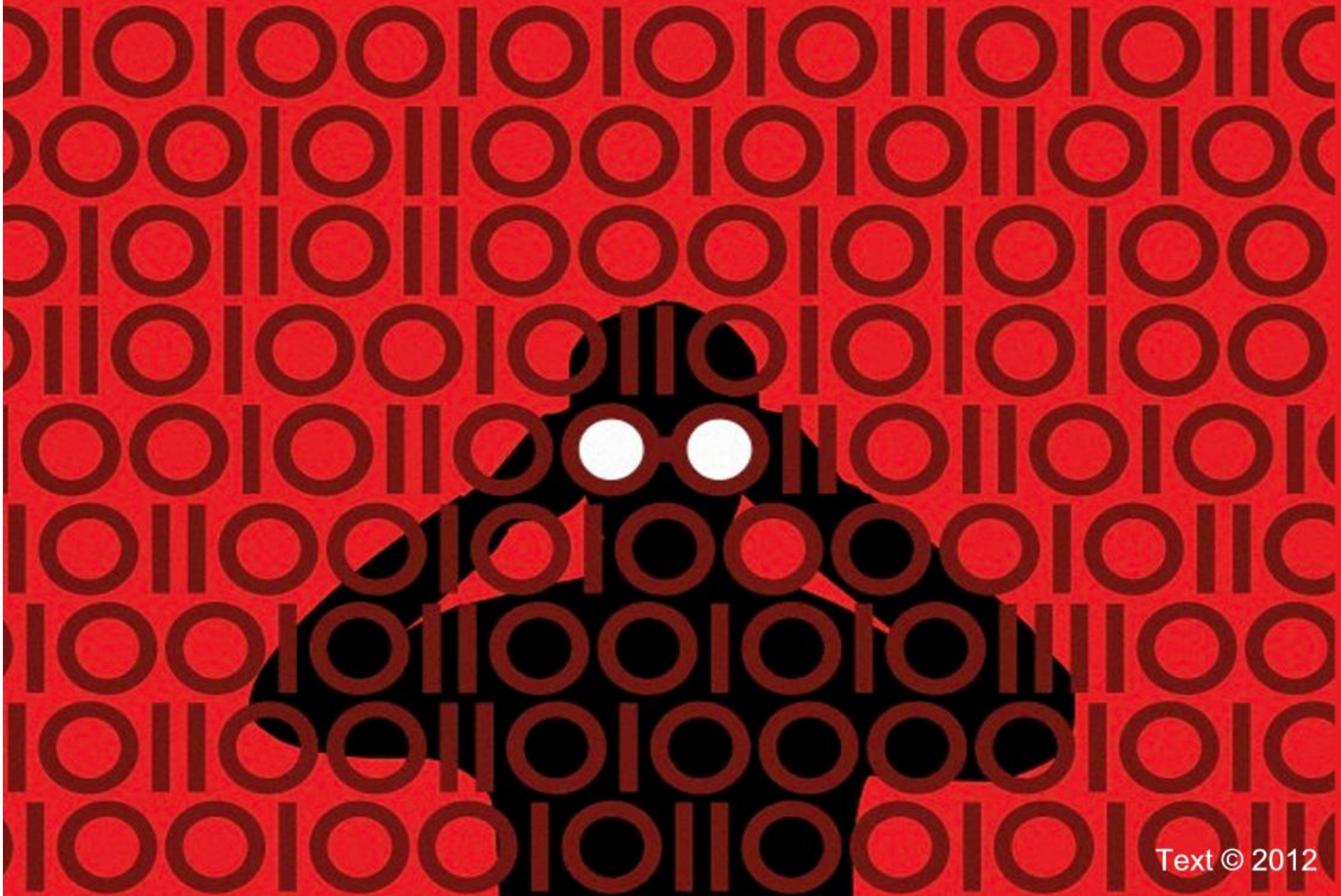
Selected Short Articles about SEARCH in Electronic Discovery

Craig Ball



Selected Short Articles about Search in Electronic Discovery

By Craig Ball
© 2012

Contents

About This Collection

I've been writing about search in electronic discovery for more than a decade, but only lately have the nuts-and-bolts of electronic search been topics of particular interest to litigators. It's beginning to dawn on some that much of what lawyers take for granted about search in e-discovery is a delusion. Spoiled by the extraordinary precision of tools like Google, Lexis and Westlaw, we've long assumed that searching our clients' electronically stored information is a task little different than searching the internet or case law.

But what we're finding is that search in e-discovery—whether by legions of reviewers or keyword search—is a lot harder than we thought. Language is complex and meaning, always elusive and nuanced, is a peculiar creature of context. As we've begun to measure search, we've seen cherished notions about its efficacy shattered by irrefutable facts and solid metrics. It turns out we've been doing a pretty lousy job of search, missing most relevant documents and barely succeeding at screening out the irrelevant and privileged ones. It's a revelation many lawyers and judges refuse to accept despite compelling evidence.

So, it’s an exciting and unsettling time for search in e-discovery. Even as lawyers come to accept keyword search as a fixture in discovery, the landscape of electronic search is shifting beneath our feet, bringing new terminology, tools and techniques. This collection of short articles—some old, most new--touches on how we’ll search tomorrow and how we can get the most out of our searches today. –**Craig Ball, 11/30/12**

About The Author

Craig Ball, of Austin is a Board Certified Texas trial lawyer, law professor (University of Texas) and accredited computer forensics expert who has dedicated his career to teaching the bench and bar about forensic technology and trial tactics. Craig hung up his trial lawyer spurs to till the soils of justice as a court-appointed special master and consultant in electronic evidence, as well as to teach and publish on computer forensics, emerging technologies, digital persuasion and electronic discovery. Fortunate to supervise, consult or serve as Special Master in some of the world's largest and most prominent electronic discovery matters, Craig greatly values his role as an instructor in computer forensics and electronic evidence to the Department of Justice and other law enforcement and security agencies. Mr. Ball also serves on the faculty of the Georgetown University Law School E-Discovery Academy and sits on the CCE Certification Board of the International Society of Computer Forensic Examiners.

Some articles reprinted from:

Available at www.craigball.com

MUSINGS ON TECHNOLOGY ASSISTED REVIEW

CRAIG BALL

Whether you call it “Predictive Coding” or Technology Assisted Review,” the time is nigh to leave much of the heavy lifting of review to machines trained to find responsive documents. These tools won’t be heuristic marvels like HAL-9000; but on the plus side, they probably won’t try to kill us.

The Streetlight Effect in Electronic Discovery

In the wee hours, a beat cop sees a drunken lawyer crawling around under a streetlight searching for something. The cop asks, “What’s this now?” The lawyer looks up and says, “I’ve lost my keys.” They both search for a while, until the cop asks, “Are you sure you lost them here?” “No, I lost them in the park,” the tipsy lawyer explains, “but the light’s better over here.”

I told that groaner in court, trying to explain why opposing counsel’s insistence that we blindly supply keywords to be run against the e-mail archive of a Fortune 50 insurance company wasn’t a reasonable or cost-effective approach e-discovery. The “Streetlight Effect,” described by David H. Freedman in his 2010 book *Wrong,* is a species of observational bias where people tend to look for things in the easiest ways. It neatly describes how lawyers approach electronic discovery. We look for responsive ESI only where and how it’s easiest, with little consideration of whether our approaches are calculated to find it.

Easy is wonderful when it works; but looking where it’s easy *when failure is assured* is something no sober-minded counsel should accept and no sensible judge should allow.

Consider *The Myth of the Enterprise Search*. Counsel within and without companies and lawyers on both sides of the docket believe that companies have the ability to run keyword searches against their myriad siloes of data: mail systems, archives, local drives, network shares, portable devices, removable media and databases. They imagine that finding responsive ESI hinges on the ability to incant magic keywords like Harry Potter. *Documentum Relevantus!*

Though data repositories may share common networks, they rarely share common search capabilities or syntax. Repositories that offer keyword search may not support Boolean constructs (queries using “AND,” “OR” and “NOT”), proximity searches (Word1 near Word2), stemming (finding “adjuster,” “adjusting,” “adjusted” and “adjustable”) or fielded searches (restricted to just addressees, subjects, dates or message bodies). Searching databases entails specialized query languages or user privileges. Moreover, different tools extract text and index such extractions in quite different ways, with the upshot being that a document found on one system will not be found on another using the same query.

But the Streetlight Effect is nowhere more insidious than when litigants use keyword searches against archives, e-mail collections and other sources of indexed ESI, That Fortune 50 company—call it All City Indemnity—collected a gargantuan volume of e-mail messages and attachments in a process called “message journaling.” Journaling copies every message traversing the system into an archive where the messages are indexed for search. Keyword searches only look at the index, not the messages or attachments; so, if you don’t find it in the index, you won’t find it at all.

All City gets sued every day. When a request for production arrives, they run keyword searches against their massive mail archive using a tool we’ll call *Truthiness*. Hundreds of big companies use *Truthiness* or software just like it, and blithely expect their systems will find all documents containing the keywords.

They’re wrong…or in denial.

If requesting parties don’t force opponents like All City to face facts, All City and its ilk will keep pretending their tools work better than they do, and requesting parties will keep getting incomplete productions. To force the epiphany, consider an interrogatory like this:

For each electronic system or index that will be searched to respond to discovery, please state:

a. **The rules employed by the system to tokenize data so as to make it searchable;**
b. **The stop words used when documents, communications or ESI were added to the system or index;**
c. **The number and nature of documents or communications in the system or index which are not searchable as a consequence of the system or index being unable to extract their full text or metadata; and**
d. **Any limitation in the system or index, or in the search syntax to be employed, tending to limit or impair the effectiveness of keyword, Boolean or proximity search in identifying documents or communications that a reasonable person would understand to be responsive to the search.**

A court will permit “discovery about discovery” like this when a party demonstrates why an inadequate index is a genuine problem. So, let’s explore the rationale behind each inquiry:

a. Tokenization Rules - When machines search collections of documents for keywords, they rarely search the documents for matches; instead, they consult an index of words extracted from the documents. Machines cannot read, so the characters in the documents are identified as “words” because their appearance meets certain rules in a process called “tokenization.” Tokenization rules aren’t uniform across systems or software. Many indices simply don’t index short words (e.g., acronyms). None index single letters or numbers.

Tokenization rules also govern such things as the handling of punctuated terms (as in a compound word like “wind-driven”), case (will a search for “roof” also find “Roof?”), diacriticals (will a search for Rene also find René?) and numbers (will a search for “Clause 4.3” work?). Most people simply *assume* these searches will work. Yet, in many search tools and archives, they don’t work as expected, or don’t work at all, unless steps are taken to ensure that they will work.

b. Stop Words – Some common “stop words” or “noise words” are simply excluded from an index when it’s compiled. Searches for stop words fail because the words never appear in the index. Stop words aren’t always trivial omissions. For example, “all” and “city” were stop words; so, a search for “All City” will fail to turn up documents containing the company’s own name! Words like side, down, part, problem, necessary, general, goods, needing, opening, possible, well, years and state are examples of common stop words. Computer systems typically employ dozens or hundreds of stop words when they compile indices.

Because users aren’t warned that searches containing stop words fail, they mistakenly assume that there are no responsive documents when there may be thousands. A search for “All City” would miss *millions* of documents at All City Indemnity (though it’s folly to search a company’s files for the company’s name).

c. Non-searchable Documents - A great many documents are not amenable to text search without special handling. Common examples of non-searchable documents are faxes and scans, as well as TIFF images and some Adobe PDF documents. While no system will be flawless in this regard, it’s important to determine *how much* of a collection isn’t text searchable, *what’s* not searchable and whether the portions of the collection that aren’t searchable are of *particular importance* to the case. If All City’s adjusters attached scanned receipts and bids to e-mail messages, the attachments aren’t keyword searchable absent optical character recognition (OCR).

Other documents may be inherently text searchable but not made a part of the index because they’re password protected (i.e., encrypted) or otherwise encoded or compressed in ways that frustrate indexing of their contents. Important documents are often password protected.

d. Other Limitations - If a party or counsel knows that the systems or searches used in e-discovery will fail to perform as expected, they should be obliged to affirmatively disclose such shortcomings. If a party or counsel is uncertain whether systems or searches work as expected, they should be obliged to find out by, e.g., running tests to be reasonably certain.

No system is perfect, and perfect isn't the e-discovery standard. Often, we must adapt to the limitations of systems or software. But you have to know what a system *can't* do before you can find ways to work around its limitations or set expectations consistent with actual capabilities, not magical thinking and unfounded expectations.

Are They Trying to Screw Me?

The title is the question posed by a plaintiffs' lawyer who called because he didn't know what to make of a proposal from opposing counsel. The lawyer explained that he'd attended a Rule 26(f) "Meet 'n Confer" where he'd tried to manifest the right grunts and signs to convey that he wanted electronically-searchable production. As neither of the lawyers conferring knew how that might achieve such a miracle, they shared a deer-in-headlights moment, followed by the usual "let me ask my client and get back to you" feint. Some years back, I defined a Rule 26(f) conference as "*Two lawyers who don't trust each other meeting to discuss matters neither understands*." That definition seems to have withstood the test of time.

Before my high-handed cynicism turns you off completely, let me explain that I appreciate that many fine lawyers didn't grow up with this "computer stuff." They earned their stripes with paper and, like me, leapt to law from the liberal arts. They're crazy busy with the constant demands of a trial practice, and ESI is just not a topic that excites their interest. Some are still recovering from the *last* time they tried to pick up pointers from a tech-savvy person and nearly drowned in a sea of acronyms and geek speak.

I feel your pain. I do. Now, let's ease the pain:

The other side proposed:

Documents will be produced as single page TIFF files with multi-page extracted text or OCR. We will furnish delimited IPRO or Opticon load files and will later identify fielded information we plan to exchange.

Are they trying to screw you? Probably not.
Are you screwing yourself by accepting the proposed form of production? Yes, probably.

First, let's translate what they said to plain English.

"Documents will be produced as single page TIFF files…."

They are not offering you the evidence in anything like the form in which they created and used the evidence. Instead, they propose to print everything to a kind of electronic paper, turning searchable, metadata-rich evidence into non-searchable pictures of much (but not all) of the source document. These pictures are called TIFFs, an acronym for Tagged Image File Format. "Single page TIFF" means that each page of a document will occupy its own TIFF image, so reading the document will require loading and reviewing multiple images in order (as compared to, *e.g.*, a PDF where the custom is for the entire document to be contained within one multipage image).

If the document they are producing is something they hold and use in its “native” electronic format, turning it into a TIFF is like lobotomizing the document. If you ever pithed a frog in high school biology, you know what it’s like to TIFF a native document. By “native,” I mean that the file that contains the document is in the same electronic format as it was when used by the software application that created or used the file. For example, the native form of Microsoft Word document is typically a file with the extension .DOC or .DOCX. For a Microsoft Excel spreadsheet, it’s a file with the extension .XLS or .XLSX. For PowerPoints, the file extensions are .PPT or .PPTX. Native file formats contain the full complement of content and application metadata available to those who created and used the document. Unlike TIFF images, native files are *functional* files, in that they can be loaded into a copy of the software application that created them to replicate what a prior user saw, as well as affording a comparable ability to manipulate the data and access content that’s made inaccessible when presented in a non-native formats.

Think of a TIFF as a PDF’s retarded little brother. I mean no offense by that, but TIFFs are not just differently abled; they are severely handicapped. Not born that way, but lamed and maimed on purpose. The version the other side retains works. They downgrade what they give you, making it harder to use and stripping it of potentially-probative content.

Make no mistake: what you lose isn’t just some hyper technical minutiae that they will label “metadata” and dismiss as irrelevant. I’m talking about <u>content</u>—what people wrote to each other *about* the document *within* parts of the document they won’t let you see. Imagine Post-It notes and marginalia on a paper document. In the era of paper, it’s like taking those notes off and throwing them away when they make copies. It’s like adjusting the photocopier settings so as to conceal what’s in the margins. We knew that was wrong with paper documents, so why can’t we see it’s just as wrong when done to electronic documents.

Do they do this because they are trying to screw you? Probably not.
Does it screw you just the same? Well, yeah.

“[W]ith multi-page extracted text or OCR.”

A native file isn’t just a picture of part of the evidence. *It’s the original electronic evidence*. As such, it contains <u>all</u> of the content of the document in an electronic form. Because it’s designed to be electronically usable, it tends to also be inherently electronically searchable; that is, whatever data it holds is encoded into the native electronic file, including certain data *about* the data, called **application metadata.** When an electronic document is converted to an image—TIFF—it loses its ability to be searched electronically and its application metadata and utility is lost. It's like photographing a steak. You can *see* it, but you can't smell, taste or touch it. You can't hear the sizzle, and you surely can't eat it.

Because converting to TIFF takes so much away, parties producing TIFF images employ cumbersome techniques to restore some of the lost functionality and metadata. To restore some electronic searchability, they extract some of the text from all the pages of the electronic document and supply it in a file sent along the TIFF images. It's called "multi-page extracted text" because, although the single-page TIFFs capture an image of each page, the text extraction spans all of the pages in the document. A recipient runs searches against the extracted text file and then seeks to correlate the hits in the text to the corresponding page image.

Note that I say, "they extract some of the text." They also leave behind some of the text. Here's where it's just dirty pool. Proponents of TIFF productions *tell* courts and opponents that they are furnishing the full textual contents in other ways, but they almost never do. They often don't give it to their own lawyers--who might be ethically bound to produce it, if they ever saw it.

If the source documents are scans of paper document, they have no electronic text to extract. Instead, the scans are subjected to a process called optical character recognition (OCR) that serves to pair the images of letters with their electronic counterparts and impart a rough measure of searchability.

"We will furnish delimited IPRO or Opticon load files…."

Whether extracted from an electronic source or cobbled together by OCR, the text corresponding to the images or scans is transferred in a so-called "load files" that may also contain metadata collected about the source documents. Collectively, the load file(s) and document images are correlated in a database tool called a "review platform" that facilitates searching the text and viewing the corresponding image. To insure that the images properly match up with extracted text and metadata, the data in the load files is "delimited," meaning that each item of information corresponding to each page image is furnished in a sequence separated by delimiters--just a fancy word for characters (like commas, tabs or semicolons) used to separate each item in the sequence. The delimiting scheme employed in the load files can follow any of several published standards for load file layout, including the most common schemes known as IPRO or Opticon.

"[A]nd will later identify fielded information we plan to exchange."

Much of the information in electronic records is fielded, meaning that is not lumped together with all the other parts of the record but is afforded its own place or space. When we fill out paper forms that include separate blanks for our first and last name, we are dividing data (our name) into fields: (first), (last). A wide array of information in and around electronic files tends to be stored as fields, in the manner in which, e.g., e-mail messages separately field information like From, To, Date and Subject. If fielded information is exchanged in discovery as fielded information, you lose the ability to filter information by, for example, Date or Sender in the case of an e-mail message or by a host of properties and metadata describing other forms of electronically stored information.

Additionally, the discovery process may necessitate the melding of various fields of information to electronic documents, such as Bates numbers, document file paths and custodians or associated TIFF image numbers. There may be hundreds of fields of metadata and other data from which to select, though not all of it has any evidentiary significance or practical utility. Accordingly, the proposal defers the identification of fielded information to be exchanged until later in the discovery process when, presumably, the parties will have a better idea what types of ESI are implicated and what complement of fields will prove useful or relevant.

Are they trying to screw you by not identifying fielded information? No. They're just buying time
Does their delay screw you? Maybe. Going back to re-collect fielded information you didn't know your opponent would seek can be burdensome and costly. Waiting too long to seek fielded information may prompt your opponent to refuse to collect and produce it.

So, are they trying to screw you by this proposal? I doubt it. Chances are they are giving you the dumbed down data because that's what they *always* give the other side, most of whom accept it, neither knowing nor caring what they're missing. It's probably the form of production their own lawyers prefer because they're reluctant to invest in modern review tools. It doesn't hurt that the old ways take longer and throw off more billable hours. Finally, no producing party is losing sleep over the stripped-away content. It's too candid, too honest, too likely to be a place where people reveal too much of what they're really thinking.

You may accept the screwed up proposal because, even if the data is less useful and incomplete, you don't have to evolve. You'll pull the TIFF images into your browser and read them one-by-one, just like good ol' paper, telling yourself that what you didn't get probably wasn't important and promising yourself that you'll get the good stuff—the native stuff--next time.

Surefire Steps to Splendid Search

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."

More 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 insights 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 craft keyword searches of ESI or implement quality control testing for same. 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?

Perhaps the most influential 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 Request for Production

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 backdrop 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 providers analyze 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, electronically stored information fairly crawls with misspellings that complicate keyword search. Merely searching for "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, often producing an unacceptably high level of false hits.

How do you convert keywords to their most common misspellings and variants? A linguist could help or you can turn to the web. Until a tool emerges that lists common variants and predicts the likelihood of false hits, try a site like **http://www.dumbtionary.com** that checks keywords against over 10,000 common misspellings and consult Wikipedia's list of more than 4,000 common misspellings (Wikipedia shortcut: **WP:LCM**).

To identify synonyms, pretend you are playing the board game Taboo. Searches for "car" or" automobile" will miss documents about someone's "wheels" or "ride." Consult the thesaurus for likely alternatives for critical keywords, but don't go hog wild with Dr. Roget's list. Question key players about internal use of alternate terms, abbreviations or slang

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 can be accomplished various ways, including culling by binary signatures, file extensions, paths, dates or sizes and by de-NISTing for known hash values. The National Institute of Standards and Technology maintains a registry of hash values for commercial software and operating system files that can be used to reliably exclude known, benign files from e-discovery collections prior to search. **http://www.nsrl.nist.gov**.

The exponential growth in the volume of ESI 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, the efficiency of search is greatly aided—and the cost greatly reduced—by *deduplicating* repetitious content *before* indexing data for search or running keywords. Employ a method of deduplication that tracks the origins of suppressed iterations so that repopulation can be accomplished on a per custodian basis.

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 the 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 custodial 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 value.

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, test search terms against data samples collected from employees or business units having nothing to do with the subject events to determine whether search terms are too generic.

8. Review the Hits

My practice when testing keywords is to generate spreadsheet-style views letting me preview search hits in context, that is, flanked by 20 to 30 words on each side of the hit. It's efficient and illuminating to scan a column of hits, pinpoint searches gone awry and select 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 unsearchable 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, look 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 only by spaces or punctuation).

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 <u>not</u> generating hits to assess whether relevant documents are being passed over.

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, but 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. 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 modify them.

Effective searches are tailored to the data under scrutiny. For example, it's silly to run a custodian's name or e-mail address against his or her 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, to 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 e-mail address than by name. In a company with dozens of Bob Browns, 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.

Search is a Science…

…but one lawyers *can* master. I guarantee these steps will wring more quality and trim the fat from text retrieval. *It's worth the trouble*, because the lowest cost e-discovery effort is the one done right from the start.

Gold Standard

Lawyers are in denial to the point of delusion with respect to the reliability of keyword search and human review. Judge John Facciola put it best when he quipped that lawyers think they're experts at keyword search because they once found a Chinese restaurant on Google.

We trust keyword search because we understand it. We trust manual review of documents because we grossly overestimate reviewers' abilities to make sound, consistent decisions about relevance. "To err is human," the Bar seems to say, "but forgive us if we'd rather not divine just how error-prone reviewers really are."

Better approaches to search are arriving as so-called "predictive coding" or "technology assisted review" (TAR) products. Still, it will be years before the rank and file embraces TAR, if only because those hawking TAR tools remain resolutely uninterested in positioning the technology for use by anyone but big corporations and white shoe law firms. Worse, the fervor among vendors to sell something, *anything* they can label predictive coding insures that tools little different from ordinary keyword search will be given a dab of lipstick and pushed out to market as TAR tools. It's messy down in the TAR pit.

Even those adopting predictive coding tools will need to compile "seed sets" of relevant documents to train their tools. So, clunky-but-comfy keyword search and manual review are likely to remain the means to cull seed sets from samples. Despite serious shortcomings, keyword search and manual review will be with us for a while.

Keyword search is the art of finding documents containing words and phrases that signal relevance followed by page-by-page (linear) review of those documents. It's often called the "gold standard" of electronic discovery.

That's ironic, because extracting and refining gold relies less on finding precious aurum than it does on dispersing all that isn't golden. Prospectors use water and chemicals to flush away all but the gold left behind. So, a true "gold standard" for keyword search would incorporate both precise inclusion (smart queries) and defensible exclusion (smart filters).

To illustrate, in one e-discovery dispute over search, the plaintiff submitted keywords to be run against the defendant's e-mail archive for a three-month interval. Unfortunately, the archive held all e-mail for all custodians, and the defendant adamantly refused to segregate by key custodian or deduplicate before running searches. The interval was narrow, but the collection was vast and redundant.

The defendant tested the agreed-upon keywords but shared only aggregate hit rates for each. Thinking the numbers too high, but unwilling to look at the hits in context, the defendant rejected the search terms. The plaintiff agreed the hit counts were daunting but asked to see examples of hits on irrelevant documents before furnishing exclusionary (AND NOT) modifications to flush away more of what wasn't golden.

The defendant refused, insisting it wasn’t necessary to see the noise hits in context to generate more precise queries. The parties were at an impasse, with one side grousing “too many hits” and demanding different search terms and the other side uncertain how to exclude irrelevant documents without knowing what caused the noisy results.

A lawyer who dismisses a search because it yields “too many hits” is as astute as the Emperor Joseph dismissing Mozart’s *Il Seraglio* as an opera with “too many notes.” Mozart replied, “There are just as many notes as there should be." Indeed, if data is properly processed to be susceptible to text search and the search tool performs appropriately, a keyword search generates just as many hits as there should be. Of course, few lawyers craft queries with the precision Mozart brought to music; so when the terms used seem well chosen for relevance, it’s crucial to scrutinize the results to learn what tailings are cropping up with the gilt-edged, relevant documents.

Keyword search is just a crude screen: “Show me items that contain these words, and don’t show me items that contain those.” High hit counts don’t always signal a bad screen. If search terms merely divide the collection into one pile holding relevant documents and one without, you’re closer to striking gold. Then, you look at what you can reliably exclude with the next screen, and the next; drawing ever closer to that elusive quarry, *documentum relevantus*.

But you must see hits in context to refine queries by exclusion. That seems so manifestly obvious, it’s astounding how often it’s not done.

When lawyers delegate keyword search, they often get back only aggregate hit counts and mistakenly conclude that’s enough information to judge searches noisy or not. If, instead, counsel get their hands dirty with the data, as by personally exploring representative samples using desktop or hosted tools, the parties could work quickly, effectively and cooperatively to zero in on relevant material. Good queries are best refined by knowledgeable people testing them against pertinent, small collections. Lousy outcomes spring from lawyers thinking up magic words and running them against everything.

It’s not just a theory. Recently, as part of an early case assessment effort, I sought to rapidly isolate relevant documents from a half million e-mail items culled from four key custodians. That’s a volume where you’d expect to see bids from service providers and mustering of review teams. It’s a project most firms would see as much more than a weekend’s work for one lawyer.

We tried something different. To start, the client exported the four key custodians’ e-mail messages for the time period of interest from its e-mail archives. Those 50 gigabytes of messaging went into a desktop processing and review tool.

Extracting and indexing the data overnight, I flagged exception items (e.g., images without extractable text and encrypted files) for further processing, then exported spreadsheets reflecting the most used e-mail addresses. I asked the custodians to flag addresses with no connection to the dispute. Meanwhile, I compiled the customary list of search terms and phrases expected to occur in relevant documents and tested these.

Documents with false hits were examined for characteristics permitting mechanical exclusion. Testing, re-testing and re-examination soon produced reliable inclusion and exclusion term lists. Weeks of evaluation took just days because the iterations and results were instantaneous.

The discards were tested, too. For example, material excluded by addresses but containing inclusion terms was carefully checked to insure the hits weren't relevant. Defensible exclusion proved as powerful as inclusion, and potentially relevant material that couldn't be excluded as tailings stayed in the collection as ore. A true "gold standard."

Did it produce a perfectly parsed set of material? Certainly not. Keyword search and human review still fall short of expectations. But it was fast, relatively cheap and afforded cautious confidence that the set produced was more relevant and less riddled with junk than what would have emerged from the usual game of blind man's buff. It was fast and cheap because the person creating and testing the inclusive and exclusive filters was elbows deep in the data and hands on with the search tool. Feedback was immediate. Quality checks could be done at once.

Ideally, e-discovery tools don't put distance between the lawyer and the evidence but, instead, extend our reach and help us get our arms around big data. A lawyer who is hands-on with the evidence and who tests and refines his or her choices is a lawyer who can explain and defend those choices. That's the real golden future of e-discovery. Welcome back, counselor.

Imagining the Evidence

As a young lawyer in Houston, I had the good fortune to sip whiskey with veteran trial attorneys who never ran short of stories. One told of the country lawyer who journeyed to the big city to argue before the court of appeals. The case was going well until a judge asked, "Counsel, are you aware of the maxim, *'volenti non fit injuria?'"*

"Why, Your Honor," he answered in a voice as smooth as melted butter, "In the piney woods of East Texas, we speak of little else."

Lately, in the piney woods of e-discovery, the topic is technology-assisted review (TAR *aka* predictive coding), and we speak of little else. The talk centers on that sudsy soap opera, *Da Silva Moore v. Publicis Groupe, and* whether Magistrate Judge Andrew Peck of the Southern District of New York will be the first judge to anoint TAR as being "court approved" and a suitable replacement for manual processes now employed to segregate ESI.

TAR is the use of computers to identify responsive or privileged documents by sophisticated comparison of a host of features shared by the documents. It's characterized by methods whereby the computer trains itself to segregate responsive material through examination of the data under scrutiny or is trained using exemplar documents ("seed sets") and/or by interrogating knowledgeable human reviewers as to the responsiveness or non-responsiveness of items sampled from the document population.

Let's put this "court approved" notion in perspective. Dunking witches was court approved and doubtlessly engendered significant cost savings. Trial by fire was also court approved and supported by precise metrics (*"M'Lord, guilt is established in that the accused walked nine feet over red-hot ploughshares and his incinerated soles festered within three days").* Whether a court smiles on a methodology may not be the best way to conclude it's the better mousetrap. Keyword search and linear review enjoy *de facto* court approval; yet both are deeply flawed and brutally inefficient.

The imprimatur that matters most is "opponent approved." Motion practice and false starts are expensive. The most cost-effective method is one the other side accepts without a fight, i.e., the least expensive method that affords opponents superior confidence that responsive and non-privileged material will be identified and produced.

Don't confuse that with an obligation to kowtow to the opposition simply to avoid conflict. The scenario I'm describing is a true win-win:

- Producing parties have an incentive to embrace TAR because, when it works, TAR attenuates the most ***expensive*** component of e-discovery: *attorney search and review*.
- Requesting parties have an incentive to embrace TAR because, when it works, TAR attenuates the most ***obstructive*** component of e-discovery: *attorney search and review*.

Producing parties don't just obstruct discovery by the rare and reprehensible act of intentionally suppressing probative evidence. It occurs more often with a pure heart and empty head as a consequence of lawyers using approaches to search and review that miss more responsive material than they find.

It's something of a miracle that documentary discovery works at all. Discovery charges those who reject the theory and merits of a claim to identify supporting evidence. More, it assigns responsibility to find and turn over damaging information to those damaged, trusting they won't rationalize that incriminating material must have had some benign, non-responsive character and so need not be produced. Discovery, in short, is anathema to human nature.

A well-trained machine doesn't care who wins, and its "mind" doesn't wander, worrying about whether it's on track for partnership. From the standpoint of a requesting party, an alternative that is both objective and more effective in identifying relevant documents is a great leap forward in fostering the integrity and efficacy of e-discovery. Crucially, a requesting party is more likely to accept the genuine absence of supportive ESI if the requesting party had a meaningful hand in training the machine.

Until now, the requesting party's role in "training" an opponent's machines has been limited to proffering keywords or Boolean queries. The results have been uniformly awful.

But the emerging ability to train machines to "find more documents like this one" will revolutionize requests for production in e-discovery. Because we can train the tools to find similar ESI using *any* documents, we won't be relegated to using seed sets derived from *actual* documents. We can train the tools with contrived examples–fabrications of documents like the genuine counterparts we hope to find.

I call this "imagining the evidence," and it's not nearly as crazy as it sounds.

If courts permit the submission of keywords to locate documents, why not entire documents to more precisely and efficiently locate other documents? Instead of demanding "any and all documents touching or concerning" some amorphous litany of topics, we will serve a sheaf of dreams—freely forged smoking guns—and direct, "show me more like these."

Predictive coding is not as linguistically fussy as keyword search. If an opponent submits contrived examples of the sorts of documents they seek, it's far more likely a similar document will surface than if keywords alone were used. As importantly, it's less likely that a responsive document will be lost in a blizzard of false hits. This allows us to rely less on our opponents to artfully construct queries. Instead, we need only trust them to produce the non-privileged, responsive results the machine finds.

There's more to documents that just the words they contain, so mocking up contrived exemplars entails more than fashioning a well-turned phrase. Effective exemplars will employ contrived letterheads and realistic structure, dates and distribution lists to insure that all useful contextual indicia are present. And, of course, care must be taken and processes employed to ensure that no contrived exemplars are mistaken for genuine evidence.

The use of contrived examples may ruffle some feathers. I can almost hear a chorus of, "How dare they draft such a vile thing!" But the methodology is sound, and how we will go about "imagining the evidence" is likely to be a topic of discussion in the negotiation of search protocols once use of technology assisted review is commonplace.

Another "not as nutty as it sounds" change in discovery practice wrought by TAR will be affording requesting parties a role in training TAR systems. The requesting party's counsel would be presented with candidate documents from the collection that the machine has identified as potentially responsive. The requester will then decide whether the sample is or is not responsive, helping the machine hone its capacity to find what the requester seeks. After all, the party seeking the evidence is better situated to teach the machine how to discriminate.

For this to work, the samples must first be vetted by the responding party's counsel for privilege and privacy concerns, and the requesting party must be willing to undertake the effort without fretting about revealing privileged mental impressions. It's going to take some getting used to; but the reward will be productions that cost less and that requesting parties trust more.

*Volenti non fit injuria means "*to a willing person, injury is not done." When we fail to embrace demonstrably better ways of searching and reviewing ESI, we assume the risk that probative evidence won't see the light of day and voluntarily pay too high a price for e-discovery.

Agatha, Hercule, Mummy and Me

Three weeks ago, skulking around the mummies in a small-but-fine museum on the University of Sydney campus, I learnt that mystery writer Agatha Christie was married to archaeologist, Max Mallowan, and that she'd assisted him in Syrian digs. Dame Agatha even used her cold cream and knitting needles to clean rare ivory artifacts. The experience found its way into her work. An exhibit of Christie-cleaned carvings included a quote from the author's fictional detective, Hercule Poirot, in Death on the Nile (1937):

> Once I went professionally to an archaeological expedition–and I learnt something there. In the course of an excavation, when something comes up out of the ground, everything is cleared away very carefully all around it. You take away the loose earth, and you scrape here and there with a knife until finally your object is there, all alone, ready to be drawn and photographed with no extraneous matter confusing it. That is what I have been seeking to do–clear away the extraneous matter so that we can see the truth–the naked shining truth.

This naturally got me thinking about the way we approach search in electronic discovery. Most lawyers use keywords to find documents responsive to discovery despite their propensity to sweep up too much chaff. We get lots of the documents we seek with keywords; unfortunately, the results come caked with the loose earth of documents containing keywords but having no connection to the case. Testing confirms this occurs with a ratio of about 20% responsive matter to 80% extraneous. That's a lot of loose earth!

The current industry practice is for keyword-culled documents to undergo horrifically expensive brute force review, i.e., bored lawyers reading each page. Such spirit crushing linear review accounts for anywhere from 50-90% of the total cost of e-discovery; consequently, when you reduce lawyer review time, you slash the biggest contributor to cost…and waste. If most of the material culled by keyword search is extraneous matter, any technique that pulls away chaff without grabbing wheat translates to significant savings of time and money while improving quality by minimizing candidates for mischaracterization.

So, maybe we should be looking at the value in a second, unique keyword pass preceding review that, like Agatha Christie's knitting needle or the archeologist's knife, clears away loose earth. This pass doesn't look for responsive documents. It employs keywords to find documents that are NOT likely to be responsive; that is, it's calculated to clear away the extraneous matter so we can see the naked shining truth.

This is "negative search." The notion of negative search isn't original with me, but neither is it much used by anyone else. Though similar in certain respects, negative search is not the same as using Boolean constructs to exclude noise hits. Boolean constructs are quite effective when artfully composed, but can be challenging to frame and tricky to execute. Negative search doesn't restrict queries in the way Boolean constructs do. Instead, negative search finds all documents containing terms deemed highly unlikely to occur within responsive documents, like “birthday cake,” “fantasy football” or “bridal shower.” These are then excluded from review. Clearly, negative search terms must be chosen wisely and ***tested carefully*** against representative samples of the collection before broad deployment. Like the NIST list, negative search terms, once compiled, can be used in subsequent cases–again with testing to guard against unexpected outcomes. So, consider if there’s a role for negative search in your next e-discovery effort and know that, in almost any collection, there’s a corpus of extraneous data that can be cost-effectively culled by negative search.

What are We Waiting For?

Winston Churchill said that, “Democracy is the worst form of government except all those other forms that have been tried from time to time.” That famous quip neatly describes keyword search in e-discovery. It stinks, yet lawyers turn to keyword search again and again, because it seems like the best option out there. It’s the devil we know.

Though keywords serve us well when searching the web, they perform poorly finding “all documents touching, concerning or relating to” an issue in litigation. The failure is particularly pronounced when keyword search is pursued in the usual fashion of opponents horse trading terms without testing them against sample data or adapting the list to ameliorate well-known flaws like misspellings, noise words and synonyms.

But that’s old news. Students of e-discovery know that keyword search is the worst form of search, and harbor no illusions that it’s better than the others that have been tried from time to time. Whether you call it advanced data analytics, predictive coding, concept search or whatever else leaps from the lips of marketing mavens, there exist techniques that, when implemented with care and judgment, do a better, less costly job than keyword search and linear review.

Yet whenever these techniques come up in conversations or articles, lawyers seem like kids inching toward the cookie jar, intently watching Mom’s face to see if it’s okay to snag some Mallomars. It may be better and cheaper, but nobody wants to give enhanced automated search much of a go until “it’s okay with Mom.”

What are we waiting for?

The answer seems to be some sort of authoritative court blessing of alternatives to keyword search. We’ve seen favorable mention of such techniques in footnotes to decisions from the most influential judges writing on e-discovery issues, but nothing opining that use of enhanced search is “court approved.”

Again, what are we waiting for?

It's not as though we held off using keyword search until a judge gave it the nod. We just did it. And, though keyword search never really got a judicial stamp of approval, neither was it summarily rejected. Again, we just did it, and in time it emerged as a standard.

Perhaps there will one day be a decision where a judge expressly cites enhanced search techniques as reliable proxies for human review or preferred alternatives to keyword search. I wouldn't hold my breath waiting for it. The American justice system doesn't favor advisory opinions. Courts expect genuine cases and controversies to drive our jurisprudence. New search techniques need to be used before they can be meaningfully addressed in reported decisions.

So, quit worrying about Mom and grab those Mallomars! If you believe enhanced automated search is better and cheaper, have the courage and wisdom to lead the way in its use.

All Wet

I was once trial counsel for the water authority of a Mexican city seeking damages for delay in the mapping of a water system serving three million customers. I learned that most water entering the pipes never reached consumers because the patchwork system was riddled with leaks—leaks difficult to repair because the water company didn't know where its pipes were buried. Repair crews made Swiss cheese of streets, but the massive leakage limited service to just a few hours a day. Those who could afford it erected tanks to hoard water. The rest suffered.

Until *Servicios de Agua y Drenaje* learned where its pipes lay, staunched the leaks and addressed local hoarding, the system stayed broken and the human and dollar costs extreme. *¡Ay, caramba!*

The thirsty señora at the spigot didn't care how hard or costly it was to collect, filter and deliver the water. She couldn't tell the water company what reservoirs and wells to tap, purification techniques to employ or pipes to use to route the water. She certainly didn't want to hear that she didn't *need* the water or hadn't used the faucet correctly. *She wanted a drink*, and felt it should flow to her in a timely and adequate way.

A judge could have ordered the water company to pump, but the cost in terms of wasted agua would have been astronomical and unsustainable. Telling the consumer to, "Find your own water or do without," was likewise untenable.

An apt metaphor for e-discovery, don't you think?

Litigants harbor immense reservoirs of ESI. Servers, like lakes and rivers, are evident and expansive. Databases and archives are vast, subterranean aquifers. Information puddles in desktops, portable devices and online storage. It's costly to preserve, tap and process, then much is lost to leaky mains:

- We don't know where our pipes are buried (lax records management).
- We let sources evaporate and sour (poor preservation).
- We poison the well (spoliation).
- We use sieves to dip and dowsing rods to explore (careless collection and search).
- We fill the tub when a basin would do (overbroad requests for production)
- We bathe in Perrier (conversion of ESI to image formats for manual review).

Through education, cooperation and improved tools and techniques, these holes are slowly getting plugged. Good thing, too, because our thirst for electronic evidence is growing fast.

Still, there's a leak in the pipes that draws no attention. Sometimes it's a trickle, and sometimes a gusher; but, if we don't find and gauge the loss, how will it ever get fixed?

This leak is blind reliance on text extraction and indexing engines as principal tools of ESI search.

Many think of electronic search in linear terms--as something that moves across the connected and collected sources of ESI comparing words and phrases to queries. Indeed, that's the way we search files on our computers and how computer forensic tools typically operate.

But most e-discovery search efforts aren't linear explorations but are instead run against an index of words extracted from the source data.

So, is that really different? Quite.

It may take hours or days to extract text and create the index, but once complete, searches run against indices are lightning fast compared to plodding linear search. That's the upside. But there's a noteworthy trade off to using indices: you may not find what you seek even though it's in the collection and you've chosen the right keyword.

It's important to appreciate how text extraction and indexing let data leak away. Text extraction tools parse data for sequences meeting the rules by which they define words. Is L33T a word? Is .DOC a word? How about 3.14159?

A simple parser might define a word as, "more than 4 but less than 14 contiguous alphabetic characters flanked by a space or punctuation." Parsers also employ rules barring certain combinations. Numbers, most punctuation and symbols are typically ignored, and common terms called "stop words" are sidelined, too. The very popular MySQL database excludes over 500 common English words and DTSearch excludes more than 120, so Shakespeare buffs can forget about finding "to be or not to be."

A more insidious shortcoming of indexed search flows from all the text that never makes it into the index.

ESI is encoded in many different ways, and it's common for encoded objects to be nested like Russian Matryoshka dolls: a Word document and a PowerPoint inside a Zip archive attached to an e-mail message within a compressed Outlook PST container file. Each nested object is encoded differently from its parent and child objects, and encoding may vary within the body of an object. Encoding is critical. In fact, next to metadata, it may be the most important thing most people don't understand about e-discovery.

When a parser processes encoded ESI, it must apply the appropriate filter to the data to convert it to plain text so it that can be indexed. If the data is encoded in multiple ways, multiple filters must be applied in the correct sequence to cycle through all different forms of encoding to reach any textual content. If no filter or the wrong filter is applied along the way, the text isn't indexed. This occurs various ways, e.g., the encoding isn't recognized, the tool doesn't support the encoding, the content isn't text or the file is corrupted, encrypted or password protected.

If a parser doesn't recognize the encoding, it may default to applying the most common textual encoding schemes to the unrecognized content in a last-ditch effort to find intelligible text. But, that doesn't always work. Foreign alphabets employ many more than our paltry 26 letters. Ideographic languages like Chinese and Japanese don't separate words with spaces. Even in English, you don't want to miss finding "résumé" when you search for "resume," so success hinges on whether the index is accent-sensitive or insensitive. Text parsers work around these challenges in various ways, but not all perform in the same way. There's many a slip between cup and lip.

Sometimes failure is hard coded into indexing applications when they're designed to pass over file types deemed unlikely to hold text or to apply only rudimentary text extraction methods. For example, Microsoft's Windows Search and Index Server have a limited capacity to index the contents of Access databases.

Finally, text extraction tools can't capture what they don't see as text. Facsimile or tiff images are classic examples of text-laden documents not captured. These and documents storing text as vector graphics must undergo optical character recognition to expose text. The same concern applies to linear search, but you can subsequently run OCR against source data. You can't do that to an index.

Maybe that's the ultimate failing of indices: *they're just a shadow of the evidence*. Because it's not the data—and failures are set in stone--you can't apply new and better ways to tease out the truth.

Is it wrong to employ indexed searches in e-discovery? Certainly not, but it's wrong to select a tool for a task it can't accomplish. So, do your homework on the parser and indexer, then test your extraction and indexing engines against representative samples of the data in the case and evaluate its performance in search. You should be prepared to disclose which encoded formats, file types and stop words are absent from the index. You need to know the capabilities and limits of the text extraction and indexing engines you deploy, because if the index won't hold water, you're up a creek.

Unlocking Keywords

The notion that words hold mythic power has been with us as long as language.

We know we don't need to ward off evil spirits, but we still say, "Gesundheit!" when someone sneezes. Can't hurt.

But misplaced confidence in the power of word searches can seriously hamper electronic data discovery. Perhaps because keyword searching works so well in the regimented realm of automated legal research, lawyers and judges embrace it in EDD with little thought given to its effectiveness as a tool for exploring less structured information. Too bad, because the difference between keyword searches that get the goods and those that fail hinges on thoughtful preparation and precaution.

Text Translation

Framing effective searches starts with understanding that most of what we think of as textual information isn't stored as text. Brilliant keywords won't turn up anything if the data searched isn't properly processed.

Take Microsoft Outlook e-mail. The message we see isn't a discrete document so much as a report assembled on-the- fly from a database. As with any database, the way information is stored little resembles the way we see it onscreen after our e-mail program works its magic by decompressing, decoding and decrypting messages.

Lots of evidence we think of as textual isn't stored as text, including fax transmissions, .tiff or PDF documents, PowerPoint word art, CAD/CAM blueprints, and zip archives. For each, the search software must process the data to insure content is accessible as searchable text.

Be certain the search tool you or your vendor employ can access and interpret all of the data that should be seen as text.

Recursion

Reviewing a box of documents that contains envelopes within folders, you'd open everything to ensure you saw everything.

Computers store data within data such that an Outlook file can hold an e-mail transmitting a zip archive containing a PowerPoint with an embedded .tiff image.

It's the electronic equivalent of Russian nesting dolls. If the text you seek is inside that .tiff, the search tool must drill down through each nested item, opening each with appropriate software to ensure all content is searched. This is called recursion, and it's an essential feature of competent search. Be sure your search tool can dig down as deep as the evidence.

Exceptions

Even when search software opens wide and digs deep, it will encounter items it can't read: password protected files, proprietary formats, and poor optical character recognition. When that happens, it's important the search software generates an exceptions log flagging failures for follow up.

Know how the search tool tracks and reports items not searched or incompletely searched.

Search Term Tips

So far, I've talked only about search tools; but search terms matter, too.
You'll get better results when you frame searches to account for computer rigidity and human frailty. Some tips:

Stemming: Computers are exasperatingly literal when searching. Though mechanized searches usually overlook differences in capitalization, they're easily confounded by variances in prefixes or suffixes of the sort that human reviewers easily assimilate (e.g., flammable and inflammable or exploded and exploding).

You'll miss fewer variations using stemmed searches targeting common roots of keywords; e.g., using "explod" to catch both exploded and exploding.

But use stemming judiciously as the more inclusive your search, the more challenging and costly the review. Be sure to include the correct stemming operator for the search tool.

Boolean Search: Just as with legal research, pinpoint responsive items and prioritize review using Boolean operators to find items containing both of two keywords, or keywords within a specified proximity.

Misspelling: It's scary how many people can't spell. Even the rare good speller may hit the wrong key or resort to the peculiar shorthand of instant messaging.

Sometimes you can be confident a particular term appears just one way in the target documents—e-mail addresses are prime examples—but a thorough search factors in common misspellings, acronyms, abbreviations and IM-speak.

Synonyms: Your search for "plane" won't get off the ground if you don't also look for "jet," "bird," "aircraft, "airliner" and "crate."

A comprehensive search incorporates synonyms as well as lingo peculiar to those whose data is searched.

Noise words: Some words occur with such regularity it's pointless to look for them. They're "noise words," the static on your ESI radio dial.

I recently encountered a situation where counsel chose terms like "law" and "legal" to cull data deemed privileged. Predictably, the results were disastrously over inclusive.

I recommend testing keywords to flush out noise words. There's irrelevant text all over a computer—in spelling dictionaries, web cache, help pages, and user license agreements. Moreover, industries have their own parlance and noise words, so it's important to assess noisiness against a representative sample of the environment you're searching.

Noise words are particularly nettlesome in computer forensic examinations, where searches extend beyond the boundaries of active files to the wilds of deleted and fragmented data. Out there, just about everything has to be treated as a potential hiding place for revealing text.

Because computers use alphabetic characters to store non-textual information, billions or trillions of characters randomly form words in the same way a million typing monkeys will eventually produce a Shakespearean sonnet. The difference is that the monkeys are theoretical while there really are legions of happenstance words on every computer. Consequently, searching three- and four-letter terms in forensic examinations—e.g., "IBM" or "Dell"—can be a fool's errand requiring an examiner to plow through thousands of false hits. If you must use noisy terms, it's best to frame them as discrete occurrences (flanked by spaces) and in a case-specific way (IBM but not iBm).

Striking a Balance

Effective keyword searching demands more than many imagine. You don't have to put every synonym and aberrant spelling on your keyword list, but you need to appreciate the limits of text search and balance the risk of missing the mark against the burden of grabbing everything and the kitchen sink. The very best results emerge from an iterative process: revisiting potentially responsive data using refined and expanded search terms.

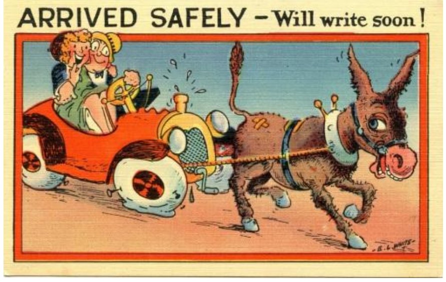
Train, Don't Cull, Using Keywords

I've been thinking about how we implement technology-assisted review tools and particularly how to hang onto the on-again/off-again benefits of keyword search while steering clear of its ugliness. The rusty flivver that is my brain got a kick start from many insightful comments made at the recent (July 2012) Carmel Valley E-discovery Retreat in Monterey, California. As is often the case when the subject is technology-assisted review (by whatever name you prefer, dear reader: predictive coding, CAR, automated document classification, Francis), some of those kicks came from lawyer Maura Grossman and computer scientist Gordon Cormack. So, if you like where I go with this, credit them. If not, blame me for misunderstanding.

Maura and Gordon are the power couple of predictive coding, thanks to their thoughtful papers and presentations transmogrifying the metrics of NIST TReC into coherent observations concerning the efficacy of automated document classification. While they're spinning straw into gold. I'm still studying it all; but from where I stand, they make a lot of sense.

Maura expressed the view that technology-assisted review tools shouldn't be run against subset collections culled by keywords but should be turned to the larger collection of ESI (i.e., the collection/sources against which keyword search might ordinarily have been deployed). The gist was, 'use the tools against as much information as possible, and don't hamstring the effort by putting old tools out in front of new ones.' [I'm not quoting here, but relating what I gleaned from the comment].

At the same Monterey conference, Judge Andrew Peck reminded us of the perils of GIGO (Garbage In:Garbage Out) when computers are mismanaged. The devil is very much in the details of any search effort, but never more so than when one deploys predictive coding in e-discovery. M*ethodology matters.*



If technology-assisted review were the automobile, we'd still be at the stage where drivers asked, "Where do I hook up my mules?" Our "mules" are keyword search.

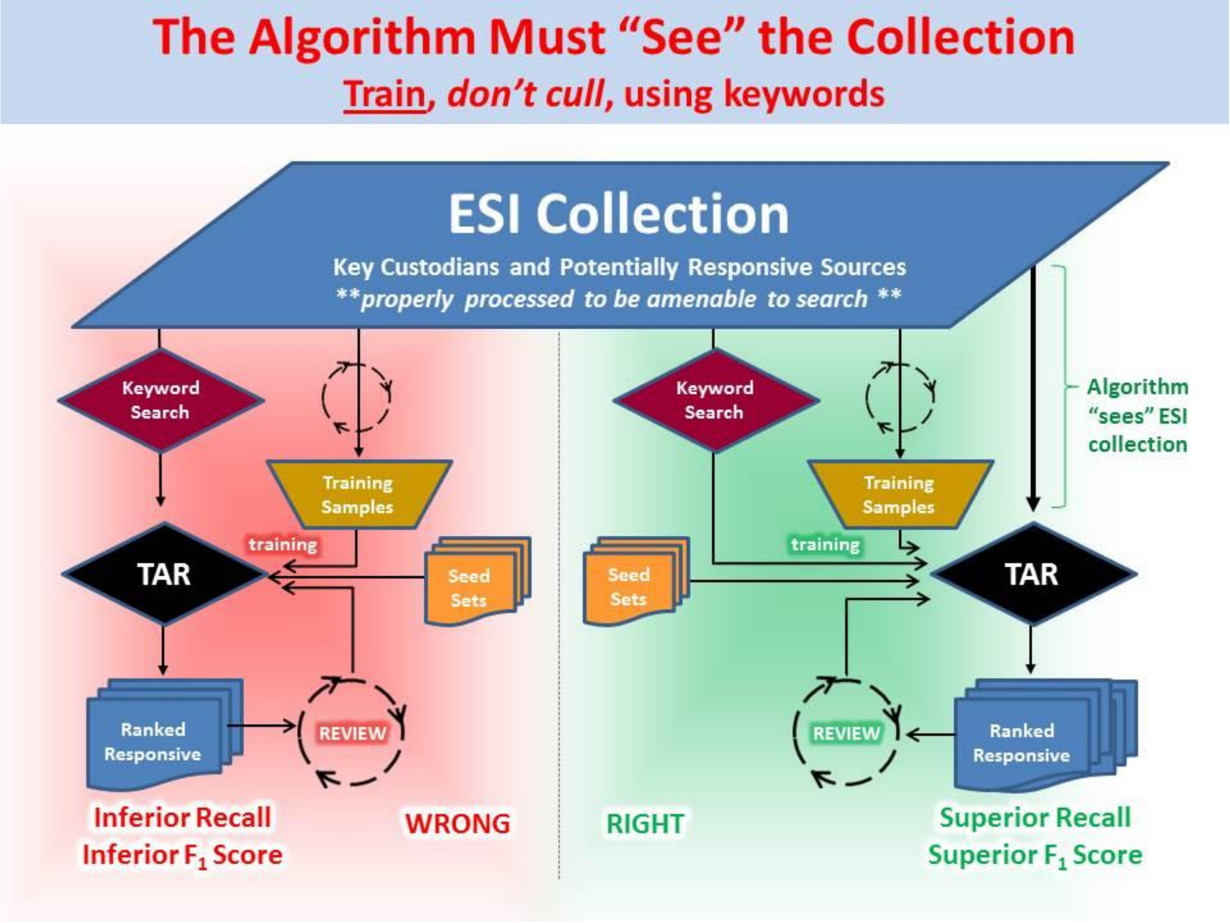
When you position keyword search in front of predictive coding; that is, when you use keyword search to create the collection that predictive coding "sees," the view doesn't change much from the old ways. You're still looking at the ass end of a mule. Breath deep the funky fragrance of keyword search. Put axiomatically, no search technology can find a responsive document that's not in the collection searched, and keyword search leaves most of the responsive documents out of the collection.

Keyword search can be very precise, but at the expense of recall. It can achieve splendid recall scores, but with abysmal precision. How, then, do we avail ourselves of the sometimes laser-like precision of keyword search without those awful recall in-laws coming to visit? Time-and-again, research proves that keyword search performs far less effectively than we hope or expect. It misses 30-80% of the truly responsive documents and sucks in scads of non-responsive junk, hiding what it finds in a blizzard of blather.

To be clear, that's an established metric based on *everyone else in the world*. It doesn't apply to *YOU*. *YOU* have the unique ability to frame fantastically precise and effective keyword searches like no one else. Likewise, all the findings about the laughably poor performance of human reviewers applies only to *other* reviewers, *not to YOU*. Tragically, not everyone has the immense good sense to employ *YOU*; so, let's take *YOU* and what *YOU* can do out of the equation until human cloning is commonplace, okay?

For all their shortcomings, mules are handy. When your Model-T gets stuck in the mud, a mule team can pull you out. Likewise, keyword search is a useful tool to pull us out of the sampling swamp and generate training sets. Using keywords, you're more likely to rapidly identify *some* responsive documents than using random sampling alone. These, in turn, increase the likelihood that predictive coding tools will find other responsive documents in the broader collection of ESI sources. *Good stuff in:good stuff out.*

With that in mind, I made the following diagram to depict how I think keyword search should be incorporated into TAR and how it shouldn't.



The Algorithm Must "See" the Collection

Train, *don't cull,* using keywords

I hope you'll agree that the interposition of keyword search to cull the collection before it's exposed to an automated document classification tool is wrong. But, in fairness, doing it the right way could come at a cost depending upon how you approach the assembly and processing of potentially responsive ESI. If you have to pay significantly more to let the tool "see" significantly more data, then quality will be sacrificed on the altar of savings. How it shakes out in your case hinges on how you handle keyword search and what you're charged for ingestion and hosting. Currently, many use keyword search via entirely separate tools and workflows to reduce the volume of information collected, processed and hosted. *Garbage In.*

Another caution I think important in using keywords to train automated classification tools is the requirement to elevate precision over recall in framing searches to insure that you don't end up training your predictive classification tool to replicate the shortcomings of keyword search. If only 20% of the documents returned by keyword search are responsive, then you don't want to train the tool to find more documents like the 80% that are junk. So when, in the illustration above, I depict keyword search as a means to train technology-assisted review tools, please don't interpret the line leading from keyword search to TAR as suggesting that the usual guesswork approach to keyword search is contemplated and you'll just dump keyword results into the tool. That's like routing the exhaust pipe into the passenger compartment. The searches required need to be narrow–precise–*surgical*. They must jettison recall to secure precision…and may even benefit from a soupçon of human review.

For the promise of predictive coding to be fulfilled, workflows and pricing must better balance the quality vs. cost equation. Yes, a technology that is less costly when introduced at nearly any stage of the review process is great and arguably superior only by being no worse than alternatives. But if that is all we seek when quality is also within easy reach, we do a disservice to justice. The societal and psychic benefits of a more trusted and accurate outcome to disputes cannot be overvalued. "Perfect" is not the standard, but neither is "screw it."