The Advantages of Early Data Assessment

Early data assessment (EDA) allows parties to search, organize and sample a collection of electronically stored information (ESI) early in a case, before the data set is fully processed. By effectively using EDA, counsel can gain valuable information to help shape a litigation strategy and promote cooperation and proportionality in discovery.

GIYOUNG SONG
DISCOVERY COUNSEL
SKADDEN, ARPS, SLATE, MEAGHER & FLOM LLP

Giyoung dedicates her practice primarily to discovery-focused litigation, advice and counsel in federal and state court-based class actions, multidistrict litigation and other disputes. She has extensive experience litigating and managing complex and sophisticated discovery issues, and regularly advises clients regarding discovery law, best practices and practical solutions.

The high costs of e-discovery often prevent the resolution of disputes on the merits, despite evidence showing that only a small fraction (0.1%) of ESI produced in a case is actually used in litigation. Absent changes in the practice of e-discovery, these costs will continue to rise exponentially, as the new digital wave is expected to bring an unprecedented data explosion.

Further, the proposed amendments to the Federal Rules of Civil Procedure (FRCP), which are due to become effective on December 1, 2015, reflect the compelling need for active judicial case management and better cooperation between parties so that the costs and burdens of discovery are proportional to the stakes in the case. The proposed amendments to Rule 1, for example, emphasize that both the court and the parties must apply the federal rules to secure the just, speedy and inexpensive resolution of cases.

Parties can employ EDA tools to handle a data population during the beginning stages of discovery, enabling the parties

to set reasonable discovery limits and ultimately process, host, review and produce less ESI. Moreover, parties can use EDA as part of an early case assessment (ECA) process to gather key information, develop a litigation budget and better manage litigation deadlines. EDA also can foster cooperation and proportionality in discovery by informing the parties early in the process about where relevant ESI is located and what ESI is significant to the case.

EDA TOOLS AND TECHNOLOGIES

EDA can aid the ECA process by shifting data analysis to an earlier point, from the post-processing phase of e-discovery (after the full data set is collected, processed and uploaded for review) to the pre-processing phase. Some technologies allow data analysis to occur even earlier in a case, before the data is collected from locations where it natively resides (for example, on desktops, laptops and e-mail servers). EDA tools are used to:

- Search the data population to identify relevant ESI.
- Reduce the volume of ESI for review by eliminating duplicate or irrelevant data.

By analyzing a larger collection of ESI at the beginning of a case and reserving full processing and review to a smaller subset, the cost of producing documents (including processing, hosting and reviewing the data) should decrease. EDA therefore can prevent e-discovery from controlling a litigation and keep it proportional to the case. However, EDA may not be practical for all cases. Whether EDA should be used depends on several factors, including the type and size of the case, the cost of EDA and the available timeframe. The overall cost of the technology should not exceed the savings it delivers.

IDENTIFYING RELEVANT ESI

EDA tools that can be used to find important and relevant ESI include:

- **Keyword searching.** Most EDA tools can search ESI using multiple keywords and phrases with:
  - Boolean operators (and, or, and not, but not);
  - proximity modifiers (w/10, w/s, w/p);
  - root expanders that find documents with the same word root (read* = reader, reading); and
  - fuzzy searches that find documents with misspellings of keywords (teater) and alternative spellings (theater/theatre).
- **Concept searching.** Some EDA tools allow concept searching, which groups ESI by ideas or subjects. Unlike keyword searches, which identify documents that contain a keyword (shoes), concept searches find documents containing synonyms of the keyword (boots, heels, sneakers) or related concepts (shoe store, podiatrist).
- **Clustering.** Some EDA tools can apply algorithms to group documents with statistically similar content, typically by the number of words that overlap from one document to another. Bayesian method looks beyond the number of common words between documents, and ranks the degree of relevance by placing a value on the words and their relationship, proximity and frequency in comparison to other documents.

Search Predictive Coding: It's Here to Stay for more on ESI search methodologies.

REDUCING THE VOLUME OF ESI FOR REVIEW

EDA tools can increase efficiency and consistency by limiting the review of ESI to unique content. Methods that limit the volume of ESI for review include:

- **De-duplication.** To de-duplicate (remove duplicate documents from the data set), the technology identifies and removes duplicates vertically across the records of one custodian or horizontally across multiple custodians. One copy is retained for review and production, and the names of the custodians whose duplicates are removed are recorded.
- **Filtering.** Filters can sort ESI by numerous criteria, including metadata such as file type, file size, date, source or sender, and recipient and subject. Metadata often is used to de-NIST or remove operating system files, program files and other non-user created files from the ESI.
- **Near de-duplication.** Near-duplicate detection is a variant of clustering that groups documents by similarity in content, such as drafts of contracts, public filings or periodic status reports that contain variations of the same form.
- **Threading.** E-mail threading technology organizes e-mails by conversation and related conversations, which can reduce redundant content. This tool can group:
  - inclusive e-mail threads (the latest-in-time e-mail messages that include the original e-mails and all the succeeding replies);
  - non-inclusive e-mail threads (the succeeding replies and original e-mails); and
  - any related e-mail threads (new e-mail or e-mail threads started from inclusive e-mail threads, for example, an e-mail forwarding an e-mail thread to a new recipient and the replies to the forwarded e-mail).

USING EDA TO SHAPE CASE STRATEGY

Because ESI generally informs counsel about the case, incorporating EDA into the ECA process can help counsel:

- Understand key facts in the case.
- Estimate a litigation budget.
- Prepare for the meet and confer.

FACT ANALYSIS

Early visibility into ESI containing important facts may dictate how the case proceeds. For example, in complex and high-stakes matters, including class actions and multidistrict litigations, it may be critical to uncover important documents quickly and accurately estimate the cost of litigation and potential liability to guide a prudent course of action. EDA may uncover evidence demonstrating individualized issues that prevent class

certification or inform other strategic considerations that arise at the class certification stage. Additionally, identifying key people and documents early in a case may help counsel spot potential risks that raise reputational and publicity concerns and impact business relationships.

EDA can facilitate budgeting by enabling the parties to quantify ESI that may be subject to discovery and therefore accurately project the cost of discovery before it occurs.

BUDGETING

EDA can facilitate budgeting by enabling the parties to quantify ESI that may be subject to discovery and therefore accurately project the cost of discovery before it occurs. For example, filtering and de-duplication tools that identify system files, inaccessible files and duplicate files, which are often excluded from discovery, impact cost estimates. The percentage of responsive and privileged ESI that can be expected based on sampling may also inform cost projections for reviewing documents and preparing privilege logs.

Search Budget Template: Document Production for a model budget template that may be used to estimate the costs of producing documents, with explanatory notes and drafting tips.

EARLY CASE MANAGEMENT

Cases are expected to proceed quickly. For example, FRCP 16(b) requires the judge to issue a scheduling order within the earlier of 120 days after the complaint is served or 90 days after any defendant has appeared. Because parties must have the meet and confer required by FRCP 26(f) at least 21 days before the scheduling order is due, they are left with less than 100 days from the complaint date to prepare for e-discovery discussions. The proposed federal rules further compress the time to prepare for meet and confers by reducing the period within which the judge must issue a scheduling order by 30 days.

Using EDA to obtain information about ESI early in a case, before data collection or processing, can add significant value by enabling the parties to make informed decisions on the scope of discovery in the case within the confines of the brief timeframe.

USING EDA TO PROMOTE COOPERATION AND PROPORTIONALITY

The meet and confer required by FRCP 26(f) provides an opportunity for the parties to tailor discovery to what is proportional to the needs of the case. State courts also require the parties to meet and confer about discovery (see, for example, *Unif. Rules for N.Y. State Trial Courts, 22 NYCRR § 202.70, R. 8*). Most federal district courts and state courts have rules and guidelines about the topics for discussion at the meet and confer, such as:

- Preservation of ESI.
- Search methodologies.
- Limits on ESI production.
- Disclosure of information withheld based on privilege.

Search Rule 26(f) Conference Checklist or see page 64 in this issue for more on the topics discussed at a meet and confer.

Knowledge gained from EDA may facilitate cooperation between the parties and provide the court with specific details to support a concrete showing that the requested discovery is proportional to the needs of the dispute.

ESI PRESERVATION

There is a trend in the law, culminating in the proposal to amend FRCP 37(e), to encourage reasonable preservation of ESI and reduce incentives for over-preservation. EDA technologies that search ESI where it natively resides can identify relevant ESI at the time a litigation hold is implemented and help avoid over-preservation that otherwise results from lack of information about the ESI.

Some EDA tools allow users to sample ESI to identify and preserve sources of relevant ESI. Each sample can be manually reviewed by a subject matter expert to classify relevant documents. Decisions applied to the sample are then extrapolated to the entire collection. Sampling can inform whether an ESI source is important enough to warrant the burden of preserving it and prevent overbroad preservation orders.

For example, in *Pippins v. KPMG LLP*, the parties were unable to agree on a sampling methodology. The producing party moved for a protective order to reduce the number of preserved hard drives from over 2,500 to 100 or, alternatively, to shift the cost of preservation to the requesting party. The court denied the producing party's motion and ordered it to preserve all 2,500 hard drives because of its refusal to sample or otherwise provide discovery about the data on the hard drives. (*279 F.R.D. 245, 249-51, 253-54 (S.D.N.Y. 2012)*.)

SEARCH METHODOLOGIES

Tools used to run searches, such as keyword or concept searches (see above *Identifying Relevant ESI*), can be used to find ESI that is needed to evaluate the merits of a case. It also can identify ESI that is responsive to a request for production or that may be withheld on the basis of the attorney-client privilege, work product protection or other grounds.

Parties commonly rely on keyword searches for linear document-by-document review and, in recent years, parties also have started using them in technology-assisted review (TAR) to select the seed sets that train the software to classify the remaining data. While keyword searches continue to be used, courts have observed their limitations and risks, and recognized that the proper selection and use of keywords involves certain technical knowledge (see, for example, *Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260 (D. Md. 2008)*).

Concept searches also can help identify important, responsive and privileged ESI that keywords may not find, by grouping conceptually similar and related ESI that do not contain the keywords. ESI grouped by concept searches may be sampled and categorized by a subject matter expert for importance, responsiveness or privilege, and that training set may be used to find similar documents and new keywords.

Search Document Requests: What to Expect in Response to an RFP for more on the use of keyword and concept searches during discovery.

EDA that employs test searches on a collection of ESI (for example, the e-mails and documents of key custodians) may be used to facilitate the parties' agreement on reasonable keywords. Test search results can be used to:

- Select the keywords that are finding the targeted ESI.
- Exclude the keywords that are finding untargeted ESI (false positives).
- Find new keywords to identify targeted ESI that the initial keywords did not find (false negatives).

In appropriate situations, the effectiveness of the keywords may be evaluated against a manually categorized ESI sample. Keywords then can be classified as:

- Good (hit on important, responsive or privileged ESI).
- Bad (hit on false positives, unimportant or non-responsive ESI).
- New (hit on false negatives or ESI tagged as important, responsive or privileged that was not hit on with the initial keywords).

ESI PRODUCTION

Requesting parties typically have limited knowledge about a producing party's ESI, so they often seek to expand discovery by broadening the pool of initial data custodians. Conversely, producing parties consistently seek to narrow the scope of discovery. Holding firm to these positions without cooperation is not in the parties' interests, especially because the court will often split the difference between the parties' requests, leaving neither party happy with the result.

Moreover, an overbroad selection of custodians frequently results in unreasonably cumulative and duplicative discovery, which is contrary to FRCP 26(b)(2)(C)(i) and the proposed amendment to FRCP 26(b)(1). Courts have limited the number of custodians subject to discovery where there is an absence of specific evidence that the additional custodians will have important, non-cumulative information (see, for example, *Kleen Prods. LLC v. Packaging Corp. of Am., No. 10-5711, 2012 WL 4498465, at *14-16 (N.D. Ill. Sept. 28, 2012)*).

Even if the information sought consists of evidence unfavorable to the producing party, the requesting party cannot reasonably expect to uncover every instance of relevant evidence (see *MBIA Ins. Corp. v. Credit Suisse Sec. (USA) LLC, No. 09-603751, 2014 WL 3543537, at *2 (Sup. Ct. N.Y. Co. July 17, 2014)*). Additionally, proportionality principles are being institutionalized in local rules and guidelines (see, for example, *Paul W. Grimm, US Dist. Court for the Dist. of Md., Discovery Order, ¶ 6(b) (Apr. 9, 2013)* (setting a presumptive limit on ESI discovery to ten key custodians)).

Information gathered from EDA may facilitate agreement between the parties on who the important custodians are and the relevant timeframe when the main events transpired, helping the parties set reasonable limits on e-discovery. Although parties generally agree that ESI from the period

ECA versus EDA

Although some e-discovery vendors use the terms interchangeably, ECA and EDA are different. ECA is a holistic case management approach designed to help counsel come to an informed and expedited decision about resolving a dispute. During ECA, counsel:

- Gather, assemble and analyze the facts and law.
- Assess the potential risks, costs and damages associated with litigation.
- Consider case logistics, such as venue and opposing counsel.

Search Case Assessment and Evaluation for more on effectively conducting a case assessment.

EDA, on the other hand, is a data management process designed to help counsel identify important and relevant ESI (and exclude unimportant and irrelevant materials) before the data is fully processed and reviewed. EDA involves an informed and expedited analysis of a company's ESI by searching, organizing and sampling the data.

immediately surrounding the key events should be searched, they often argue about the timeframe peripheral to the core date range. Front-loaded discovery may be avoided by focusing on when key custodians communicated about key events.

Sorting the key custodians' e-mails by date, sender, recipient, domain name, keywords and concepts may help identify:

- Other custodians with whom the key custodians communicated.
- Which other custodians communicated about topics of interest.
- Which other custodians prepared, sent or received key documents.
- Whether key custodians or other custodians used previously unidentified e-mail addresses, such as personal e-mail accounts.

EDA technologies that search ESI where it natively resides can identify relevant ESI at the time a litigation hold is implemented and help avoid over-preservation.

PRIVILEGE

The local rules and guidelines of many federal and state courts require parties to address the disclosure of withheld privileged information before the initial court conference. For example, Rule 11-b of the Commercial Division of the New York Supreme Court requires the parties to discuss excluding categories of information from privilege logs, as well as the use of category privilege logs (which identify categories of withheld documents) instead of traditional document-by-document logs (which list each privileged document individually) (see *Unif. Rules for N.Y. State Trial Courts, 22 NYCRR § 202.70, R. 11-b*).

Similarly, a pilot program in the US District Court for the Southern District of New York presumptively excludes certain categories of documents from privilege logs, including:

- Communications exclusively between a party and its trial counsel.
- Work product created by trial counsel after commencement of the action.
- Internal communications within a law firm, legal assistance organization, governmental law office, or legal department of a corporation or another organization.

(*Standing Order M10-468, In re Pilot Project Regarding Case Mgmt. Techniques for Complex Civil Cases in the S.D.N.Y., at 6 (S.D.N.Y. Nov. 1, 2011)*.)

Identifying categories of ESI to be excluded from or included in a privilege log can be difficult when ESI has not been reviewed. Early analysis of ESI may facilitate an agreement between the parties, before privilege review takes place, about the disclosure of withheld information on privilege logs, such as whether certain ESI may be excluded from the logs and whether the parties can use category logs.

More specifically, EDA can be used to organize ESI, so that the ESI can be sampled to confirm whether the documents fall within a defined category and, in the process, counsel can formulate a description for that category. The parties can then use this information to discuss what information to exclude from or include in the privilege log. For example:

- Clustering ESI by subject may help identify categories of communications that parties can agree to exclude from or include in the privilege log.
- Near-duplicate detection may help organize draft versions of similar documents into categories for privilege log entries.
- Threading may help group e-mail conversations to correspond to categories of entries on privilege logs.
- Searches combining keywords with sender, recipient and domain names may help find categories of potentially privileged communications.

Search Privilege Log for a sample privilege log that may be used in federal civil litigation, with explanatory notes and drafting tips.