Surefire Steps to Splendid Search
Craig Ball

Hear that rumble? It's the bench's mounting frustration with the senseless, slipshod way lawyers approach keyword search.

It started with Federal Magistrate Judge John Facciola's observation that keyword search entails a complicated interplay of sciences beyond a lawyer's ken. He said lawyers selecting search terms without expert guidance were truly going "where angels fear to tread."

Federal Magistrate Judge Paul Grimm called for "careful advance planning by persons qualified to design effective search methodology" and testing search methods for quality assurance. He added that, "the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented."

Most recently, Federal Magistrate Judge Andrew Peck issued a "wake up call to the Bar," excoriating counsel for proposing *thousands* of artless search terms.

> Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI. Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of 'false positives.' It is time that the Bar—even those lawyers who did not come of age in the computer era—understand this.

No Help

Despite the insight of Facciola, Grimm and Peck, lawyers still don't know what to <u>do</u> when it comes to effective, defensible keyword search. Attorneys aren't *trained* to carefully craft appropriate keywords or implement quality control testing for searching ESI. And their experience using Westlaw, Lexis or Google serves only to inspire false confidence in search prowess.

Even saying "hire an expert" is scant guidance. Who's an expert in ESI search for your case? A linguistics professor or litigation support vendor? Perhaps the misbegotten offspring of William Safire and Sergey Brin?

The most admired figure in e-discovery search today—*the Sultan of Search*—is Jason R. Baron at the National Archives and Records Administration, and Jason would be the first to admit he has no training in search. The persons most qualified to design effective search in e-discovery earned their stripes by spending thousands of hours running searches in real cases--making mistakes, starting over and tweaking the results to balance efficiency and accuracy.

The Step-by-Step of Smart Search

So, until the courts connect the dots or better guidance emerges, here's my step-by-step guide to craftsmanlike keyword search. I promise these ten steps will help you fashion more effective, efficient and defensible queries.

1. **Start with the request for production**
2. **Seek input from key players**
3. **Look at what You've Got and the Tools you'll Use**
4. **Communicate and Collaborate**
5. **Incorporate Misspellings, Variants and Synonyms**
6. **Filter and Deduplicate First**
7. **Test, test, test!**
8. **Review the hits**
9. **Tweak the queries and retest**
10. **Check the discards**

1. Start with the RFP

Your pursuit of ESI should begin at the first anticipation of litigation in support of the obligation to identify and preserve potentially relevant data. Starting on receipt of a request for production (RFP) is starting late. Still, it's against the background of the RFP that your production efforts will be judged, so the RFP warrants careful analysis to transform its often expansive and bewildering demands to a coherent search protocol.

The structure and wording of most RFPs are relics from a bygone time when information was stored on paper. You'll first need to hack through the haze, getting beyond the "any and all" and "touching or concerning" legalese. Try to rephrase the demands in everyday English to get closer to the terms most likely to appear in the ESI. Add terms of art from the RFP to your list of keyword candidates. Have several persons do the same, insuring you include multiple interpretations of the requests and obtain keywords from varying points of view.

If a request isn't clear or is hopelessly overbroad, push back promptly. Request a clarification, move for protection or specially except if your Rules permit same. Don't assume you can trot out some boilerplate objections and ignore the request. If you

can't make sense of it, or implement it in a reasonable way, tell the other side how you'll interpret the demand and approach the search for responsive material. Wherever possible, you want to be able to say, "We told you what we were doing, and you didn't object."

2. Seek input from key players

Judge Peck was particularly exercised by the parties' failure to elicit search assistance from the custodians of the data being searched. Custodians are THE subject matter experts on their own data. Proceeding without their input is foolish. Ask key players, "If you were looking for responsive information, how would you go about searching for it? What terms or names would likely appear in the messages we seek? What kinds of attachments? What distribution lists would have been used? What intervals and events are most significant or triggered discussion?" Invite custodians to show you examples of responsive items, and carefully observe how they go about conducting their search and what they offer. You may see them take steps they neglect to describe or discover a strain of responsive ESI you didn't know existed.

Emerging empirical evidence underscores the value of key player input. At the latest TREC Legal Track challenge, higher precision and recall seemed to closely correlate with the amount of time devoted to questioning persons who understood the documents and why they were relevant. The need to do so seems obvious, but lawyers routinely dive into search before dipping a toe into the pool of subject matter experts.

3. Look at what You've Got and the Tools You'll Use

Analyze the pertinent documentary and e-mail evidence you have. Unique phrases will turn up threads. Look for words and short phrases that tend to distinguish the communication as being about the topic at issue. What content, context, sender or recipients would prompt you to file the message or attachment in a responsive folder had it occurred in a paper document?

Knowing what you've got also means understanding the forms of ESI you must search. Textual content stored in TIFF images or facsimiles demands a different search technique than that used for e-mail container files or word processed documents.

You can't implement a sound search if you don't know the capabilities and limitations of your search tool. Don't rely on what a vendor tells you their tool can do, test it against actual data and evidence. Does it find the responsive data you already know to be there? If not, why not?

Any search tool must be able to handle the most common productivity formats, e.g., .doc, docx, .ppt, .pptx, .xls. .xlsx, and .pdf, thoroughly process the contents of common container files, e.g., .pst, .ost, .zip, and recurse through nested content and e-mail attachments.

As importantly, search tools need to clearly identify any "exceptional" files unable to be searched, such as non-standard file types or encrypted ESI. If you've done a good job collecting and preserving ESI, you should have a sense of the file types comprising the ESI under scrutiny. Be sure that you or your service provider analyzes the complement of file types and flags any that can't be searched. Unless you make it clear that certain files types won't be searched, the natural assumption will be that you thoroughly searched all types of ESI.

4. Communicate and Collaborate

Engaging in genuine, good faith collaboration is the most important step you can take to insure successful, defensible search. Cooperation with the other side is not a sign of weakness, and courts expect to see it in e-discovery. Treat cooperation as an opportunity to show competence and readiness, as well as to assess your opponent's mettle. What do you gain from wasting time and money on searches the other side didn't seek and can easily discredit? Won't you benefit from knowing if they have a clear sense of what they seek and how to find it?

Tell the other side the tools and terms you're considering and seek their input. They may balk or throw out hundreds of absurd suggestions, but there's a good chance they'll highlight something you overlooked, and that's one less do over or ground for sanctions. Don't position cooperation as a trap nor blindly commit to run all search terms proposed. "We'll run your terms if you agree to accept our protocol as sufficient" isn't fair and won't foster restraint. Instead, ask for targeted suggestions, and test them on representative data. Then, make expedited production of responsive data from the sample to let everyone see what's working and what's not.

Importantly, frame your approach to accommodate at least two rounds of keyword search and review, affording the other side a reasonable opportunity to review the first production before proposing additional searches. When an opponent knows they'll get a second dip at the well, they don't have to make Draconian demands.

5. Incorporate Misspellings, Variants and Synonyms

Did you know Google got its name because its founders couldn't spell googol? Whether due to typos, transposition, IM-speak, misuse of homophones or ignorance, ESI fairly crawls with misspellings that complicate keyword search. If you don't search for common spelling variants and errors, you'll overlook responsive items. "Management" will miss "managment" and "mangement."

To address this, you must either include common variants and errors in your list of keywords or employ a search tool that supports fuzzy searching. The former tends to be more efficient because fuzzy searching (also called *approximate string matching*)

mechanically varies letters by substitution, insertion, deletion and transposition, often producing an unacceptably high level of false hits.

While every word processor application flags misspelled terms, how do you convert keywords to their most common misspellings and variants? A linguist could help, or you might probe for propensities by having key custodians type keywords as they are read aloud. More likely, you'll turn to the web. Until an online tool emerges that lists common variants and predicts the likelihood of false hits, you might visit a site like www.dumbtionary.com that checks a keyword against more than 10,000 common misspellings or consult the Wikipedia list of over 4,000 common misspellings.

If you've ever played the board game *Taboo,* you know there are many ways to communicate an idea without using obvious word choices. Searches for "car" or" automobile" will miss documents about someone's "wheels" or "ride." You've got to consult the thesaurus, but don't go hog wild with Dr. Roget's list. Identify and include *likely* alternatives for critical keywords. Also, question key players about alternate terms, abbreviations or slang used internally to reference the same topics as the search terms.

6. Filter and Deduplicate First

Always filter out irrelevant file types and locations *before* initiating search. Music and images are unlikely to hold responsive text, yet they'll generate vast numbers of false hits because their content is stored as alphanumeric characters. The same issue arises when search tools fail to decode e-mail attachments before search. Here again, you have to know how your search tool handles encoded, embedded, multibyte and compressed content.

Filtering irrelevant file types is done various ways, including by de-NISTing for known hash values and culling by binary signatures, file extensions, paths, dates or sizes.

The exponential growth in the volume of information seen in e-discovery doesn't represent a leap in productivity so much as an explosion in duplication and distribution. Much of the data we encounter are the same documents, messages and attachments replicated across multiple backup intervals, devices and custodians. Accordingly, deduplicating repetitious content before indexing data for search or running keywords greatly aids search efficiency and reduces cost. Be sure any method of deduplication employed tracks the origins of suppressed iterations so that repopulation can be accomplished on a per custodian basis, if needed.

Applied sparingly and with care, you may even be able to use keywords to *exclude* irrelevant ESI. For example, the presence of keywords "Cialis" or "baby shower" in an

e-mail may reliably signal the message isn't responsive; but testing and sampling must be used to validate such exclusionary searches.

7. Test, test, test!

The single most important step you can take to assess keywords is to test search terms against representative data from the universe of machines and data under scrutiny. No matter how well you think you know the data or have refined your searches, testing will open your eyes to something unforeseen and likely save a lot of wasted time and money.

The nature and sample size of representative data will vary with each case. The goal in selection isn't to reflect the average employee's collection but to fairly mirror the collections of employees likely to hold responsive evidence. Don't select a custodian in marketing if the key players are in engineering. Often, the optimum choices will be obvious, especially when their roles made them a nexus for relevant communications. Custodians prone to retention of ESI are better candidates than those priding themselves on empty inboxes. The goal is to flush out problems *before* deploying searches across broader collections, so opting for uncomplicated samples lessens the values.

It's amazing how many false hits turn up in application help files and system logs; so early on, I like to test for noisy keywords by running searches against data having nothing whatsoever to do with the case or the parties (e.g., the contents of a new computer). Being able to show a large number of hits in wholly irrelevant collections is compelling justification for limiting or eliminating unsuitable keywords. Similarly, a company may want to test search terms against data samples collected from employees or business units having nothing to do with the events of concern to determine whether search terms are too generic to be of value.

8. Review the Hits

My practice when testing keywords is to generate spreadsheets letting me preview search hits in context; that is, flanked by perhaps 20-30 words on each side of the hit. It's very efficient and illuminating to scan a column of hits for searches gone awry while selecting particular documents for further scrutiny. Not all search tools support this ability, so check with your service provider to see what options they offer.

Armed with the results of your test runs, determine whether the keywords employed are hitting on a reasonably high incidence of potentially responsive documents. If not, what usages are throwing the search off? What file types are appearing on exceptions lists as unable to be searched due to e.g., obscure encoding, password protection or encryption?

As responsive documents are identified, review them for additional keywords, acronyms and misspellings

. Are terms that should be finding known responsive documents failing to achieve hits? Are there any consistent features in the documents with noise hits that would allow them to be excluded by modifying the query?

Effective search is an iterative process, and success depends on new insight from each pass. So, expect to spend considerable time assessing the results of your sample search. It's time wisely invested.

9. Tweak the queries and retest

As you review the sample searches, you're looking for ways you can tweak the queries to achieve better precision without adversely affecting recall. Do keyword pairs tend to cluster in responsive documents such that using a Boolean AND connector will reduce noise hits? Can you approximate the precise context you seek by controlling for proximity between terms?

If very short (e.g., three letter) acronyms or words are generating too many noise hits, you may improve performance by controlling for case (e.g., all caps) or searching for discrete occurrences (i.e., the term is flanked by spaces).

10. Check the discards

Keyword search must be judged both by what it finds and what it misses. That's the "quality assurance" courts demand. A defensible search protocol includes limited examination of the items *not* generating hits to assess whether relevant documents are being passed over. This examination of the discards will be more exacting for your representative sample searches as you seek to refine and gain confidence in your queries. Thereafter, random sampling should suffice.

No court has proposed a benchmark or rule-of-thumb for random sampling, and there's more science to sampling than simply checking every hundredth document. If your budget doesn't allow for expert statistical advice, and you can't reach a consensus with the other side, be prepared to articulate why your sampling method was chosen and why it strikes a fair balance between quality assurance and economy. The sampling method you employ needn't be foolproof, but it must be rational.

Remember that the purpose of sampling the discards is to promptly identify and resolve ineffective searches. If quality assurance examinations reveal that responsive documents are turning up in the discards, those failures must receive prompt attention.

Search Tips

Defensible search strategies are well documented. Be sure to record your efforts in composing, testing and tweaking search terms and the reasons for your choices along

the way. Spreadsheets are handy for tracking the evolution of your queries as you add, cut, test and tweak them.

Effective searches are *tailored to the data under scrutiny*. For example, it's silly to run a custodian's name or address against their own e-mail, but sensible for other collections. It's often smart to tier your ESI and employ keywords suited to each tier or, when feasible, limit searches to just those file types or segments of documents (i.e., message body and subject) likely to be responsive. This requires understanding what you're searching and how it's structured.

When searching e-mail for recipients, it's almost always better to search by-mail address than by name. In a company with dozens of Raj Patel's, each must have a unique e-mail address. Be sure to check whether users employ e-mail aliasing (assigning idiosyncratic "nicknames" to addressees) or distribution lists, as these can thwart search by e-mail address or name.

I guarantee these steps will improve your keyword searches but…

If you tell a court, "Craig Ball said to do it this way," you might hear, "Who?" or "Who cares?" Yet, until we know exactly what courts regard as sufficient and whose imprimatur matters, these techniques will help wring more quality and trim some of the fat from text retrieval. Don't forget, the least costly approaches to e-discovery are those done right from the start.