

Don't Take a Beating On Your Hit Rate

Use Effective Keyword Searches to Manage Discovery

By Diane Barrasso and Richard P. Rollo

Faced with ever-increasing litigation costs, in-house lawyers are searching for effective and legally defensible means of limiting the costs of electronic discovery. Many vendors advocate utilizing keyword searches to identify (and limit) the documents to be reviewed during discovery. If done correctly, the approach can reduce the total volume of documents reviewed by attorneys (the proverbial haystack) in search of those that are relevant to the litigation (the proverbial needles). However, utilizing keyword searches in this manner — if done incorrectly — has the potential to cost you a fortune.

Legal teams can effectively incorporate search techniques into their best practices by considering critical

issues before they review a single page. Doing so will only eliminate a major nightmare: excessive costs associated with over-collection and technical challenges that will require teams of project specialists to resolve.

PRE-DISCOVERY CONCERNS

The December 2006 amendments to the Federal Rules of Civil Procedure and the sheer volume of electronically stored (and discoverable) information have underscored the need to creatively craft and discuss document production protocols.

Negotiate

Notwithstanding the new federal rules, most aspects of eDiscovery are still negotiated informally among adversaries. Early on in pre-trial discovery conversations, you should propose a concise set of search terms to begin the dialogue. Keep your list short and clearly articulate the reasons for including or excluding particular terms. Be open minded to your adversary's proposals, but guard against overreaching — especially in situations where your adversary has little incentive to be reasonable (such as stockholder plaintiffs in derivative litigation involving unequal discovery cost). If you are unable to agree upon a mutually acceptable list of terms, you should consider submitting

competing proposals to the court *before* conducting document review.

Test Your List

Before agreeing on a list of search terms, run the terms through a preliminary document database, compute the number of hits (*i.e.*, documents returned) for each term, and evaluate the aggregate results. Search terms resulting in an inordinately high number of hits are likely flawed. During your pre-trial discovery conversations, consider sharing the foregoing analytics with your adversary to advance the dialogue.

Sample the Results

Employ a common sense approach to determine whether the search results, both hits and non-hits, are accurate. Randomly select a set of sample documents to analyze and verify the search results. If you find a disproportionate number of mis-hits (*i.e.*, responsive documents which are not a hit, or non-responsive documents that are a hit), adjust your methodology and retest.

Ensure that the samples account for the entire “family” of an e-mail chain (*i.e.*, the e-mail is the “parent” and its related attachments are “children”; collectively, they are referred to as a “family”) because it is rarely appropriate to break up the family for review and production (with the exception of withholding for privilege, in which case you would indi-

Diane Barrasso (dbarrasso@barrassoconsulting.com) is founder and president of Barrasso Consulting, a company specializing in the management, collection and review of documents in complex litigation. **Richard P. Rollo** is a counsel at Richards, Layton & Finger, P.A., in Wilmington, DE. The views expressed herein are those of the authors only, and not those of Barrasso Consulting, Richards, Layton & Finger, P.A., or its clients.

cate that a document is missing from the family).

GENERAL SEARCH GUIDANCE

There are some general rules of thumb that are common to all search-based approaches.

Monitor Syntax

Familiarize yourself with the search protocols unique to the search engine (and vendor) being utilized. Specifically, identify the appropriate wildcard symbols used to detect multiple variants of a word (e.g., "*" or "!") or near variants (e.g., fuzzy searching) and the proximity symbols (e.g., "near", "w/5", "w/s", etc.). To confirm accuracy of the syntax, screen the search terms through the vendor conducting the searches.

The analytics recorded in the sampling phase at the outset will reveal certain grammatical anomalies. If the results do not meet your expectations, probe the vendor and modify your search.

Generally Avoid .log and Extreme File Types

A hard drive or server is home to countless files that will contain one or more of your search terms, but have no relevant information. To ease the burden of this review, analyze file extensions, such as ".log," against the search terms to determine the likelihood of value. As each case is different, however, evaluate when unusual files should be considered and when they can be avoided.

This is another area for negotiation with your adversary. During your pre-trial discovery conversations, propose a list of file types to automatically exclude from the document review. Most vendors maintain such a list, which should provide a good starting point for your discussions.

Isolate Non-Searchable Files

Certain electronic files that might contain relevant information, such as images of text rather than text itself, cannot be searched yet need to

be reviewed. JPEGs, GIFs, older PDFs and many others fall into this category and should be isolated for special treatment.

Handle Embedded Files

Use an appropriate software tool that can fully search email attachments, contents within compressed material (such as .zip files) and embedded files.

Beware of Foreign Languages

Choose search tools according to the need for analysis of material in a foreign language. This may alter the syntax and complicate the hit-rate from usual file types. Crafting a native language search for foreign documents will expedite a complex aspect of the review as long as the tool can support the process and the languages are identified in advance.

Monitor Cut-off Date

Restrictions For Parent/Child Attachments

Searches need to anticipate families of e-mails and use the entire family as part of the overall document count. When restricting the production to a certain time period, for example, reviewers must consider the date of the entire family.

The primary date for consideration is usually the one listed on the final e-mail string. If that document is within the date range, it may pull documents outside of the date range. For these reasons, the legal team must make decisions on family handling before the process is finalized. With date restrictions, the conservative approach is to include an entire family if the parent or any child falls within the relevant range.

DOCUMENT HARVESTING AND COLLECTION

Harvesting data and preparing it for page-by-page attorney review takes on many forms these days — from collecting everything that isn't nailed down (entire hard drives, for instance) to more surgical collection approaches that even employ the use

of e-mail analytic tools. The goal should be to focus the collection to minimize the downstream processing costs, attorney review and document production.

To be clear, however, "harvesting and collection" is not the same as preservation (e.g., retaining out-of-cycle back up tapes, discontinuing any automatic deletion/destruction for relevant custodians, etc.). Preserving broadly remains the best (and safest) approach. How you choose to efficiently cull that data into a manageable (and defensible) subset of documents to be reviewed in discovery is a cost/benefit analysis to be conducted on a case by case basis after consulting with your legal team.

Again, there is no single blueprint; however, here are some approaches to consider when attempting to reduce your "preservation set" of documents to a manageable "review set" of documents.

Let Custodians Guide You

Most employees organize their electronic files into topical folders much like how they handle paper documents. These folders, assuming they are in a logical structure, can provide a roadmap for pinpointing relevant documents. As an alternative (or in addition) to negotiating search terms, some legal teams have successfully persuaded adversaries to use the folder structure as a means to identify (and limit) the documents to be reviewed during discovery. One such approach is to provide a printout of the directory structure on relevant computers and let your adversary identify the folders/documents to be examined during discovery. The approach is generally only successful in quid pro quo situations (e.g., two large corporations as adversaries) where both parties bear similar potential discovery costs.

Refine the Dataset

Use search terms to refine a dataset into the documents most

relevant to a particular matter. Desktop or server tools can further reduce the document pool that a vendor needs to process. They also identify those items that can be excluded immediately from review and production.

SEARCH-ASSISTED REVIEW

The power of searching can be put to good use even after harvesting and collection of documents is well-focused. Properly constructed searches help to locate trade secrets, privilege details, or otherwise protected information. Here are some extrapolations on the general and more obvious guidance for searching as specifically relevant to review.

E-mail Footers Can Mislead Keywords

In the current business environment of hypersensitivity to corporate governance, companies are increasingly placing disclaimers in the footer of each email message, which may contain the words "privilege," "legal," "confidential" and others that would have historically triggered a more detailed review. In universal application, however, these terms have little or no value in identifying truly sensitive or privileged message threads.

For this reason, reviewers must be attentive to the language that the company under scrutiny uses in its email messaging system. Again, sample the hit-rate of privilege search terms and documents that a test search produced.

To avoid mis-hit documents containing the word "legal," consider using related phrases such as "ask legal," "check with legal," or "legal near approv*." Excluding blanket footers from a search would be too risky since some of the truly privileged documents identified for collection will have these legends. Ideally, try to construct a search that would ignore rather than exclude the footers.

Broken E-mail Families Provide Context

You can potentially lose continuity in an email family if only one member of that group is the result of a search hit. That fragmentation can also distort the timeline, meaning and message of a particular chain of correspondence. Likewise, the trail of what family member(s) generated the hit(s) can help focus and streamline the review of the entire family.

Group Like Documents, Organize Conceptually

Assembling similar categories of documents may make content review easier. You are familiar with the terms, custodians and meanings. You can also determine the time frame and context more easily. Material can be aggregated by file path within a custodian, by subject line amongst custodians, or even by algorithm-based context clues, such as those provided by specialized review tools for content analysis and concept organization. This technique has the potential of accelerating reviewer analysis, particularly when used to search entire employee hard drives or mailboxes with excessive volumes of data.

An alternative approach is to organize electronic documents in chronological order before conducting your review. This approach works particularly well with emails. By examining events as they unfolded, it allows reviewers to comprehend the story line better.

Identify Personal Material

Irrelevant personal information often appears in a document collection and should be treated carefully. Well-organized individuals store personal documents of all sorts (income tax returns, mortgage applications, family messages, etc.) on their company computers. To minimize the chance of inadvertently producing such materials, you should conduct searches for key

terms related to potentially personal information (e.g., "joke"). In addition, consider examining the directory structure on relevant computers or email accounts to identify folders that appear to contain personal information. Once identified, these folders should receive a higher level of scrutiny.

CONCLUSION

While keyword searching has been responsible for streamlining modern discovery collection and review, the popularity of the technique has actually led to its downfall. The more familiar lawyers are with keywords, the more terms they want to search within a dataset. But, every time a search term fails to account for a prefix or a suffix, an abbreviation or local parlance, misspellings or other lexicographic variations, it produces more results. These results drive up the cost, delay production and frustrate the process.

Success comes with understanding that while keyword searches are not a magic bullet, the most effective cases are those in which short lists are negotiated in advance, tested and modified based on established best practices.

The result is that you will be searching for needles in smaller haystacks.

