



Top Ten Technical Questions from Litigators on Collecting, Reviewing, and Processing Electronic Data

November 2008

Stacy O'Neil Jackson, IE Discovery, Inc.

“It took man thousands of years to put words down on paper, and his lawyers still wish he wouldn't.” – Mignon McLaughlin

These days, the wish for clients to stop putting words on paper has been fulfilled...in a way. It's true that most clients no longer generate huge volumes of paper, but—as we all know—it's not because they've stopped capturing information. Instead, they've simply shifted the information to a new format: electronic data.

Despite the tedium inherent to sifting through box after box of paper, many attorneys yearn for those familiar days. That's understandable—the task of understanding electronic data management is overwhelming for most non-technical personnel.

Fortunately, formulating an efficient and effective process for collecting, reviewing, and processing electronic data isn't a shot in the dark. In fact, the technical questions at the forefront of most litigator's minds have straightforward answers that are proven by years of experience. The purpose of this article is to share those answers, so that you, too, understand the basic formula for a successful e-Discovery effort.

Question 1: Who should I include on my e-Discovery team?

E-Discovery requires the same approach as almost any other complex undertaking. You don't need to personally have the answer to every question—you just need to surround yourself with a group of people who collectively do.

A well-structured e-Discovery team includes representatives from three core departments: Legal, Information Technology (IT), and Records Management. The job of the legal experts is to thoroughly understand litigation details and, in turn, the types of documents and data that are relevant to the collection. The IT and Records Management teams should understand where those relevant documents and data live. Finally, the IT experts should know what tools and processes are required to accurately extract documents and data while remaining mindful of technical issues, such as preserving metadata.

Question 2: What technical issues are involved in maintaining the chain of custody?

Since electronic data is easily altered, an effective e-Discovery strategy must include the retention of accurate and comprehensive chain of custody records. A defensible approach to chain of custody records includes:

- **Keep data intact throughout processing.** Running virus scans and implementing write-protection on drives and files are two good ways to ensure that data remains unchanged.
- **Make a complete copy of all data.** Copies should capture all data on the drive, including ambient data (residual data from files that have been deleted but not erased) and metadata (data about data, such as the file creation date).
- **Use a reliable copying process.** The copying process must meet industry standards and be capable of independent analysis. Additionally, your copies should be tamper-proof—for example, you should save files to a limited-access, password-protected network location or to a write-protected hard drive.
- **Ensure every media item in the collection is physically secure.** Effective security processes include labeling items with the collection date, time, and source; storing items in physically secure locations that allow limited access; and performing forensic analysis only on working copies (not on the original file) whenever possible.
- **Document the entire process.** Perform an inventory at every step—from intake through final delivery—and keep an extensive chain of custody log that accounts for every file or storage container that you collect and process. The report should include the name of the original storage media, the total number of files contained in the media, the number of those files that were converted to image, the number of images rendered from the conversion, and a complete list of those files you did not convert along with an associated explanation.

Question 3: How do I design a sound collection strategy?

A well-designed collection effort is vital to an efficient and cost-effective e-Discovery strategy. Over-collection—in other words, collecting many more files than your case actually requires—may save some time and money upfront, but these savings are often offset by the high cost of reviewing irrelevant and duplicate data. On the other hand, the risk of under-collecting data is that you may have to re-collect as case issues change. Re-collection is usually much more expensive than gathering relevant data on the first attempt.

Follow these guidelines to strike the right balance:

- **Understand the types of data and documents that will comprise your collection.** For example, if your collection includes email, you'll need to know what email systems are involved, and the custodians and date ranges on which your collection should focus. If your collection includes electronic files, you'll need to know what types of files are likely to be relevant (e.g., PDFs, spreadsheets) and the folders and file paths required to access those files.
- **Create a comprehensive list of relevant search terms.** This list should include basic search terms and their common variations, such as alternative spellings and formats.
- **Know where relevant files are likely to exist.** Before your collection effort begins, you should understand the possible locations of relevant files. Will you be searching across a network? On local hard drives? On tape and backup drives? In custodians' desk drawers? Each of these locations will take a different collection approach.
- **Do not allow document custodians to forward email to you for collection purposes.** Doing so may cause spoliation or privilege challenges. Instead, ask custodians to upload relevant files directly to a centralized location with limited access or have them create a PST file.

Question 4: Once I've got a comprehensive collection, how should I deduplicate it?

Almost every large collection requires some level of automated deduplication. Otherwise, manual review becomes far too burdensome in terms of both time and money.

Usually, the decision of how to cull exact duplicates (documents with a 100% match between content, metadata, and format) is easy. Deduplicating software uses hash-based algorithms to reliably identify exact duplicates, which can often be segregated from production without manual review. More difficult is the decision of how to handle near duplicates—documents that are similar, but not exactly the same.

Discarding near duplicate documents without manual review is often a risky decision. For example, the 0.05% difference between two contracts might be the overall contract amount. Similarly, the slight difference between two emails might be the one sentence that gives rise to attorney-client privilege. When it comes to near duplicates, the best approach is usually to rely on automated processes to identify similar documents, and then manually review those documents to make a final determination on near-duplicate status.

Question 5: How do I develop a strategy for search technologies and processes?

The volume of electronic data involved in today's typical case is enormous, and continues to grow exponentially with every passing year. Simply put, it's no longer practical to rely exclusively on manual review to identify relevant documents. Instead, organizations must leverage automated search technologies—but not without careful consideration.

There are a variety of search techniques above and beyond simple keyword and Boolean search, including fuzzy logic, conceptual search, and text mining. Each of these techniques has associated benefits and risks, and it behooves the user to understand those issues prior to creating a search strategy. The most comprehensive resource for information on search technologies is *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*.

Question 6: Should I use an automated review process to cull documents?

As noted in Question 5, the massive volume of data involved in most of today's complex litigation necessitates the use of automated search and review techniques. Following are three reasons why the answer to, "Should I use automated review to cull documents?" is generally, "Yes."

- **Speed:** While the volume of data involved in today's typical case grows ever larger, production deadlines remain the same. Using automated tools to cull obviously irrelevant documents dramatically reduces time spent on manual review.
- **Cost:** Manual review is usually the most expensive aspect of discovery. In fact, the cost for a junior-level associate at many law firms is over \$200 per hour, often pushing the cost to review a single gigabyte (1GB) of data to over \$30,000.* It's easy to see how manually reviewing every document in a collection that exceeds a terabyte (1,024 gigabytes) of data—as many collections these days do—is almost always cost prohibitive.
- **Accuracy:** According to a 1985 study by David Blair and M.E. Maron, human review efforts are, on average, only about 20% accurate at identifying relevant documents in a large

* The Sedona Conference. (2007). The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery. *The Sedona Conference Journal (Volume 8, Fall 2007)*, 192.

collection.[†] Automated review processes, while certainly not perfect, tend to obtain significantly higher accuracy rates. Additionally, automated processes can be iteratively honed if initial accuracy rates are inadequate.

Question 7: Should I use an automated review process to assist with manual review?

While automated review has a number of benefits, it's generally not a replacement for manual review. Instead, the best approach is to use automated review to cull obviously irrelevant documents and identify term matches in relevant and privileged documents. Then, relevant documents can be conceptually grouped for a speedier and more accurate manual review.

A well-designed automated review will visually identify (i.e., highlight) term hits in a document, allowing the reviewer to immediately identify and focus on potentially relevant content, rather than spending time combing through every word in a document looking for a potential hit. Not only does term highlighting speed the manual review process, but it also improves accuracy, making it far less likely that a human reviewer will miss the one critical paragraph on page 48 of a 50 page document.

Likewise, conceptual grouping enables the reviewer to analyze documents with associated content in a logical sequence, so that patterns are more easily identifiable. Conversely, conceptual grouping also enables the reviewer to more quickly eliminate irrelevant documents, since they are also grouped together.

Question 8: What data should reviewers see during the review?

When it comes to review data, what your reviewers don't see can hurt you. At the same time, asking reviewers to analyze too much data can really slow the review process down. As with your deduplication strategy, you must be careful to strike the right balance.

Following is a list of data types that you should generally consider including in a review.

- **Family Relationships:** In this context, "family relationships" refer to associations between documents. For example, an email and its attachment comprise a document family, with the email considered the parent document, and the attachment considered the child. Often, the content of one document in a family will affect how other documents in the family are marked. Because it provides such critical information, there is generally no reason to exclude familial data from a document review.
- **Custodian/Source:** Clearly, understanding the source of a document is critical to both relevance and privilege reviews.
- **Match Terms:** If your review includes an automated component, make sure to visually highlight the relevance and privilege term matches within each document to improve both speed and accuracy.
- **Document Coding:** Examples of document coding data that can assist reviewer(s) include the document creation date, the title of the document, names associated with the document, and whether the coding was human or automated.

Question 9: What format should I use for review and production?

[†] Blair, David C. and M.E. Maron. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. *Communications of the ACM (Volume 28, Issue 3)*, 289 – 299.

To save both time and money, make your determinations about the format(s) in which you will review and produce documents as early in the process as possible. Also, to limit your legal exposure, use the same format(s) for review that you choose for production.

Following is a list of the review and production formats you may choose to employ, including the associated pros and cons:

- **Image:** A benefit of converting files to image is that it deletes metadata, which can sometimes include embarrassing or sensitive information. Additionally, image-only files are a universally viewable, and therefore highly efficient, format. If you decide to use this format, remember to secure agreement from opposing counsel on how each side will handle databases, large spreadsheets, and files that can't be processed—for example, executable (.exe) files and files that are unreadable due to viruses or damage.
- **Text:** Because it is universally viewable, text is a well-accepted format for production. Additionally, it can be time- and cost-efficient choice, since text files don't require you to perform OCR (optical character recognition) on the files produced by opposing counsel. The downside is that text can present redaction challenges.
- **Image/Text Combo:** When producing in image-only format, both sides need to perform OCR on the images to obtain searchable text. Unfortunately, OCR has a couple of significant drawbacks: it's very costly for high-volume productions, and the best OCR engines are only about 95% accurate. So, it may be in both parties' best interests to produce both image and text files to reduce processing costs for both sides. If you use this option, remember to pay close attention to redactions—if you redact an image and then hand over complete text, you've created a huge problem for yourself.
- **Native Files:** It's becoming increasingly common to produce files in their native format, because it's a cost-effective approach that also enables automated review. Native files do have a few disadvantages, though, including the risk of exposure from metadata and the requirement to have either the native application or a universal file viewer to open a document. Also, you can't redact native files.

Question 10: How can I measure the relative costs against the legal benefits of my technical choices?

Let's face it—legal discovery is expensive. Fortunately, though, you *can* control costs by making sound technical choices. To design a cost-effective discovery strategy:

- **Estimate the entire process.** Build realistic, scenario-based estimates around the options you're considering. Then, choose the most cost-effective options that are viable for your particular situation. For example, if you save on processing costs by reviewing and producing in native format instead of image and text, will that slow your reviewers down so much that your savings are obliterated by increased review costs?
- **Negotiate up front as much as possible.** At the meet and confer, you should—at the very least—define the project's scope, determine how you'll handle duplicates, agree on keyword lists, and identify the production format. By securing agreement with opposing counsel, you'll help prevent expenses from popping up unexpectedly.
- **Get started early.** Allocate as much time as possible to the discovery process to avoid rush fees from vendors, multi-shift staffing of legal reviewers, and possible legal sanctions.

While these ten topics don't address every possible technical question a litigator may have about collecting, processing, and reviewing electronic data, they are a good start to formulating a sound e-Discovery strategy.

If you follow the advice outlined in this article and you still find yourself floundering, the best step you can take is to start asking questions. Remember, that's why you created such a well-rounded e-Discovery team! There's a good chance that if you meet a serious challenge, someone—be it your IT lead, your Records Management expert, one of your fellow attorneys, or even an external e-Discovery provider—will have the answer you seek.