E-Discovery—Managing the Risks to Get the Reward: Best Practices for Construction Lawyers and Their Clients

By Melissa Beutler Withy, Patrick R. Kingsley, and Peter Bogdasarian

Melissa Beutler Withy

Patrick R. Kingsley

Peter Bogdasarian

In 1965, Gordon Moore observed that the number of components per integrated circuit would double every year;[1] in 1975, he revised this forecast to doubling every two years.[2] These observations are commonly known as "Moore's Law" and reflect the belief that the speed and capability of computers will increase every couple of years, even while the cost of computers trends down.

Whether or not Moore's Law remains an accurate observation concerning the pace of technological advancement,[3] the effects of the last 50-some years are all around us. Most Americans (81 percent) own a smartphone, while nearly three-quarters of U.S. adults own a desktop or laptop computer, and roughly half own a tablet computer.[4] The result of this technological advancement is the complete transformation of how companies do business in the modern era. Face-to-face meetings and phone calls now compete with email, texting, and other methods of communication. Digital files now replace old-fashioned banker's boxes, and an entire warehouse's worth of documents can now sit on a single shared server.

These technological developments gave rise to an entire new specialization inside litigation—electronic discovery or "e-discovery." In this article, we will provide a general overview of the e-discovery life cycle once litigation has commenced. We will start with why mastering this area is important to counsel, both in-house and external, and provide a general overview of some of the significant concepts and terms deployed in the context of e-discovery. A discussion of the process of identifying and preserving data then follows, with examples of potential pitfalls and things to avoid. Once the data have been identified and preserved, the discussion then turns to how best to collect and retrieve it for future use. Finally, we will discuss review and production.

Why Go Through the Process at All?

When confronted with new and complex ideas and processes, it is common for people to examine means to short-circuit the whole thing and skip to the result. Lawyers are no different, and many lawyers, when faced with discovery questions, including questions about e-discovery, think about the shortcut. In the last decade, many lawyers have learned hard lessons when trying to take a quick and easy approach to e-discovery. Perhaps the following will illustrate.

In Klipsch Group, Inc. v. ePRO E-Commerce Ltd., plaintiff Klipsch, a headphone manufacturer, sued DealExtreme.com, a subsidiary of defendant-appellant ePRO, "alleging that it was selling counterfeit Klipsch headphones."[5] The amount in controversy was only $25,000. Throughout the discovery process, ePRO "engaged in persistent discovery misconduct: it failed to disclose the majority of the responsive documents in its possession, restricted a discovery vendor's access to its electronic data, and failed to impose an adequate litigation hold even after the court directed it to do so," allowing "custodians of relevant electronic data to delete thousands of documents and significant quantities of data."[6]

Klipsch moved for discovery sanctions, and the federal district court issued an order granting in part and denying in part Klipsch's motion, issuing a $2.7 million monetary sanction to compensate Klipsch for its corrective discovery efforts and a corresponding asset restraint in that amount, permissive and mandatory jury instructions, and an additional $2.3 million bond to preserve Klipsch's ability to recover damages and fees at the end of the case.[7] The district court "concluded that Klipsch had shown that ePRO had willfully spoliated relevant Unstructured ESI" and "deemed ePRO to be a dissipation risk in light of its persistent failures to comply with court orders or discovery protocols."[8]

ePRO, naturally enough, balked at a monetary sanction for $2.7 million for a case where the amount in controversy was $25,000, arguing the sanction was "so out of proportion to the value of the evidence uncovered by Klipsch's efforts or to the likely ultimate value of the case as to be impermissibly

punitive and a violation of due process."[9] The court disagreed, holding that the sanctions were "calculated to make Klipsch whole for the extra cost and efforts it reasonably undertook in response to ePRO's recalcitrance."[10] Not only were the sanctions and fees "carefully limited to costs Klipsch incurred in direct response to ePRO's misconduct," but also "Klipsch obtained approval from the magistrate judge prior to each of its substantive efforts, and in each case, that approval was given only after ePRO had already squandered an opportunity to correct its own errors."[11]

The court also held that "there is no special rule requiring parties to suffer an opponent's open and notorious discovery misconduct in small value cases."[12] Even if the "likely valuation of actual damages" in controversy was $25,000, the court held that "tethering monetary sanctions to the ultimate amount in controversy would restrict the court's discretion to a number that remains speculative and indeterminate."[13] The court ultimately held that "the proportionality that matters here is" not whether the amount of sanctions were proportionate to the actual damages in controversy, but rather that "the amount of sanctions [were] plainly proportionate . . . to the costs ePRO inflicted on Klipsch in its reasonable efforts to remedy ePRO's misconduct."[14]

The Klipsch case reinforces the premise that issue of e-discovery cannot simply be ignored. In that case, inattention and a lack of respect for the process of e-discovery turned a $25,000 case into one over 100 times larger in size. Thus, it is imperative for lawyers to spend sufficient resources thinking about and managing the e-discovery process. This warning and advice are not just for junior lawyers or paralegals who may be in the weeds on discovery. Lead trial lawyers need to be intimately involved. If a paralegal or junior lawyer is directing the collection process, they will probably err on the side of massive overcollection for defensibility purposes—at the sacrifice of efficiency and feasibility. But an in-house lawyer, working in conjunction with lead trial counsel, is better positioned to make judgment calls that will add efficiency to the process.

Similarly, a paralegal or vendor "checklist" for collection may trigger requests for broad categories of information that may be related to the subject matter of the litigation but have no bearing on the types of evidence the parties are actually seeking. For example, a client may frequently have dozens of boxes of invoices related to a project stored in the accounting department. Unless the cost or timing of a subcontractor is relevant, scanning and producing this information would simply burden the rest of the discovery process. The involvement of an in-house lawyer or lead trial counsel can help steer this process toward efficiently gathering only probative information.

In the new age of e-discovery, lead trial counsel can no longer delegate gathering and retrieval of discovery information to the paralegal and avoid even thinking about it until later. Rather, lead trial counsel needs to be involved in the key decisions early and often and needs to be proactive in communicating and negotiating with opposing counsel regarding scope and strategy.

Significant Terms and Fundamental Concepts

The Lingo

There are certain terms that are so basic to the practice of e-discovery that we believe it makes sense to discuss them right at the start.

ESI. ESI is an acronym for "electronically stored information," a term codified in the Rules of Federal Civil Procedure in Rule 37(e), which will be discussed in more detail below. Generally speaking, it refers to any readable or usable data or information stored electronically, such as on a computer or a hard drive/

Metadata. Metadata is "an electronic 'fingerprint' that automatically adds identifying characteristics" to electronically stored information, "such as the creator or author of the file, the name of individuals who have accessed or edited the file, the location from which the file was accessed, and the amount of time spent editing the file."[15] A ready example of what we mean by that can be found by right-clicking on a document on a computer's desktop and selecting the "properties" option. A small screen will pop up that will be filled with metadata fields: the size of the file, the date it was created, the date the file was last modified, etc. Because metadata can provide sensitive and relevant information about ESI, it can prove crucial in the course of litigation discovery, as demonstrated in Lawrence v. City of New York.[16]

In Lawrence, plaintiff Lawrence alleged that "NYPD officers entered her home without a warrant, pushed her to the floor, damaged her property, and stole more than $1,000 in cash."[17] Lawrence "provided photographs that she claimed depicted the condition of her apartment several days after the incident."[18] However, when defendants "checked the photographs' metadata, they learned that 67 of the 70 photographs had been taken . . . two years after the incident."[19] The time stamp from the metadata ultimately convinced the court that the photographs were staged, and the efforts to introduce them were "an attempted fraud on this court."[20] Based on this finding, the court dismissed Lawrence's claims with prejudice.[21]

Custodian. A custodian of electronically stored information is defined as "a person having administrative control of a document or electronic file."[22] An example of a data custodian of an email "is the owner of the mailbox which contains the message."[23] A custodian may not necessarily be the author of a document. For example, Anna may have a file she received from Bart on her computer. In this instance, Anna would be the custodian (because it is her machine) and not Bart.

Deduplication. Deduplication is "a technique for eliminating redundant data in a data set" by eliminating the additional copies of the repeated data and only storing a single copy.[24] Deduplication is typically used when processing email to reduce the overall volume that is necessary to review. In a typical e-discovery project, deduplication is conducted at either the custodial or global level. Global deduplication compares each file to the entire data set and preserves only the first instance of a unique document. This approach eliminates the largest number of documents, but because it will save only one copy, it can lead to an attorney team losing the context for a particular

document. For example, a particular file (Hotdoc.txt) may have been collected first from the computer of the custodian Anna, and then, at a later point, a duplicate copy is collected from the computer of custodian Bart. Deduplicating Bart's data on a global basis will remove the copy collected from Bart's machine. If Bart's ownership of the file is important (for example, if a particular document will only be reviewed and potentially produced if it was collected from the computer of custodian Bart), then either the data should not be deduplicated globally or additional steps should be taken in the processing stage to preserve the lost custodial information through a field (commonly called "Additional Custodians").

Custodial deduplication only compares an uploaded file to the set of documents collected from a particular custodian.[25] As with global deduplication, custodial deduplication will preserve the first instance of a unique document, but duplicates will exist across custodians. As a result, more copies of the same data will make it into review. As noted above, this may be a desirable result if the custodial source of a document is particularly important, but it can also lead to an increased burden on the legal team, in terms of both having more documents to review and having a greater burden as to maintain consistency in the coding of reviewed documents.

De-NISTing. De-NISTing refers to the processing of files to remove system files and other file types that are highly unlikely to have evidentiary value. NIST refers to the National Institute of Standards and Technology, who publish a subproject called the National Software Reference Library.[26] For example, the Windows 10 operating system uses between roughly 25 and 40 gigabytes of space, and there will be no value added in the typical litigation by reviewing any of these files.

Families. The term "family" in the context of e-discovery refers to all parts of a group of documents that are connected to each other.[27] The most common usage of the term "family" is to refer to an email family. For example, Anna sends Bart an email and attaches three spreadsheets. We would refer to these four documents collectively as constituting a single family. Preserving the family relationship among documents is important. When conducting deduplication, the best practice is to treat the entire family as a single unique document and to only remove it as a duplicate if all of the documents in the family are an exact match to another family. Take the example given above of Anna sending Bart an email with three spreadsheets. Let's say that Anna kept a copy of each of the three spreadsheets on her computer and did not make any further changes to them. We would not want to remove either the attachments to Anna's email or the stand-alone copies of the spreadsheets from the review set as duplicates of one another because it would be difficult to reconstruct potentially necessary information (like verifying Anna's lack of changes to the copies on her computer) after it was deduplicated in that manner.[28]

Format. Common formats for ESI include native, TIFF, and PDF. Native format means the original format of the rule. Some documents may only be accessible in their native format (for example, a movie file or an Excel spreadsheet). Native files are ordinarily not Bates stamped; instead, the control number is usually applied by providing a placeholder that references the underlying file. Documents in a TIFF format have been converted from the native file to a TIFF image, with one image used for each page of the original document. TIFF images are ordinarily supplied with a load file (discussed below) that provides additional information concerning the documents. TIFF images are not adequate substitutes for videos or large spreadsheets. PDFs are an Adobe format and represent a compromise between the single-page TIFF format and the mixed native format. A document is converted to a single PDF document that contains all of the relevant pages and that can be Bates stamped. As with TIFF images, PDFs are not adequate substitutes for videos or large spreadsheets. The most common format in sophisticated e-discovery matters is to use TIFF images with placeholders for native files that cannot readily be converted to a TIFF image and a load file to supply metadata and searchable text.

Forensic Collections and Imaging. The forensic collection of data is aimed at deploying tools to collect data that will preserve and retain metadata. "Imaging" is the process of converting the contents of a computer into a digital file that can be restored later or, for the purpose of e-discovery, can be processed and loaded to a review tool. The term can encompass a significant range of methodologies and processes. For example, a physical image will take a complete picture of a computer's hard drive, including empty space, system files, what is known as "slack space" (the unused space in a cluster assigned to data),[29] and deleted files. This is expensive and incredibly data intensive (for example, taking a physical image of a one terabyte hard drive means a terabyte of incoming data). Outside of a situation involving fraud or the possible destruction of data, a physical image is often unnecessary. A "logical" image is a more focused collection that captures only active data. Deleted files, slack space, and unused space will not be captured. Finally, a "targeted" image involves the capture of either selected files or folders. It offers the most focused collection, with the additional caveat that there is the potential to overlook responsive data if the collection team is not adequately informed about what to look for.

Load File. A load file refers to a file supplied with an ESI production to allow it to be loaded to a review tool. The load file will ordinarily link the images to particular records (for example, assigning TIFF images ANNA000001 through ANNA000006 to the first document in the production) and will provide additional metadata (identifying ANNA000001 as an email sent from Anna to Bart on October 18, 2018, etc.).

Spoliation. The destruction or alteration of a document that destroys its value as evidence in a legal proceeding.

Proportionality, Relevancy, and Privacy

As with all discovery, proportionality, relevance, and privacy concerns are important to e-discovery. On December 1, 2015, among other changes, a rewritten Rule 26(b)(1) of the Federal Rules of Civil Procedure went into effect. The amended Rule 26(b)(1) confines discovery to "any nonprivileged matter that is relevant to any party's claim

or defense."[30] This eliminates the old authority for a court to order discovery of "any matter relevant to the subject authority," although this authority was rarely exercised in practice. The amendment also deleted the clause allowing information to be discovered "if it appears reasonably calculated to lead to the discovery of admissible evidence." The authors of the amended Rule 26(b)(1) sought to refocus the federal courts on the question of whether discovery is "proportional to the needs of the case." The amended rule offers the following considerations to guide the proportionality analysis: "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." As stated in the committee notes, "[t]he present amendment restores the proportionality factors to their original place in defining the scope of discovery."

The change to Rule 26(b)(1), and its effect on discovery, is worth an entire article in and of itself and is beyond the scope of what can reasonably be covered in this article. Instead, we will provide one example of the impact of the amended Rule 26(b)(1) in an e-discovery context. In Russell v. Kiewit Corp., the plaintiff filed a motion seeking to compel the defendants to produce the email file covering his entire employment with the defendants in its native .pst format.[31] He also requested that the defendants produce "all prior versions of this file (allegedly nine in total) to ascertain whether the files were altered or overwritten at any time during or after plaintiff's employment." The plaintiff argued under the pre-amendment language of Rule 26(b)(1) that granting his motion would lead to the discovery of admissible evidence and that proportionality would be satisfied because producing the .pst format would allow him to more efficiently review the file. The court rejected the request as overly broad and not proportional to the needs of the case. Cases like Russell v. Kiewit Corp. demonstrate that some courts, armed with revised text in Rule 26(b)(1), are pushing back against burdensome e-discovery not properly tailored to counsel's needs.

Preservation of Relevant ESI in Anticipation of Litigation

Understanding when a duty to preserve documents is triggered and then acting to correctly preserve those documents is the single most important step counsel can take with respect to e-discovery. If counsel preserves the necessary documents, then there is always the possibility of returning to the original source to correct later mistakes, accidents, and oversights and/or to adjust the overall approach. But if counsel do not act in time and the data are deleted or otherwise lost, it can be ruinously expensive (as shown by Klipsch) or even impossible to put things right again.

An amended Rule 37(e) also went into effect on December 1, 2015, and lays out the possible penalties:

> (e) Failure to Preserve Electronically Stored Information. If electronically stored information that should have been preserved in the anticipation or conduct of litigation is lost because a party failed to take reasonable steps to preserve it, and it cannot be restored or replaced through additional discovery, the court:
>
> (1) upon finding prejudice to another party from loss of the information, may order measures no greater than necessary to cure the prejudice; or
>
> (2) only upon finding that the party acted with the intent to deprive another party of the information's use in the litigation may:
>
> (A) presume that the lost information was unfavorable to the party;
>
> (B) instruct the jury that it may or must presume the information was unfavorable to the party; or
>
> (C) dismiss the action or enter a default judgment.[32]

The amendments to Rule 37(e) set a higher threshold for the trigger of sanctions. A party must take "reasonable steps" to preserve information, and there is the opportunity to restore or replace the information through additional discovery. Even when a party has failed to meet these standards, the most onerous sanctions (often called "death knell" sanctions) are only appropriate when a party "acted with the intent to deprive another party of the information's use in the litigation." As the first sentence of Rule 37(e) suggests, the best way to avoid ever having to parse the possible sanctions is to act when the duty to preserve is invoked. "When litigation is 'reasonably foreseeable' is a flexible fact-specific standard that allows a district court to exercise the discretion necessary to confront the myriad factual situations inherent in the spoliation inquiry."[33] The reasonable anticipation of litigation is not the same as a potential for litigation. The duty to preserve can be triggered by the notice of conduct underlying a potential claim (such as sexual harassment)[34] or the retention of counsel and/or other experts in anticipation of a suit.[35] One takeaway from all these cases is that reasonable and defensible policies and procedures for determining when the obligation to preserve documents is triggered are far superior to ad hoc practices.

In re Abilify (Aripiprazole) Products Liability Litigation is illustrative of the analysis a federal district court could perform under the revised rule when asked to sanction the other party for failure to preserve electronic information. The plaintiffs in In re Abilify moved for spoliation against one of the defendants, Otsuka America Pharmaceutical, Inc. (OAPI), when the defendant informed them it could not produce email for three custodians for the period 2002–2006 because the company maintained a 60-day retention policy for email during that period.[36] The plaintiffs advanced

the argument that OAPI should have reasonably anticipated litigation during the period in contention because of industry-wide events (including clinical trials, litigation with the DOJ concerning off-label promotion of the drug, and litigation concerning other drugs). The court rejected the plaintiff's argument as "highly problematic because it improperly places too much emphasis on events other than those generated by the plaintiff or those similarly situated to the plaintiff."[37]

Implementation of the Legal Hold

An important part of the preservation of ESI is the "legal hold" communication. Once counsel has made the decision to send out a legal hold,[38] the hard work begins. An effective legal hold must describe the documents to be preserved with particularity and clarity, encourage its recipients to think laterally about potential sources of documents, avoid potential minefields of misunderstanding, and be short enough that the recipients will actually read it. For lawyers, who often believe that verbosity is its own defense, this last element can prove the most challenging. An effective legal hold should begin with a description of the circumstances. A litigation has been filed (or an investigation commenced) and there is a need to preserve documents. Effective immediately and until further notice, the recipients must preserve and not destroy, alter, or delete any and all documents related to the subject matter. Having made an overarching point about why all this is happening; the legal hold should then address the following topics: (1) who to communicate with when there are questions regarding the legal hold, (2) the definition of a document, (3) the scope of the legal hold notice, and (4) how to preserve the documents. The legal hold should also direct the recipient to keep it confidential. As we will discuss below, the legal hold may also request the recipient to acknowledge receipt. The purpose behind defining a document is to assist the recipient of the legal hold to think about something other than the classic filing drawer stuffed with papers or the spreadsheet put together with Microsoft Excel. Twenty-first century society devotes an enormous amount of intellectual and financial resources to innovating new methods by which two people can communicate,[39] and if people aren't talking to one another, then the computers are doing it through things like server logs and the Internet of things.

The scope of the legal hold notice should define both what and when are covered by the legal hold. Defining what is covered by the legal hold is a matter of finesse. Ideally, the what should aim to be somewhat broader than any potential document request, under the logic set forth above that it is better to keep something and not need it than to need something and not have it. On the other hand, in all but the most exigent of matters, the temptation must be resisted to keep every document ever created at the company because a thoughtlessly broad legal hold can have a very real effect on the bottom line—not only because storing and preserving data cost money but also because failure to comply with a legal hold can be a means of advancing an argument for spoliation. The discussion of how to preserve the documents should describe the obligations falling on the recipient, as well as any special considerations. The recipient should be asked to confirm the location of and retain all documents in their possession, custody, or control. This should include directing the recipient to consider other potential storage areas (their mobile devices, a home office, or even the trunk of their car). They should be reminded not to alter, delete, or destroy any documents covered by the legal hold, including maintaining documents as they are filed and assembled (for instance, not removing staples or paper clips), and not disposing of drafts. If ephemeral data like instant messages, handwritten notes, or text messages on a custodian's cell phone are in play, the recipient should be directed to reach out to the point of contact to address these data because any delay may cause them to be lost. Finally, there should be instructions for the recipient on what to do if their position at the company changes, or if they depart, so that their documents are not lost.

A company's IT department (or outside IT vendor) should also ordinarily receive a form of the legal hold. The legal hold sent to IT should discuss the need to ensure automatic deletion mechanisms are turned off or otherwise disabled and instructions on what to do with company-owned laptops and mobile devices when custodians subject to the legal hold depart the company or are upgraded to new technology. If the company has a records management department, they should also receive the legal hold. Consideration should also be given to whether there are documents held by subsidiaries, third-party vendors, and other agents that can reasonably be considered to fall within the company's possession/custody/control because of contractual arrangements and that are necessary to retain. Common examples of these sorts of actors include parent and subsidiary companies, auditors, outside counsel, and terminated or departed employees.

Finally, there is a decision to make with respect to how to record the dissemination of the legal hold. One option is to have the recipients acknowledge receipt, either by returning a signature page or through email.[40] If counsel elects to require acknowledgments, then they need to be obtained from all recipients of the legal hold. It is rare for counsel to know on the very first day who the appropriate custodians or departments should be. In this instance, it is better for counsel to act immediately to begin protecting data and to refine the legal hold as they receive additional clarity. For example, a preliminary notice may be sent to what appears to be the relevant group of custodians based on an organization chart or client interview. The content and audience of this notice can then be refined at a later point with a second communication.

One practical procedure is, upon notice that there is the potential for a claim, to assemble a list of persons who may possess information related to the matter. Then, in a chart form, conduct brief five-minute-long interviews with all custodians to inquire whether they have any of the three types of information above related to the dispute or any additional document or information that should be considered.

If they have such information, then discuss a strategy to copy and preserve it so that it is not deleted or destroyed. This procedure is likely sufficient to show that the client took reasonable efforts to preserve information. It is not a time-consuming process, but it does require being proactive.

In taking these actions, counsel should understand that its role is to provide oversight and accountability to the legal hold process.[41] Counsel should make a practice of journaling to record significant developments in the process—the distribution of legal holds, new information impacting existing holds, and similar decision points. Counsel should also exercise both diligence and their professional skepticism throughout their involvement with the legal hold process. For example, Anna may be vehement in her custodial interview that Charlie was a minor player, but when counsel begins reviewing her emails, it turns out Charlie is a more significant participant than Bart (whom Anna swore was her main contact). In this case, counsel should have a flexible mind and consider that Charlie should receive the legal hold notice (as well as other potential follow-up measures, such as interviewing Charlie).

In EPAC Technologies, Inc. v. HarperCollins Christian Publishing, the defendants either negligently or with "gross inattention" failed to preserve relevant physical evidence and ESI.[42] Among other issues, when the general counsel of the defendant sent his instructions on preservation to business personnel, he did not communicate with the IT department to determine what sources of data might bear on readily foreseeable claims and defenses and whether those data were at risk of destruction, and to instruct the IT department to take the necessary steps to preserve it. Although the legal hold directive instructed employees to disable the automatic deletion of email and take other necessary steps to prevent the loss of data, it was not sent to employees capable of carrying out the necessary tasks and was ignored by the recipients. The court concluded that it was "hard to imagine a circumstance in which [defendant's] steps to preserve ESI would have been considered reasonable."[43] In connection with other remedial measures, the court held that the jury would receive an adverse inference instruction and that the plaintiffs would be permitted to re-depose key witnesses on issues reasonably related to information belatedly produced by the defendant at the defendant's expense, excluding attorney fees.

Sources of Data

The following is a (nonexhaustive) discussion of some of the more common kinds of data encountered in litigation matters, as well as some thoughts for counsel to consider with respect to each category.

Email will often be front-of-mind, for it is the dominant record of communications in the modern era. Counsel should be prepared to explore (potentially with the experience of an IT specialist) whether and how its client archives its email. Are messages maintained by custodians on their desktops or on a virtual desktop? Is there a centralized archiving system? Is there continuous capture at the server level, where messages are recorded when they are sent to the server? If an archiving system exists, how does it address duplicates—is there automatic deduplication, and, if so, does this impact how counsel needs to search the email messages? These (and more) are all questions counsel will want to gather the answers to as soon as possible. While it may have been less common five or 10 years ago, most clients' internal email systems now likely have built-in ability to perform basic culling—segregating emails by custodian, date range, and key words. Having the client cull their own email production with these parameters will greatly shrink the size of the data set that gets transferred to a law firm or vendor for processing and hosting. The client also likely has the ability to deduplicate the emails against each other in searching multiple custodians. Make sure counsel understands and discusses with the client what their capabilities are. Without a short and proactive discussion, the client could fail to use these tools that are easily available to them.

Paper was king for much of human history, and even in the 21st century there is still far too much of it. As mentioned above, custodians may have paper files in their offices, in their homes, or in storage (on-site or off-site). For paper boxes, prepare a detailed index of the boxes and their individual folders. Then, on the matrix, highlight for opposing counsel which files are believed responsive and therefore should be produced. Request opposing counsel to identify any other files that they request (with the invitation that any additional files can be requested later to avoid forcing overcollection). The best practice is to scan paper files in a manner that preserves the relationships among the documents (for example, if a Post-it note has been affixed to a document, scan the document with and without the note) and then to load them to a review database in a manner that will preserve their origin and other relevant information. In extreme cases, where the sourcing can get quite complex, this may call for the tabulation of additional metadata—for example, sourcing documents to the folder of origin in a filing cabinet ("2017 accounting records")—or naming documents in the review system in a manner that makes them easy to trace (using the prefix BARTHC for Bart's hardcopy documents).

For documents on a shared network server, prepare a "file tree" that identifies each folder and a few key levels of subfolders within each level. Then, on the matrix, highlight for opposing counsel which folders are believed to be responsive and therefore should be produced. Request opposing counsel to identify any other files that they request. Then, ensure that files are created in a manner that preserves this file structure.

If there are relevant data contained on mobile devices (such as text messages), counsel should make its collection a priority. Mobile devices are often configured to automatically delete these messages and custodians may not have the technological wherewithal to disable automatic deletion. In addition, the popularity of Bring Your Own Device (BYOD) initiatives means that the mobile devices in question often belong to the custodians themselves

and are at risk of replacement due to either breakage or normal upgrade cycles.

Document management systems are a particular challenge and, if they house relevant information, may require counsel to work with an outside vendor to determine how best to collect them. Examples include specialized construction project management software such as Procore as well as other databases like Salesforce or accounting programs. In some instances, it may be more efficient for counsel to negotiate an inspection or some other form of access to the underlying database than to expend a great deal of resources trying to export the data in a forensically sound manner.

Finally, there is the question of what to do about the company's backups. The first step is for counsel to determine whether there is information that must be preserved that is only available in backup data. This decision will often be informed by determining whether the backups in question are used to archive data or if they are true disaster recovery backups. Disaster recovery backups are normally quite expensive to preserve and access, but they also present the greatest risk of loss because most organizations doing tape backups will recycle the tapes on a regular basis. If there are data that are only contained on backups, the best approach is to put the opponent on notice and then, if they demand its production, to explore the possibility of cost-shifting.

Make sure that databases are created in a manner that allows counsel to easily segregate paper, shared network documents, and emails. This makes searching and review more strategic. That way, counsel can segregate the emails for word searching without triggering hits in documents that were not intended to be searched, such as the specifications, submittals, or subcontracts. If counsel have already reviewed the paper boxes for subject matter, then there is likely no need for privilege review or relevance review of these documents prior to production. Similarly, with shared network drive documents, based on the folder they were in, counsel can often make mass decisions regarding privilege and relevance that can avoid costly review time.

Spoliation

A discussion regarding how and what to preserve would be incomplete without a brief discussion of the ramifications for failure to preserve. From the beginning of e-discovery, spoliation and the failure to preserve data have proven to be a well-trodden battleground. The question then is what effect, if any, amended Rule 37(e) has had upon the frequency and severity of fights over spoliation. A survey of some recent cases suggests that the amendment has had a beneficial effect overall, but that parties should not be complacent and that where there is adequate evidence of intentional destruction of data, a court will not hesitate to apply sanctions under the Rule.

IBM v. Naganayagam concerned a litigation against a former employee for breach of contract, seeking money for rescinded stock options and equity awards.[44] During his deposition, the defendant made reference to strategic plans he prepared during his employment at the plaintiff. The defendant also learned during the depositions of certain of defendant's former co-workers that the plaintiff had not instituted a legal hold over their emails or documents. When the defendant sought the production of the strategic plans from the plaintiff and they could not be produced, he moved for spoliation sanctions consisting of an adverse inference instruction and other sanctions. The court analyzed the allegations under Rule 37(e) and found that the defendant had merely alleged negligent, rather than intentional, behavior on the part of the plaintiff and that this did not meet the standard for an adverse inference instruction. In addition, the court determined that the defendant had failed to establish how the alleged spoliation would be prejudicial to his interests and so denied the defendant's motion for spoliation sanctions.

Wakefield v. Visalus, Inc., saw the plaintiff file a motion for sanctions alleging spoliation of the defendant's call records, which the plaintiff claimed would have provided conclusive proof as to at least one of the elements of its class claim.[45] Plaintiff first learned that these call records were not being preserved as early as December 2016, but waited until February 2019 to file her sanctions motion, more than 15 months after the close of discovery. The plaintiff's unreasonable delay rendered her motion untimely, and the court denied it.

In University Accounting Service, LLC v. Schulton, a defendant admitted at his deposition that on three occasions after his receipt of a document subpoena, he destroyed responsive evidence on his personal computer and in his personal cloud storage account.[46] The court found the admission to constitute facts sufficient to establish that he acted with the intent to deprive the plaintiff of the information's use in the litigation, at least to the extent of depriving the plaintiff of the ability to prove what information the defendant had taken with him when he left the employment of his former company (another defendant). The court ruled that because the defendant acted with intent, it would provide the jury with a permissive inference spoliation instruction.

Plaintiffs in Hernandez et al. v. City of Houston uncovered a series of misrepresentations by the defendant city about its discovery process.[47] These misrepresentations included misreporting the number of potentially responsive documents; representing to the court that the defendant had reviewed documents generated by the plaintiff's search terms when, in fact, they had not; representing that they had issued a litigation hold; and obfuscating the wiping of hard drives of former employees who departed the city's employment after the commencement of the litigation. The court found the pattern of misrepresentations to be either intentional or the result of deliberate indifference and that an adverse inference finding would be the appropriate sanction.

In Waymo LLC v. Uber Technologies, a self-driving start-up and subsidiary of Alphabet Inc. accused Uber of misappropriating its trade secrets.[48] Waymo alleged that a former employee secretly downloaded a significant amount of confidential data and used this to launch a rival company that Uber later acquired. The matter involved a grueling

discovery process with numerous allegations against Uber related to the spoliation of data in its possession, and, ultimately, Waymo moved for an adverse-inference instruction under Rule 37(e).

In considering the motion, the court first examined whether a reasonable party in Uber's circumstances would have foreseen the litigation at the time when it conducted the acquisition. Uber itself had previously argued that it foresaw the acquisition in connection with its claims of joint-defense and common-interest privilege, it had retained litigation counsel to obtain advice regarding its potential liability exposure, and it had commissioned a due diligence investigation to determine its potential exposure. The court therefore found that Uber had a duty to preserve the relevant evidence.

The court then considered whether Uber bore responsibility for the spoliation of evidence. Uber conceded the evidence in question had been lost and could not be restored or replaced but contended no sanction was appropriate because of Waymo's delay in bringing the motion, the irrelevance of the spoliated evidence, and the fact that it had acted in good faith.[49] The court rejected each of these defenses, finding that the delays were due to Waymo's reasonable investigation of the circumstances under which evidence had been lost and that the facts reasonably suggested the spoliated evidence was relevant to the matter. The court reserved the question of bad faith (and the decision on a potential adverse inference instruction) until after Waymo presented its case-in-chief in trial.

Collection and Retrieval of ESI

Generally, the point at which the parties begin collecting documents is when the expensive decisions are made in the e-discovery process. Decisions are made about what to collect (often in connection with negotiations with opposing counsel), and these decisions will define the scope of the most expensive stage of e-discovery: the review and production. Furthermore, vendors may be engaged to assist with collection, processing, and hosting, crystalizing costs that were previously only viewed abstractly. For these reasons, counsel should understand that planning for the collection and retrieval of data is an important opportunity to impact both the cost and scope of the matter. To that end, counsel should begin taking affirmative steps to put themselves and their client in the best position possible to conduct early and proactive negotiations with opposing counsel about scope, strategy, and cooperation.

This planning and preparation begins with developing an understanding of the client's data. What types of data does the client have, how much of the data does it have, and of that segment, how much of the data are potentially relevant to the dispute at hand? Also, what capabilities does the client have to search and retrieve these data, and how can these capabilities best be leveraged to suppress costs and maximize efficiencies? Both to protect against spoliation arguments and to ensure counsel is getting all responsive data collected, a brief interview with each custodian is helpful. The interviews do not need to be lengthy, just consistent and recorded.

The interview can be recorded in a simple template that ensures questions will be asked in a consistent manner across custodians; this will assist counsel in identifying potential areas for follow-up (which, in some instances, may require an additional call to a custodian if new information is learned). The template should record the date of the interview, the interviewee and their position, the identity of the counsel and/or paralegals handling the interview, a brief discussion of the interviewee's knowledge of the documentary evidence, information on the potential data sources (this can be drawn from the discussion above), the interviewee's receipt of the legal hold, and any follow-up questions (along with updates reflecting their resolution). Counsel should also consider having the custodian sign the interview template and a short litigation hold acknowledgment. In many cases, for the first time in a later deposition, a key witness discloses or remembers he or she had a notebook or file of documents that the custodian did not previously disclose. This is obviously ripe grounds for righteous indignation by the other side and potential motions to extend discovery or for sanctions. Having a record that counsel took reasonable, earlier, due diligence steps to identify whether this information existed can be a valuable safeguard.

Commonly, at this stage in the process, counsel will be drawn into the decision of whether the client should self-collect or if a third-party vendor should be engaged to assist with the collection. There is no bright-line rule here. Instead, counsel must balance cost, efficiency, and expertise (the risk of spoliation and the possibility of prejudice) in its decision-making process. For example, if metadata is extremely important, perhaps because of a risk of fraud, then counsel will likely desire to engage a third-party vendor with experience in conducting forensic collections. Moreover, to the extent that discovery may be a subject of dispute in the litigation, counsel should also consider the possibility that a party may be required to provide testimony or make representations regarding its processes for collection and whether it wants to expose client personnel to being called upon to discuss those subjects.

In most construction cases, there is normally little dispute regarding the authenticity of the data produced. The scope of the data, including unintended gaps and oversights, may be a more common discussion. In considering whether to have a client self-collect its data or retain a forensic vendor to collect the data, consider the likelihood of a potential dispute over the collection process. If counsel has a construction case where the collection process may be an issue, it may be advisable to have a forensic firm collect the data so that they can testify to the project. But in most construction cases, simply having an IT professional journal and track what they collected and how is likely sufficient basis to provide a discussion and defense if the collection

method becomes an issue. The risk of spoliation of sensitive data is also a factor to consider. Yet, most standard email retrieval searches or collection from shared network drives does not present a high risk of dispute.

Vendors can provide more than additional labor. Sophisticated e-discovery vendors can also assist with (i) inventorying and understanding the client's IT systems, (ii) advising on the most forensically sound manner of capturing different data sources (such as mobile devices), (iii) auditing and certifying the performance of other vendors, (iv) journaling the collection process, (v) processing and reporting on the documents gathered during the collection process, (vi) hosting the data on a platform so they can be accessed and reviewed, and (vii) preparing documents for production.

While forensic data collection has a key role in other types of civil cases, especially involving those of fraud and concealment or destruction of documents, forensic data collection may not be important in a construction case, especially between sophisticated construction parties with company email systems and shared network drives. While a small portion of data in any construction case may reside only locally on the hard drive of a project manager, usually these data are small in relation to the cost of imaging entire computers and likely only incidental to or duplicative of other information retrieved from company-wide sources. Construction counsel should work together to decide whether forensic imaging of computers is necessary—the cost of both retrieving and the subsequent cost of sorting and hosting may provide the more persuasive factor in the answer. However, in a case where the other party is a small subcontractor that uses only Gmail or other noncompany email and does not use a shared network drive for files, imaging the hard drive of the key representatives may be the only way to retrieve data, much of which may have been deleted or discarded during normal use.

As alluded to in the introduction, at some point in any matter, counsel will need to decide how to limit the data that will be reviewed, produced, and presented. These limitations may be applied when the data are gathered, or counsel may wait until the data have been collected and processed to apply the search terms to identify the set for review. Applying the limitations at the point of collection will reduce the up-front cost of gathering the data, but in turn this raises two risks: (1) Counsel will have less visibility into the collection process and may not be in a position to detect any mistakes and (2) by taking too little up front, counsel may have to go back and get more (for example, discovering that Anna and Bart commonly referred to a project by a codename that was not initially defined as a search term). The better informed counsel is before beginning the process, the greater the opportunity for counsel to mitigate these two risks.

Search Terms and Other Means of Culling

Understanding ways to intelligently cull the data, including fashioning intelligent search terms, is key to managing costs and usability of the information. Search terms and other methods of culling a data set can be based on both keywords and metadata, often in combination. For example, most litigation will apply some sort of date restriction to a data set. Data can also be limited to particular participants in an email exchange. For example, a data set could be limited to only messages exchanged within the company or perhaps to include messages exchanged only with certain specified external email domains. As a best practice, it is useful with a large email set to run reports to identify large families (those containing 10+ attachments) and the most frequent senders and recipients. If custodian Anna receives five digests a day from her favorite news sources and this information isn't relevant to the matter, it may be most efficient to address these documents up front rather than including them in the review set.

Keywords are perhaps the most common method for culling a data set to the essential documents (their popularity no doubt enhanced by serving as a cornerstone of legal research for several generations of incoming lawyers). There is no single formula by which counsel can arrive at the perfect keywords to deploy in a matter, but there are a number of positive steps worth incorporating into any approach. The first is to understand what a keyword search may not find. For example, if the keywords are being run in a native environment (such as inside a custodian's Outlook mailbox), there may be an issue with indexing.[50] Some programs may not index attachments, and thus they may not be searched. Moreover, scanned documents, such as PDFs or hardcopy documents, will ordinarily not contain searchable text until they have been processed.[51] Handwritten notes, in particular, can be nearly impossible to process in an efficient manner and will almost always require manual review to interpret. Finally, documents with access restrictions (such as those requiring a password to access) will ordinarily not be indexed (and they may not be accessible at all without the password).

The second step is to consider the syntax available with which to construct the keywords. Sophisticated e-discovery systems will present counsel with a dizzying variety of options, supplementing the standard Boolean AND/OR searches with finely detailed proximity-based searching (looking for words in proximity to other words). These proximity-based searches can be extremely powerful tools, if they are understood correctly, but often present their own challenges with respect to syntax.[52] A question any counsel should ask when considering a proximity-based search is whether the proximity connector is counting words or spaces—there is a significant difference in meaning between looking for "build" within five characters (including spaces) of "dam" and "build" within five words (of any length and excluding spaces) of "dam." Counsel should also appreciate whether the tool will automatically "stem" words (meaning that it automatically adds a prefix or suffix) or if the root requires the incorporation of a wildcard to achieve the same result. Returning

to our example, some programs may consider "build," "builds," and "building" to be three different keywords, while others will automatically extend "build" to look for all three (and if that was a troubling thought, consider the potential impact of "dam" versus "dams" versus "damn"). Counsel should determine whether there are specific "noise words" that are excluded from indexes and that may not be searched.[53] For example, an application may not index words like "all" or "them," which will cause a search for "sell them all" to return just the word sell (in some systems) or the word "sell" in proximity to any other two noise words.

Finally, there is the design of the keywords themselves. Absolutely, unequivocally bad keywords exist, but determining why particular words are so terrible will often require some measure of experimentation on the part of counsel. Terms that relate to the general business of a company, such as "construction" or "project," will likely be pointless when run over a company whose day-to-day business is exactly that. Short keywords, such as "IT" that can also be in commonplace usage ("it"), will ordinarily be ineffective. Similarly, many special characters, such as the period in an email address or an ampersand, are indexed as spaces, causing a search for "anna@company.com" to be treated as three separate terms ("anna," "company," and "com"). Even a somewhat more focused term, such as "architect," may require additional context (such as the name of a project) to return relevant material. In some instances, counsel may need to run even "bad" terms, simply to build a record as to why these terms are ineffective and/or burdensome.

If at all possible, keywords should be approached as an iterative process. A party should be prepared to test its keywords and to exercise its professional skepticism as to whether the data set makes sense. This skepticism can cut both ways. A party whose keywords returned just 500 documents from a search of 500,000 documents for a multiyear construction project should consider whether that is a reasonable result given what is known about the custodians, their roles in the project, and the frequency of their communications. Or to provide another example, let's say that the keywords return roughly a thousand messages per month except for April 2018, where there are just five documents for 30 days. Did something significant happen in that month to curtail discussion, or is the collection somehow incomplete? Also, if counsel is already aware of significant documents at this stage of the proceedings (whether obtained in connection with the engagement, from the pleadings, from custodial interviews, or from some other source), efforts should be made to determine whether the keywords are effective in identifying these documents.

Another tool available to counsel once data are in hand is to undertake a sampling review. By taking a statistically significant random sample from an identified data set and subjecting the documents to analysis, counsel can often get out in front of potential issues and reach conclusions about the data set at an earlier stage. While a full exploration of how and when to apply statistical sampling is beyond the scope of this paper, it is worth emphasizing that counsel should be able to reach preliminary conclusions about pretty much any data set by reviewing a randomly generated sample of approximately 400 documents.[54] Sampling may help counsel develop an appreciation for potential issues such as (a) the potential presence of privileged communications, (b) keywords that are returning false hits (such as the code name for a project returning company-wide financial and reporting documents that are not the subject of the litigation), (c) syntax that is over- or underinclusive, or (d) the relative occurrence of responsive documents in the data set. Sampling can also be proposed as a means of resolving a discovery dispute.[55]

Counsel in the modern era should expect that at some point, they will be required to sit down and "show their work," whether in the context of negotiations with opposing counsel or through representations regarding the performance of the discovery process on behalf of their client. To this end, counsel should continue to maintain the practice of journaling established during the legal hold process. Record the sources of documents, the locations searched to obtain them, and the information gathered through interviews and other communications. Identify keywords, their frequency, and any modifications that were made to improve their functionality (such as adding a limiter to generic keywords to tie the documents to a specific project).

Finally, in communicating with the other side, accept that there is a give and take and that demonstrating the burden created by bad search terms is most convincing when counsel can marshal the facts in support. Informing a court and opposing counsel that a bad term will return a million additional documents (with the corresponding figure for the expenditure of attorney hours for their review and production based on the review of a random sample) is a much more convincing argument than waving one's hand and saying, "these aren't the terms you're looking for."

The ESI Protocol

It is important to have discussions with opposing counsel about how to treat each of the different categories of documents to ensure that the production set that counsel receives from the other side is as organized and usable as the production set that counsel is assembling with its client's documents. Similarly, the ESI protocol is also helpful to memorialize the agreement of the parties on what to gather and produce, prior to embarking on the exercise of producing the document. While a portion of an ESI protocol includes "technical mumbo jumbo" on production format that can be delegated to vendors for review, the ESI protocol is the place to include key agreements on custodians to produce, data ranges, search terms as well as how to address different types of documents. So,

the lead trial lawyer needs to focus on those portions of the protocol, making sure to be engaged in the plan and parameters that will shape the case. And an effective conversation about the scope of an ESI protocol cannot happen until the trial team has a decent understanding of the size, location, and type of client data that exist.

Pre-production Review

Effective pre-production review should consider the source of the documents. For example, scanned paper files pulled from a segregated project archive likely need little to no pre-production review in the ESI platform. If during collection files were verified to include boxes and boxes of "submittals," those boxes can be mass-coded as "responsive" without paying for lawyers and paralegals to review documents independently.

Similarly, network drive files can also likely be mass-coded as responsive after documents are spot-checked to confirm their contents (and spot-checking for contents is much easier prior to ESI processing when the files are still in their native folder structure with the benefit of Windows Explorer to review). External emails to known project custodians also need little pre-production review. If the email went to an external project custodian, the document is not privileged. Also, if the document went to an external project recipient, it is already in the hands of other parties and, even if not strictly relevant to the project, is likely to turn up in another party's production set. Mass-coding emails to and from known external project custodians is an effective way to limit required pre-production review.

Internal emails require more care and pre-production review. Do not rely on search terms to determine that internal emails are responsive and appropriate for production. For example, the key word search including the project name of the project director or project executive may trigger a number of documents that counsel does not want produced in the litigation—for example, company-wide financial projections where the project is listed as one of many other projects or company-wide client lists and directories where the name of the project is an incidental word in a larger document. To avoid producing this sensitive client information, an effective pre-production review strategy for internal emails needs to be crafted and deployed.

Also, if a project manager on the assigned project also worked on two or three other projects during the same time period, an effective strategy needs to be deployed to cull out emails related to the other projects. Sometimes the best and only strategy is review of every email that wasn't mass-coded out in the external email coding—but the cost and time of that need to be evaluated in terms of the volume of remaining information and the risk of sensitive information. For example, emails to and from known (i) external project custodians on the unrelated project or (ii) internal project custodians who did not work on the subject project can be mass-coded as nonresponsive easily. Another category of internal emails that may be considered for mass-coding as nonresponsive can include those the front desk receptionist sends when there are donuts in the breakroom or a car in the parking lot with its lights on. But internal emails on the unrelated project may need to be reviewed individually, especially if the emails are to and from an internal custodian who was also involved in both projects.

The best way to structure the ultimate strategy may be by (i) mass-coding known related external custodians as "responsive," (ii) mass-coding known internal custodians not related to the project as "nonresponsive," (iii) mass-coding known unrelated external custodians as "responsive," and then (iv) evaluating a sample of the remaining internal emails to determine further culling and review strategies, including individual review.

Review and Production

While the nuts and bolts of operating electronic discovery platforms and working with e-discovery vendors can be delegated to qualified outside counsel, the in-house lawyer plays an essential role in ensuring that the process is conducted efficiently and that trial counsel is appropriately engaged in the process. The days of trial counsel simply delegating the entire task of collecting, sorting, and culling ESI to paralegals are over. The multiplying size of ESI as well as vendor costs makes the e-discovery budget frequently as large as the attorney fee budget. Having an in-house counsel who understands the company records and information and a trial counsel who can assist in making strategic decisions regarding collection and production are key to effective management.

A "just collect everything" approach is outdated and inefficient—and extremely expensive. Overcollection results in large data sets that must be processed and hosted in order to review for production—and this multiplies both vendor costs and attorney fee time to review. A "data dump" approach to production inevitably means that confidential company financial and other strategic and proprietary information are produced to avoid comprehensive review. The best strategy is for in-house counsel to be strategically engaged at the beginning of the process and ensure that lead trial counsel pays attention to the process and manages it strategically.

Joint Review or Production

In a case with several parties, it makes significant sense to structure an agreement to use a single vendor to deduplicate all the project email sets against each other to ensure that the parties only have one copy of each email in the ultimate production set and review set. It is common for one party to deduplicate its own data. It is more progressive to work with other parties on such deduplication. In one representative case, the deduplicate exercise described above, alone, eliminated 30 percent of the data size. It reduced 950 GB of information by 300 GB—a savings of nearly a $100,000 in ESI processing and hosting costs.

There were inevitably significant additional savings in outside lawyer spend because the lawyers did not have to trip through a duplicate volume of data that would have expanded emails by one-third in their review, coding, and preparation. Similarly, having all parties' data hosted by a single vendor, who can control permissions to give the same effect of each party having an independent vendor, can provide significant savings on hosting charges. It can also save time related to production and loading of new sets because the data are all located on a single server.

Master Set of Project Documents

In construction cases, especially those with multiple parties, it is highly practical for the parties to stipulate producing and using a single set of key project documents—such as contracts, plans, specifications, submittals, RFIs, change orders, etc. Even if there is a dispute regarding how such documents should be applied and interpreted, in most cases the contents and authenticity of the documents are not in dispute. Basic project documents can encompass thousands of pages of documents—and if produced by each party, this will bloat the resulting data set. Stipulating to a single set of key project documents not only reduces the size of the data set, it also streamlines eventual deposition and trial exhibits because the parties have already agreed to one version of the documents instead of having each party separately designate as a trial exhibit a document that is identical but for the different party's Bates number (based on who produced it).

Further, trying to stipulate to the documents early on may save the parties cost and time of discovering that they do have a dispute regarding "what is the contract" or "which version of the plans and specs governs." At that point, counsel can agree to disagree and remove the disputed document from the master stipulated set. So, the up-front dialogue isn't wasted time even if it cannot lead to an agreement to stipulate to a single set of master project documents. In conjunction with this strategy, developing a phased production agreement may be helpful. At the beginning of a project, if it does not appear that thousands of pages of daily safety meeting minutes or welding inspection reports are relevant to the dispute, agree not to produce them yet; in the event that they do become implicated in this dispute, they can be produced later. This avoids the need for the old-school "I need every project-related document now, regardless of whether it relates to the dispute" approach.

Quality Control Review

The quality-control process is intended to determine the accuracy and consistency of the coding applied during the initial pre-production review. In designing the quality control review, counsel should identify their worst outcome and ensure their process will address it. Examples of potential worst outcomes are (1) producing a privileged document, (2) failing to produce a responsive document, or (3) producing a nonresponsive document. For example, to address the potential production of a privileged document, counsel should put in place additional searches of documents slated for production. These searches should be run over both metadata (sender/recipient fields) and the body text of messages and, at the very least, should include the names of counsel identified either in preparation for or during the review. Counsel should also consider searching for specific terms. Unfortunately, a term like "privileged" is often used in generic email footers affixed to every message sent or received by a company, but by targeting the substance of the legal advice, counsel can often get to a similar result.

If counsel's principal concern relates to responsiveness, then quality control resources should be dedicated towards a review of the appropriate set (documents coded as nonresponsive to identify additional responsive documents and documents coded responsive to identify false positives). A common and efficient approach is to review a random and statistically significant sample of the nonresponsive documents. Upon finding an error, the next step is to attempt to ascertain why a document was erroneously coded. Was the error simply the result of a missed click, or does it reflect a lack of understanding, either on the part of the reviewer or in the review protocol itself? If the latter, counsel should devise a plan to correct the issue (most likely through targeted searches).

Another method counsel can use to ascertain the review team's understanding of responsiveness is to undertake a consistency review. A consistency review looks at duplicate, near-duplicate, and threaded documents to determine whether they have been coded in a consistent fashion. If counsel knows an agreement is responsive, a search of the database for copies of the agreement can be instructive—if half of the copies are coded responsive and the other half are marked nonresponsive, then counsel clearly has an issue to address. Finally, once counsel has arrived at the final set for production and prepared it, counsel should undertake a sanity check. Are the expected number of documents in the production set? Are redactions visible on the final production documents, and have they been applied as reflected in the system? Was anything in the final production set marked as privileged by the review team?

Post-production Review

The pre-production review is a more streamlined strategy that can use mass-coding, as discussed above. Perhaps the most expensive step from the lawyer-spend side is how the trial teams manage post-production review to prepare for depositions and trial. The eyes-on-every-document approach to review is not appropriate for large construction cases. While in a case with less than 100,000 documents, a trial team may be able to put eyes on every document, in cases with 500,000 documents or in excess of 1,000,000 documents, it is not in the client's best interest to put eyes on every document and manually "code" them to various subject matter categories. Similarly, the strategy

of preparing for a deposition by having the trial team individually review every email to or from a particular custodian is also not cost-effective in large cases.

Sophisticated e-discovery platforms contain a multitude of tools than can be deployed to make review more efficient. Counsel should be aware of the possibility of deploying any and/or all of these tools to assist with the review:

- Email threading (identifying email relationships and grouping them so they can be viewed as a coherent conversation);
- Cluster review (using analytics to categorize and group "like" documents);
- Technology-assisted review; and/or
- Timeline-based review.

Having the team trained and ready to deploy these tools can both expedite the speed and lower the cost of the review. In large email sets, key word searches or very targeted date range review may be much more appropriate than hours of manual review and coding.

It is also important to consider the role of lead trial counsel in structuring a review and coding protocol and procedure. Does lead trial counsel understand the volume of data and the decisions being made by the review team? Has lead trial counsel at least spent some time in the data set to understand the types of documents encountered to identify trends that may streamline review and coding? If trial counsel delegates the entire review and coding process to junior timekeepers without the experience or skill set to make judgment calls, the client is inevitably going to be billed for time spent on overinclusive and unnecessary coding.

Similarly, are the people spending time doing the review and coding ultimately positioned to take the institutional knowledge they are gaining and adding that knowledge to the trial team in an efficient manner? Assigning a team of associates that are not going to work the case to trial to spend hundreds of hours reviewing and coding in a single month is a lost opportunity for the institutional knowledge of the trial team. If the timeline dictates a mass-coding effort by timekeepers who are not intended to work on the case long term, the client's interests are likely better served by employing contract review attorneys rather than associates on loan from other cases at the firm.

Finally, if it is important enough to have eyes on a significant volume of documents in review and coding, isn't it self-evident that some of those hours be spent by key trial team members? If the first time that a deposing lawyer looks at emails in a case is the week before the deposition, and is limited to a predesignated set assembled by others, what has been missed in the process? Even if the deposing lawyer only spends one hour for every 10 to 20 hours spent by the remainder of the review team, isn't it helpful for the deposing lawyer to have some exposure to the data set that is being reviewed and coded to identify trends and scope of the documents that exist to help guide the reviewing team? This level of involvement can provide real-time direction on what documents to focus on and which are likely not relevant, and can identify potential gaps or issues with the review set. This type of review can also provide essential and effective guidance on how to review and code documents to achieve that lawyer's desired objective later.

Technology-Assisted Review

Technology-assisted review (or TAR) encompasses a number of different approaches to deploying algorithms to make the review and categorization of documents more efficient, more effective, and more consistent.[56] The discussion below is a general introduction to the subject but is not a replacement for specialized advice.

The algorithms analyze syntactical patterns, the frequency and patterns of terms, and other characteristics of the documents to rank a document on a scale that typically runs from zero to 100, with the expectation that the farther the document falls to one side of the scale, the easier it is to categorize. This process is known as "training." The initial set may be selected, or it could be a random sample taken from the larger population.

In some reviews, the ranking will be used to "promote" documents. Documents the system ranks towards the responsive end of the scale will be advanced to the front of the review queue, getting the review team to the responsive documents faster and earlier in the life of the matter. At a certain point, the ranking may also be used to cut off the review of documents. In other reviews, the ranking will be deployed to inform the actual categorization of documents. Documents at or above a certain rank will be categorized as responsive, and those falling below a certain rank will be categorized as nonresponsive.[57]

It is also possible that a third category will exist that consists of documents where the system cannot make an informed call, and these will be sent to a human team for review. This is ordinarily an iterative process, in which the system will rank documents, counsel will review the results of the ranking, and then the process will be repeated until counsel has the necessary confidence in the rankings. Ultimately, what happens is to teach the system to code like the reviewer(s). For this reason, it is important that the reviewers providing the input to the system have as accurate a command of the subject matter and review as possible as inconsistency on their part will become the system's inconsistency.

Adding to this challenge is that a ranking itself is not determinative of the outcome of the review. A score of 80 in one matter may indicate that a document is very likely to be responsive, while in another matter (or even, for example, the same matter but a different custodian), it may offer little to no insight into the responsiveness of the document. The methodology by which counsel arrives at an appreciation of the accuracy of the TAR process is

known as validation and requires the random sampling of populations to both inform counsel and feed data back into the system. To provide an example, consider the situation where counsel believes documents with a ranking below 50 are nonresponsive. To establish this fact, counsel will review a statistically sound random sample of the documents falling below the cutoff to determine whether they are classified correctly and what further steps may be required to enhance the accuracy of the rankings and review.

When deployed correctly, TAR can provide massive efficiencies, reducing the time and cost of the review. Studies also show that TAR can greatly enhance the consistency of the review, ensuring that like documents are coded in a similar fashion.[58] Finally, TAR can provide the ability to craft a more transparent process than human review through the ability to test the effectiveness of the algorithms. TAR tools are complex and operate at an intersection of math, technology, and law that can provide any number of challenges for counsel. While courts are prepared to accept the use of TAR technology,[59] if counsel does not fully understand the tools it is deploying and how to properly validate their results, then TAR can magnify the effect of mistakes made during the review.

In In re Domestic Airline Travel Antitrust Litigation, the parties deployed TAR to assist with the document review in a multidistrict class action.[60] The three defendants produced approximately six million documents to the plaintiffs. One of the defendants produced more than 3.5 million of these documents to the plaintiffs. Unfortunately, due to errors in this particular defendant's validation of its TAR process, only 17 percent (600,000) of the documents it produced were responsive to the plaintiffs' requests. Plaintiffs were therefore placed in the unanticipated position of reviewing an extremely large document set. Plaintiffs moved for an extension of discovery five months before the fact discovery deadline. The court granted the request, noting that the plaintiffs had shown appropriate diligence, that there was no prejudice to the nonmoving parties, and that scheduling-related factors did not cut against the plaintiffs (for example, that there would be no need to conduct additional discovery based on the extension).

Conclusion

By now, counsel should recognize that many of the most important decisions in any e-discovery process will be made long before the commencement of litigation. An organization's records management and information governance programs will have an enormous impact on its ability to respond in an agile fashion to the demands of e-discovery. If electronic files are kept in a centralized manner where they are well segregated by project/matter, then an organization will face a greatly reduced burden in identifying and collecting the necessary data for its counsel. Similarly, having sound policies and procedures addressing the invocation and implementation of a legal hold will prevent having to engage in an ad hoc exercise when a problem does come along. The same principles hold true for outside counsel.

The best time to build out processes for identification, preservation, and collection of electronic data is prior to the commencement of litigation, when there is time to identify and develop the needed in-house resources and to build relationships with vendors who can supply the needed expertise. As discussed above, TAR can be essential to making document review less expensive and more efficient, but learning how to operate and deploy a complex tool while keeping to tight litigation deadlines can create significant risks. Any lawyer can advance the argument that they can learn the necessary skills along the way, but having those skills in place on day one is a value differentiator.

Melissa Beutler Withy is vice president and general counsel at Big-D Construction. Patrick R. Kingsley is a partner and Peter Bogdasarian is counsel at Stradley Ronon Stevens & Young, LLP in Philadelphia. A version of this article was presented at the Forum's 2020 Midwinter Meeting in Tucson, Arizona.

Endnotes

1. Gordon E. Moore, Cramming More Components onto Integrated Circuits, 38 Elec. 8 (Apr. 19, 1965), available at https://drive.google.com/file/d/0By83v5TWkGjvQkpBcXJKT1I1TTA/view.

2. Gordon E. Moore, Co-founder, Intel Corp., Speech at Int'l Electron Devices Meeting, IEEE: Progress in Digital Integrated Electronics (1975), available at http://www.eng.auburn.edu/~agrawvd/COURSE/E7770_Spr07/READ/Gordon_Moore_1975_Speech.pdf.

3. Tom Simonite, Moore's Law Is Dead. Now What?, MIT Tech. Rev. (May 13, 2016), https://www.technologyreview.com/s/601441/moores-law-is-dead-now-what/.

4. Mobile Fact Sheet, Pew Rsch. Ctr. (June 12, 2019), https://www.pewinternet.org/fact-sheet/mobile/.

5. 880 F.3d 620, 624 (2d Cir. 2018).

6. Id. at 623.

7. Id. at 624, 626.

8. Id. at 626.

9. Id. at 630.

10. Id. at 631.

11. Id. at 632.

12. Id.

13. Id.

14. Id. at 635.

15. What Is Metadata?, Harv. L. Sch. Info. Tech. Serv., https://hls.harvard.edu/dept/its/what-is-metadata/.

16. No. 15-cv-8947, 2018 WL 3611963, at *1 (S.D.N.Y. July 27, 2018).

17. Id.

18. Id.

19. Id.

20. Id. at *4.

21. Id. at *8. The court ultimately decided not to award attorney fees and costs on the grounds that Lawrence is unemployed and the award would be uncollectible.

22. Chad Main, What Is an ESI or Data Custodian?, Percipient, https://percipient.co/what-is-an-esi-or-data-custodian/.

23. Id.

24. Jennifer Deming Burnham, Understanding Data Deduplication—and Why It's Critical for Moving Data to the Cloud, Druva (Mar. 24, 2015), https://www.druva.com/blog/a-simple-definition-what-is-data-deduplication/.

25. The scope of deduplication can also be configured using other criteria, such as a time period or source.

26. National Software Reference Library (NSRL), NIST, https://www.nist.gov/software-quality-group/national-software-reference-library-nsrl (last visited Aug. 2, 2019).

27. 52 e-Discovery Terms That You Should Know, CDS Legal, https://cdslegal.com/wp-content/uploads/2012/07/52-ediscovery-terms-you-should-know.pdf (last visited July 25, 2019).

28. For an example of why this is potentially significant, consider a matter that turns on the question of whether Anna made additional changes to the spreadsheets after sending them to Bart.

29. For more information on slack space, please see Michelle A. Poolet, What Is Slack Space?, IT Pro Today (Jan. 19, 2010), https://www.itprotoday.com/sql-server/what-slack-space.

30. Fed R. Civ. P. 26(b)(1).

31. Case No. 18-2144-KHV (D. Kan. June 4, 2019).

32. Fed R. Civ. P. 37(e).

33. Micron Tech., Inc. v. Rambus Inc., 645 F.3d 1311, 1320 (Fed. Cir. 2011).

34. Broccoli v. Echostar Commc'n Corp., 229 F.R.D. 506 (D. Md. 2005), appeal dismissed, 164 F. App'x 374 (4th Cir. 2006).

35. Silverstri v. Gen. Motors Corp., 271 F.3d 583, 590 (4th Cir. 2001).

36. 2018 WL 4856767 (N.D. Fla. Oct. 5, 2018).

37. Id. The court also rejected the contention that FDA regulations required OAPI to preserve emails during the period in contention.

38. There are many possible names applied to what the authors are calling a legal hold, such as a "litigation hold," a "preservation notice," or a "document retention notice."

39. As a very circumscribed example, consider the following list: Zoom, Skype, WhatsApp, Facebook Messenger, Bluejeans, LinkedIn, WeChat, Viber, LINE, Telegram, and Weibo.

40. Outlook's voting feature can be a helpful tool for simplifying this process.

41. Zubulake v. UBS Warburg, 229 F.R.D. 422, 432 (S.D.N.Y. 2004) ("A party's discovery obligations do not end with the implementation of a 'litigation hold'—to the contrary, that's only the beginning. Counsel must oversee compliance with the litigation hold, monitoring the party's efforts to retain and produce the relevant documents.").

42. 2018 WL 1542040 (M.D. Tenn. Mar. 29, 2018).

43. Id.

44. 2017 WL 5633165 (S.D.N.Y. Nov. 21, 2017).

45. 2019 WL 1411127 (D. Or. Mar. 27, 2019).

46. 2019 WL 2404512 (D. Or. June 7, 2019).

47. 2018 WL 4140684 (S.D. Tex. Aug. 30, 2018).

48. 2018 WL 646701 (N.D. Cal. Jan. 30, 2018).

49. The court rejected one other instance of alleged spoliation by finding that the individual controlling the information had asserted their Fifth Amendment rights, placing them outside of Uber's possession, custody, and control.

50. Indexing, in general, is the automatic organization of information to enhance retrieval by a search algorithm. This allows the algorithm to search the index rather than having to trawl the underlying data.

51. The process itself is known as optical character recognition (OCR) and reflects the mechanical or electronic conversion of images of text into machine-encoded text that can be searched.

52. For example, some systems will process w/n as within n words, while other systems require the entry of near/n to accomplish the same effect.

53. Most e-discovery platforms in common usage will have a list of noise words attached to the data. These word lists can often be customized to remove particular noise words, in exchange for a reduction in performance.

54. See, e.g., Sample Size Calculator, Raosoft, http://www.raosoft.com/samplesize.html. Reviewing a set of 4,000 documents to a 95 percent confidence level with a 5 percent margin of error (and assuming a 50 percent response distribution, which gives the largest sample size) requires review of 351 documents. For 40,000 documents, that number increases to 381 documents. For 400,000 documents, the recommended sample is 384 documents, and for sets of 4,000,000 or more, 385 documents.

55. See, e.g., Updateme Inc. v. Axel Springer SE, 2018 WL 4952588 (N.D. Cal. Oct. 11, 2018) (in dispute over search term, ordering defendants to review random sample of unreviewed documents containing that term and to report results of sampling review).

56. Technology-assisted review may also be referred to using other terms like predictive coding or machine learning.

57. TAR can also be deployed to inform deposition and trial preparation, in which the categories would instead concern relevant and nonrelevant information.

58. See Maura R. Grossman & Gordon V. Cormack, Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, 17 Rich. J.L. & Tech. 11 (2011), http://scholarship.richmond.edu/jolt/vol17/iss3/5.

59. See Da Silva Moore et al. v. Publicis Groupe, 287 F.R.D. 182 (S.D.N.Y. Feb. 24, 2012), adopted sub nom. Moore v. Publicis Groupe SA, 2012 WL 1446534 (S.D.N.Y. Apr. 26, 2012); see also Rio Tinto PLC v. Vale S.A., 306 F.R.D. 125, 127 (S.D.N.Y. 2015) (finding the acceptability of TAR to be "black letter law").

60. 2018 WL 4441507 (D.D.C. Sept. 13, 2018).