webpages that Google believes to have the highest probability of being responsive to the topic a user wishes to learn about, based upon the keywords typed into the search field by the user.

This concept can be contrasted with the "find" function from Microsoft Word which only returns a text result that exactly matches the text a user inputs into the search box. An algorithm-based search like Google, on the other hand, can be programmed to include associated concepts. For example a search for "police dog" might bring up a document that does not contain the words "police" or "dog" but instead contains the phrase "canine law enforcement training."

Most e-discovery software suites contain some capability to perform algorithm-based searches that can be turned on or off by the user performing the search. Two examples of the algorithm-based searches that typically appear in e-discovery software are "phonetic matching," which returns documents with words that sound like the keywords used in a search, and "concept matching," which returns documents that contain topics similar to the queried keywords.

Benefits:

Algorithms can be used to expand the number of documents that result from a keyword search by returning documents that contain closely related concepts, spellings, or meanings.

Burdens:

The only burden associated with the most basic use of algorithms (e.g. to expand search results) is the cost of acquiring capable e-discovery software.

On the other hand, there is a significant cost burden associated with the use of algorithms to incorporate computer learning into the e-discovery process. This technique, commonly known as Technology or Computer Assisted Review, is only practical in complex litigation where the significant expense is outweighed by the efficiencies created by culling out non-responsive documents and thereby reducing the time and money spent reviewing the larger set of documents.

Term/Concept: ***Artificial Intelligence (AI)*; see *Technology Assisted Review (TAR)***

Technical Definition:

A subfield of computer science focused on the development of intelligence in machines so that the machines can react and adapt to their environment and the unknown. AI is the capability of a device to perform functions that are normally associated with human intelligence, such as reasoning and optimization through experience. It

attempts to approximate the results of human reasoning by organizing and manipulating factual and heuristic knowledge. Areas of AI activity include expert systems, natural language understanding, speech recognition, vision, and robotics.

Practical Definition:

In the e-discovery context, artificial intelligence is used to manage the analysis and review of very large sets of documents by combining computer-generated algorithms with human learning to facilitate computer learning. This process is known as Technology Assisted Review or TAR as it relates to its use within the e-discovery field. *See* ***Technology Assisted Review*** for more information.

Term/Concept: ***Boolean Search***

Technical Definition:

Boolean searches use keywords and logical operators such as "and," "or," and "not" to include or exclude documents containing the specified terms from a search, and thus produce broader or narrower search results.

Practical Definition:

Boolean searches allow for searches that are more complex than a basic keyword search, which only returns the documents that contain the specified keyword. The most basic Boolean searches involve the use of the "AND," "OR," and "NOT" operators.

Benefit:

As opposed to keyword searches, Boolean searches offer a much higher degree of specific searching. For example, rather than using a search that returns all the documents in a data set containing the word "dog," while searching a pet store's records, the parties could agree to the review of all documents which contain all of the words "dog" and "bite" and "customer." This would return very specific results with a high likelihood of responsiveness.

Burden:

The biggest burden associated with keywords and Boolean searches is that a party must know in advance what words its opponent uses to refer to the matter at issue. For example, a plaintiff named Fizzy Bottling Company may not know that its opponent's employees almost

universally refer to the plaintiff as "FizBot." A keyword search that did not include the term "FizBot" would miss the vast majority of communications regarding the plaintiff in such an instance. Furthermore, an e-discovery software suite is typically needed to perform Boolean searches.

Term/Concept: ***Coding***

Technical Definition:

Coding is the automated or human process by which specific information is captured from documents. Coding may be structured (limited to the selection of one of a finite number of choices) or unstructured (a narrative comment about a document).

Practical Definition:

Coding is the process of reviewing documents and marking them as responsive, non-responsive, or privileged. Parties may also assign other attributes to the documents, like noting its relevancy to a particular issue, for easier future access.

Term/Concept: ***Computer Aided Review*** **or** ***Computer Assisted Review*** **or** ***Predictive Coding***

See Technology Assisted Review (TAR)

Term/Concept: ***Cull*** (verb)

Technical Definition:

Culling is the process of removing or suppressing from view, a document from a collection to be reviewed or produced.

Practical Definition:

Culling is the process of removing non-responsive materials from a review set so that reviewers do not waste time reviewing the items.

For example, a supervising attorney may determine that a batch of e-mails received from the defendant in a wrongful termination suit contains years of e-mails that occurred before the plaintiff interviewed for the position. The supervising attorney would then "cull" all of the e-mails and other documents that pre-date the plaintiff's interview for the position by either marking them all non-responsive or deleting them from the review set entirely.

Benefit:

Culling saves the client and its attorneys time and money that would otherwise be spent on reviewing documents with very little chance of being responsive.

Burden:

There is virtually no burden associated with culling files that clearly lack responsiveness other than the time required to perform the exercise. That time, of course should be more than offset by the time saved by avoiding review of non-responsive documents.

Term/Concept: ***Data Verification***

Technical Definition:

The assessment of data to ensure it has not been modified from a prior version. The most common method of verification is hash coding by using industry-accepted algorithms such as MD5, SHA1, or SHA2.

Practical Definition:

Data verification is the process of detecting a copied file's digital fingerprint and comparing it to the fingerprint of the original file to determine whether the copy is identical to the original.

Benefit:

Data verification allows the parties to verify the completeness of a file after transfer to ensure that no data was lost during transmission.

Burden:

The burden is minimal; programs that perform data verification are free and widely available.

Term/Concept: **Data Verification - *Checksum* or *MD5 Hash* or *Digital Fingerprint***

Technical Definition:

A value calculated on a set of data as a means of verifying its authenticity compared to a copy of the same set of data, usually used to ensure data was not corrupted during storage or transmission.

Practical Definition:

A checksum or MD5 number is the alphanumeric representation of a file's digital "signature" or "fingerprint." The checksum or MD5 hash of any two files can be compared to ensure that one file is an authentic and complete copy of the other. A computer program looks at every bit of information within a file and creates a short alphanumeric code that represents that content's fingerprint. Any differences, no matter how small, will result in a different fingerprint.

Benefit:

Checksums and MD5 hash codes are the basic building blocks for detecting duplicate files within a data set. Removing duplicate files can greatly reduce the time and money spent reviewing the same document multiple times.

Burden:

The primary burden is the cost of using e-discovery software.

Term/Concept: **Data Verification** – ***Hash Coding***

Technical Definition:

A mathematical algorithm that calculates a unique value for a given set of data, similar to a digital fingerprint, representing the binary content of the data to assist in subsequently ensuring that data has not been modified. Common hash algorithms include MD5 and SHA.

Practical Definition:

Hash coding is the process of using a computer program to generate a file's digital fingerprint or "hash code."

Term/Concept: ***De-Duplication*** **(noun) or** ***De-Dupe*** **(verb)**

Technical Definition:

De-Duplication is the process of comparing electronic files or records based on their characteristics and removing, suppressing, or marking exact duplicate files or records within the data set for the purposes of minimizing the amount of data for review and production. De-duplication is typically achieved by calculating a file or record's hash value using a mathematical algorithm. De-duplication can be selective, depending on the agreed-upon criteria.

Practical Definition:

De-duplication is the process of identifying files with the same content and removing the duplicates from a review set to prevent multiple reviews of the same file. A computer program scans the contents of the files and generates a checksum or MD5 hash number for all of the files. Then, a human groups all of the files with the same checksum or MD5 hash number together and marks all but one of the files as a duplicate. Then, the duplicates are either (1) disregarded by the producing party or (2) coded uniformly and, if responsive, produced to the requesting party.

Benefit:

Duplicate files can be a major problem in the review phase of e-discovery, especially where the review of multiple e-mail accounts from within the same company. For intra-company e-mails, at least two copies (sent and received) are stored on a company's servers. The numbers of copies increase for each company employee that is a recipient of the communication. Accordingly, company-wide announcements can result in hundreds or thousands of duplicates. Furthermore, it is common for businesses to circulate copies of reports via e-mail as attachments.

In a case in which a 20-page sales report is circulated to a sales team of 51 employees, the use of de-duplication software could remove 1,000 pages of material from a review set if all of the employees' e-mail communications are being reviewed. This is true for each instance that a report is disseminated. If the report were sent weekly, the total number of pages removed from review would be 52,000. Assuming an attorney-review rate of 50 pages per hour, de-duplication would save the client 104 attorney review hours just for this single report type.

Burden:

De-duplication is often available in e-discovery software suites at no extra cost. The process can take from several minutes to hours for the computer to complete, depending on the size of a review set.

Term/Concept: **De-Duplication** - ***Case-wide De-Duplication*** **or** ***Cross Custodial De-Duplication*** **or** ***Global De-Duplication*** **or** ***Horizontal De-Duplication.***

Technical Definition:

> The process of eliminating duplicates to retain only one copy of each document per case. For example, if an identical document resides with three custodians, only the first custodian's copy will be saved.

Practical Definition:

> Case-wide de-duplication is the removal of all the identical documents loaded into the e-discovery system across all systems and custodians. This differs from "vertical" or "custodian-based" de-duplication, which only removes identical files possessed by a single custodian.

Benefit:

> De-duplication can greatly reduce review time, especially with regard to the review of e-mails with numerous recipients that are custodians whose e-mails will also be reviewed.

Burden:

> The primary burden associated with the process of de-duplication is the cost of obtaining or licensing an e-discovery software suite that includes the feature.

Term/Concept: **De-Duplication** – ***Vertical De-Duplication*** **or** ***Custodian-based De-Duplication***

Technical Definition:

> Vertical or custodian-based de-duplication is the process through which duplicate ESI, as determined by matching hash values, is eliminated within a single custodian's data set.

Practical Definition:

> Vertical or custodian-based de-duplication is a form of de-duplication that is limited to removing duplicate files possessed by a single custodian.

Benefit:

Vertical or custodian-based de-duplication has time and cost saving benefits, but they are limited in comparison to horizontal or case-wide de-duplication.

Burden:

There is no burden above the cost to obtain e-discovery software.

Term/Concept: **De-Duplication** – ***Metadata Comparison***

Technical Definition:

The process of comparing specified metadata as the basis for de-duplication without regard to content.

Practical Definition:

Metadata Comparison de-duplication is the process of comparing the properties of multiple files to determine whether they are duplicates without looking at the files' content.

For example, a document reviewer may place all of the files into a single folder and only review one copy of a file if the names, extension, sizes, and dates associated with multiple files are identical.

Benefit:

This type of de-duplication does not require any additional costs, such as those associated with using an e-discovery software suite to perform de-duplication.

Burden:

Employing metadata comparison as the sole means of de-duplication without at least peeking at the file's contents is not entirely reliable because the contents have not actually been examined, either by a person or electronically.

Term/Concept: ***E-mail Thread or String – Exclusive Review of the Most Inclusive E-mail***

Technical Definition:

A thread is a series of technologically related communications, usually on a particular topic. Threads can be a series of bulletin board messages (for example, when someone posts a question and others reply with answers or additional queries on the same topic). A thread

can also apply to emails or chats, where multiple conversation threads may exist simultaneously.

Practical Definition:

Most modern e-mail applications, such as Gmail and Microsoft Outlook, allow users to view e-mail conversations as a "thread." E-mails are viewed as a "thread" when all of the replies to the original message are grouped together and placed in sequential order to allow for easy review of the entire conversation.

Most e-mail programs automatically thread the replies by including all of the reply messages in a "quoted text" section below the primary message content. Thus, most e-mails contain the whole conversation thread in the quoted text. Although users may have the ability to disable the inclusion of the e-mail thread in their replies' quoted text section, most users do not disable this feature.

Accordingly, it has become common for parties to agree that they must only review the most recent e-mail, provided that it includes the entire e-mail thread.

Benefit:

Parties agree to only reviewing the most recent e-mail in a thread because it saves a substantial amount of time and money during the review process. A party can perform threaded e-mail review using commonly available e-mail programs, such as Gmail and Outlook. This type of review does not require e-discovery software.

Burden:

The main burden is the risk of user-error in not selecting and reviewing the most inclusive e-mail thread. However, the danger of losing responsive information in this scenario is generally minimal. E-discovery software assembles e-mail threads by identifying "near duplicates" and grouping them together. E-mails that do not contain the quoted text of previous e-mails in the thread will not be identified as "near duplicates." Further, native e-mail programs, such as Gmail or Outlook, provide a threaded conversation view whereby the primary e-mail messages can be easily reviewed even if the most recent e-mail does not contain the quoted text.

Term/Concept: ***Keyword Search***

Technical Definition:

A search using any specified word, or combination of words with the intent of locating certain results.

Practical Definition:

Keyword searches are the most basic ESI searches that can be performed. Unless keywords are combined with an advanced search technique, like Boolean searching, the search only returns documents that include the keyword, exactly as spelled, and nothing more or less.

Benefit:

Keywords are useful in that they allow parties to pinpoint ESI pertaining to relevant topics using search terms designed to return a broad range of documents.

Burden:

Keyword searches tend to be simultaneously over and under inclusive. Accordingly, parties tend to use multiple keywords to make up for their inherent under-inclusiveness which creates more work for both parties. Accordingly, a solution is to combine keywords with Boolean and other more advanced search techniques.

A significant burden associated with keywords is that, in order to return a reliable set of responsive documents, a party must know or be able to accurately guess in advance what words are used by its opponent in the documents it seeks.

Term/Concept: ***Linear and Non-Linear Review***

Technical Definition:

These terms refer to the two types of review that can be performed by humans. Linear review workflow begins at the beginning of a collection and addresses information in order until a full review of all information is complete. Non-linear review workflow is to prepare only certain portions for review, based either on the results of criteria, such as search terms, computer assisted review results or some other method, to isolate only information that is likely responsive.

Practical Definition:

> Linear review involves review of an entire set of ESI from start to finish, without skipping over anything or reviewing the documents out of order.
>
> Non-linear review is any other type of review in which documents are not reviewed in chronological order and in which technology or human judgment can be used to cull out non-responsive ESI.

Benefit:

> Non-linear review is both cost-efficient and accurate, largely because documents concerning the same topic are reviewed in groups. For example, non-linear review allows e-mails to be reviewed as grouped conversations.
>
> In fact, studies have shown that non-linear review is more accurate than linear review because reviewers are not forced to bounce around and review different topics. *See* Bennet B. Borden, *The Demise of Linear Review*, available at http://upc.utah.gov/materials/2015civil/The_Demise_of_Linear_Review.pdf.

Burden:

> Non-linear review does not necessarily require e-discovery software, but e-discovery software is required to take full advantage of the technologies that provide the biggest efficiency and accuracy boosters. For instance, a party may review e-mails as conversation chains by simply reviewing e-mails natively in an e-mail program such as Outlook or Gmail. However, e-discovery software can be used to further group similar conversation chains together regardless of the custodian so that a reviewer can examine all of e-mails on a topic in rapid succession.

Term/Concept: ***Natural Language Search***

Technical Definition:

> A manner of searching that permits the use of plain language without special connectors or precise terminology, such as "Where can I find information on William Shakespeare?" as opposed to formulating a search statement (such as "information" and "William Shakespeare").

Term/Concept: ***Near Duplicates***

Technical Definition:

(1) Two or more files that are similar to a certain percentage, for example, files that are 90% similar may be identified as near duplicates; used for review to locate similar documents and review all near duplicates at one time; (2) The longest email in an email conversation where the subparts are identified and suppressed in an email collection to reduce review volume

Term/Concept: ***Over-inclusive***

Technical Definition:

When referring to data sets returned by some method of query, search, filter, or cull, results that are returned overly broad.

Practical Definition:

Search queries that are too over-inclusive return more documents than are responsive. A proper search should be over-inclusive to an acceptable degree.

Term/Concept: ***Proximity Search***

Technical Definition:

A proximity search is a search syntax written to find two or more words within a specified distance from each other.

Practical Definition:

For example, a user can search "coffee /20 burn" to retrieve all documents which contain the words "coffee" and "burn" within twenty words of each other.

Benefit:

Proximity searches allows for more tailored results than a typical Boolean search on a topic. This increases the likely responsiveness of the results. A normal Boolean search on this topic would likely take the form of "coffee AND burn," and it would return every document that contains the two terms.

Burden:

None, other than construction of the search in a manner most likely to return an appropriate review set.

Term/Concept: ***Technology Assisted Review (TAR)* or *Predictive Coding***

Technical Definition:

> TAR is the process of prioritizing or coding a collection of ESI using a computerized system that harnesses human judgments of subject matter expert(s) on a smaller set of documents and then extrapolates those judgments to the remaining documents in the collection. Some TAR methods use algorithms that determine how similar (or dissimilar) each of the remaining documents is to those coded as relevant (or non-relevant, respectively) by the subject matter experts(s), while other TAR methods derive systematic rules that emulate the expert(s) decision-making process. TAR systems generally incorporate statistical models and/or sampling techniques to guide the process and to measure overall system effectiveness.

Practical Definition:

> TAR is a process whereby computers and humans team up to accomplish document review in a faster than normal manner by culling out non-responsive materials and only reviewing documents with an agreed upon likelihood of being responsive. The process is completed as follows:
>
> A computer observes a human while he or she reviews sample sets of documents and marks them either responsive or non-responsive. During that session, the computer records the various data points associated with the documents marked responsive and non-responsive by the human. Using e-mail review as an example, the computer records the dates the e-mails were sent and received, the senders, the recipients, and the conversation topics included in the body of the e-mails. After the computer observes the human's review and marking of a statistically significant number of e-mails, the computer uses an algorithm to search the text and various data points associated with all of the remaining e-mails in the set. Then, the computer assigns a number to every single e-mail that indicates the percentage probability that the e-mail is responsive based on the information learned during its observation of how the human marked the e-mails.
>
> The e-mails can then be sorted and reviewed according to their likely responsiveness. The producing party then reviews the e-mails in order from the most likely to be responsiveness to the least likely to be responsive. This continues until the burden or expense of reviewing the e-mails outweighs the likely benefit or until the percentage

probability reaches a level agreed upon by the parties. For example, the parties might agree that the producing party would only be required to review documents with a 25% chance or greater of being responsive. The parties' agreement would obviously depend on the size of the case and the number of potentially responsive e-mails.

Benefit:

The greatest benefit of TAR is that it can be used to pare down or "cull out" documents that are unlikely to contain responsive information through the use of computer learning. This can greatly reduce the amount of time needed to perform the first pass relevancy-related review of a set of documents.

Burden:

There is a significant cost burden associated with the use of algorithms to incorporate computer learning into the e-discovery process. Accordingly, this technique, commonly known as artificial intelligence review, is only practical in complex litigation when the significant expense is outweighed by the efficiencies created by culling out non-responsive documents and thereby reducing the money paid to attorneys to review the larger set of documents. There is no bright line indicating the point at which the use of artificial intelligence and TAR becomes cost-efficient. Some practitioners have suggested that a minimum expense level of $25,000 in costs associated with first-pass relevancy review is the lowest expense level at which TAR is a viable alternative to traditional review.

Term/Concept: ***Under-inclusive***

Technical Definition:

When referring to data sets returned by some method of query, search, filter, or cull, results that are returned incomplete or too narrow.

Practical Definition:

An under-inclusive search is a search that returns less than all of the responsive ESI. Parties should not rely on a single under-inclusive search when identifying responsive ESI for review.

Glossary: Step 3, Production

Term/Concept: ***Data Verification***

Technical Definition:

The assessment of data to ensure it has not been modified from a prior version. The most common method of verification is hash coding by using industry accepted algorithms such as MD5, SHA1, or SHA2.

Practical Definition:

Data verification is the process of detecting a copied file's digital fingerprint and comparing it to the fingerprint of the original file to determine whether the copy is identical to the original.

Benefit:

Data verification allows the parties to verify the completeness of a file after transfer to ensure that no data was lost during transmission.

Burden:

The burden is minimal as programs that perform data verification are free and widely available.

Term/Concept: **Data Verification - *Checksum* or *MD5 Hash* or *Digital Fingerprint***

Technical Definition:

A value calculated on a set of data as a means of verifying its authenticity to a copy of the same set of data, usually used to ensure data was not corrupted during storage or transmission.

Practical Definition:

A checksum or MD5 number is the alphanumeric representation of a file's digital "signature" or "fingerprint." The checksum or MD5 hash of any two files can be compared to ensure that one file is an authentic and complete copy of the other. A computer program looks at every bit of information within a file and creates a short alphanumeric code that represents that content's fingerprint. Any differences, no matter how small, will result in a different fingerprint.

Benefit:

Checksums and MD5 hash codes are the basic building blocks for detecting duplicate files within a data set. Removing duplicate files

can greatly reduce the time and money spent reviewing the same document multiple times.

Burden:

The primary burden is the cost of using e-discovery software.

Term/Concept: **Data Verification** – ***Hash Coding***

Technical Definition:

A mathematical algorithm that calculates a unique value for a given set of data, similar to a digital fingerprint, representing the binary content of the data to assist in subsequently ensuring that data has not been modified. Common hash algorithms include MD5 and SHA.

Practical Definition:

Hash coding is the process of using a computer program to generate a file's digital fingerprint or "hash code."

Term/Concept: ***Form of Production***

Technical Definition:

This term refers to the specifications for the exchange of documents and/or data between parties during a legal dispute. It is used to refer both to file format (e.g., native vs. imaged format with agreed-upon metadata and extracted text in a load file) and the media on which the documents are produced (paper vs. electronic).

Practical Definition:

The phrase "form of production" refers to the method and manner in which the parties produce the ESI. The parties may agree to a paper or digital production. If a paper production is agreed to, the parties may further agree to provide print-outs of metadata associated with the printed files at their discretion. The term also refers to whether a party produces the files in their native format (e.g. a word document being produced as a .doc file) or in some alterative format (e.g. a word document being produced as a PDF file).

The parties may agree to allow one another to strip the metadata from the files being produced or require the production of metadata. The parties may agree to production in the form of a "load file" bundle that includes or excludes metadata. A "load file" usually consists of the digital files in whatever format the parties have agreed to in

combination with a database file that allows the recipient's e-discovery software to easily sort the files received.

Benefit:

Parties should be encouraged to agree upon the form of production early to prevent last minute discovery disputes.

Burden:

There may be costs associated with producing ESI in a load file that is compatible with the requesting party's e-discovery software. Accordingly, it may be appropriate to shift the costs associated with meeting a request for production of a load file that is compatible with the system of the requesting party. This is especially true where the producing party would not otherwise utilize e-discovery software to assist in its review of ESI.

Term/Concept: ***Load File***

Technical Definition:

A file that relates to a set of scanned images or electronically processed files that indicates where individual pages or files belong together as documents, to include attachments, and where each document begins and ends. A load file may also contain data relevant to the individual documents, such as selected metadata, coded data, and extracted text. Load files should be obtained and provided in prearranged or standardized formats to ensure transfer of accurate and usable images and data.

Practical Definition:

A load file is a file that contains all of the ESI from a production coupled together with database files that allow the recipient's e-discovery software to load the data on the recipient's system and to sort the ESI easily. In some cases, the load file constitutes the entire ESI production.

Benefit:

Producing ESI in the form of a load file makes it very easy for the recipient to access and evaluate the data produced. This results in significant efficiencies for the recipient.

Burden:

When the parties both use e-discovery software capable of outputting load files compatible with the other party's system, there is little burden. However, not all producing parties use e-discovery software. Further, the software a party uses may not be capable of outputting a load file compatible with the recipient's system. Thus, cost-shifting may be appropriate in situations when a recipient requests production as a specific load file and the producing party either does not use an e-discovery software suite or uses a software suite that is not capable of producing the requested load file type.

Term/Concept: ***Make-Available Production***

Technical Definition:

Process by which a generally large universe of potentially responsive documents is made available to a requesting party; the requesting party selects or tags desired documents, and the producing party produces only the selected documents.

Term/Concept: ***Metadata***

Technical Definition:

The generic term used to describe the structural information of a file that contains data about the file, as opposed to describing the content of a file.

Practical Definition:

Metadata is the information stored in a file other than the file's contents. In other words, it's the "data about the data." Metadata is also known as the "properties" of a file.

There are two types of metadata: user generated metadata and system generated metadata. User generated metadata varies by file type and generally can include the author, title, editor, the person who last accessed the file, the edits a person made, and user descriptions of the file, among other things. System generated metadata are fields that are automatically generated by the device's operating system. These fields may indicate the following information about a file: Date Created, Date Last Accessed, Date Last Modified, Date Sent (emails and other messages), Date Received (emails and other messages), the location of a file on a particular hard drive, and GPS coordinates

reflecting where the record was created (photos or videos), among other things.

When using the Microsoft Windows operating system, system generated metadata can be accessed by right clicking a file and selecting the "properties" list item. On Mac computers, this is accomplished by pressing the Ctrl button while clicking the mouse button and selecting the "get info" item. The process for accessing user-generated metadata varies depending on the file type. Using Microsoft Word as an example, user-generated metadata is accessed by opening a Word file and selecting the "info" menu item from the "file" dropdown list.

Benefit:

The production of metadata allows the recipient to glean file information that is otherwise undetectable from the contents of the file. Furthermore, production of metadata can occur with little or no logistical cost to the producing party by simply producing the files in native format.

Burden:

The production of metadata can involve challenges, particularly regarding the accuracy of user-generated metadata. As opposed to system generated metadata, user-generated metadata can be manipulated by anyone in the file's chain of custody.

Term/Concept: ***Migration***

Technical Definition:

Moving ESI from one computer application or platform to another; may require conversion to a different format.

Term/Concept: ***Native Format Production***

Technical Definition:

Electronic documents have an associated file structure defined by the original creating application. This file structure is referred to as the native format of the document. Because viewing or searching documents in the native format may require the original application (for example, viewing a Microsoft Word document may require the Microsoft Word application), documents may be converted to a neutral format as part of the record acquisition or archive process. Static format (often called imaged format), such as TIFF or PDF, are

designed to retain an image of the document as it would look viewed in the original creating application but do not allow metadata to be viewed or the document information to be manipulated unless agreed-upon metadata and extracted text are preserved. In the conversion to static format, some metadata can be processed, preserved, and electronically associated with the static format file. However, with technology advancements, tools and applications are increasingly available to allow viewing and searching of documents in their native format while still preserving pertinent metadata. It should be noted that not all ESI may be conducive to production in either the native format or imaged format, and some other form of production may be necessary. Databases, for example, often present such issues.

Practical Definition:

Native format production is the process of producing ESI in its original format and as it existed on the producing party's computer system. It allows the recipient to view and manipulate the files just like the original user. However, the system generates metadata indicating the date of last modification that should reveal any attempts to manipulate the file.

Benefit:

Native format production allows the recipient to detect user generated metadata that is otherwise undetectable in a static format production involving PDF or TIFF files. For example, a Microsoft Excel file might contain a hidden tab with important information. A native format production would allow the recipient to view any hidden tabs within the spreadsheet.

Burden:

Authenticity can become an issue since documents can be easily manipulated after production.