WHITE PAPER

WHERE ANGELS FEAR TO TREAD: THE PROBLEMS OF KEYWORD SEARCH IN E-DISCOVERY

By J.R. Jenkins, MLIS, Senior Product Marketing Manager at FTI Technology

FTI CONSULTING | TECHNOLOGY

The heavy reliance on keyword search in e-discovery places an enormous burden on today's legal teams. Inconsistencies in language, inefficiencies in search techniques and software user interfaces, which conceal more than reveal, place the attorney in a difficult position: determining what is relevant in a compressed timeline using obsolete tools and tactics. These outdated tools are a key factor behind the spiraling costs and risks associated with e-discovery.

With data collections growing at an epic pace, legal teams must sift quickly through documents and e-mail at rates that would have seemed unrealistic just 5 years ago. Megabytes of data grew quickly to gigabytes which have grown into terabytes – the typical matter now contains millions of documents and e-mail messages driving the average cost of a matter to more than 2 million dollars[1]. The ability to reduce and review these documents within agreed upon timeframes have placed incredible pressure on legal teams. While advances in search technologies have benefited many, the legal review space finds itself using outmoded tools and methods to analyze, review, and tag documents within a collection. Simply put, keyword search strategies break when thrown against the mountains of data which define today's legal matters.

At the same time the e-discovery market has captured the attention of many search vendors that have quickly branded their technologies as solutions for the legal domain. Very few of these vendors bring an intimate understanding of the legal discovery process; rather they bring generic search technologies which do little to aid the modern legal team. Users of these systems are asked to formulate complex Boolean search queries, surround themselves with statisticians and linguists, and have clairvoyant insight of a matter's details before it begins. These keyword search solutions deliver data rather than knowledge, and hindsight where foresight is needed.

These facts have not gone unnoticed by the legal community. The Sedona Conference has published several white papers which contain practical recommendations for dealing with these challenges and several key opinions have been written by the bench on these topics. In this paper – first in a three part series - we will examine the difficulties of working with keywords during a matter, the many challenges associated with relying on keyword search, and the cost and risk associated with attempting to "guess" one's way to a relevant document set.

In the second and third part of this series we will discuss how intelligent combinations of visual analytics and concept clustering can help legal teams mitigate inherent risks of keyword search strategies and, in fact, restore keyword search to a more proper place within the e-discovery process.

THE INTERSECTION OF E-DISCOVERY COST AND RISK

E-discovery and document review are expensive, difficult, and fraught with risk. Whether dealing with frequent in-house investigations or "bet the company" litigation, the legal team which can most quickly obtain and understand the potential importance of critical details about the matter earns both tactical and strategic advantages. Acquiring this knowledge then, becomes a race to discover key information as expeditiously as possible.

Just one process in nine (as defined in the EDRM model), the document review phase is estimated to account for up to 80% of the total cost of e-discovery, and according to IDC, can exceed the storage costs associated with a matter by more than 1400x[2]. The heavy reliance on human review – examining documents one at a time – is most responsible for this cost. Given this fact, one would assume that the tools used by legal teams would accelerate their unique talents. This is decidedly not the case.

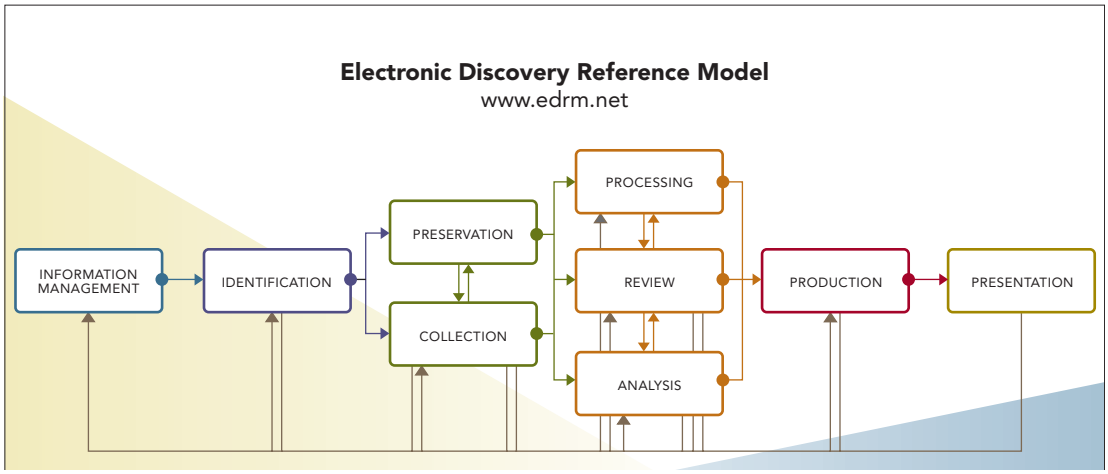

Figure 1: The EDRM Model - Review accounts for up to 80% of costs

Most legal reviews today rely heavily on search engines and emphasize keyword-based search strategies, placing a software interface between the attorney and the documents which make up a matter. This unnatural barrier forces counsel to deploy hunt and peck methods in their quest for relevant documents, while the software does little to accelerate the process. Search results are displayed in a simple relevance ranked order and any critical connections or shared characteristics between the documents are not revealed. More importantly, potentially relevant documents which do not satisfy search parameters can remain completely hidden from view for the length of the matter and whose late discovery can have disastrous consequences.

Getting around this barrier is a constant tradeoff between risk and cost. With exhaustive searching comes higher cost and less risk; with cursory searching comes higher risk and less cost.

TOO MUCH EMPHASIS ON SEARCHING – NOT ENOUGH ON FINDING

Keyword search can be effective if the collection is small, well understood and easy to access. However, modern legal review environments are dynamic, big, and unknown. The following table contrasts the "ideal" keyword search environment with the "reality" of e-discovery.

"IDEAL" KEYWORD SEARCH	REALITY OF E-DISCOVERY
The document collection is stable and static	Ad-hoc document collections
The size of the collection is small	Constantly changing corpus as new documents emerge
Heavy use of the index has created knowledge around the documents that aid in retrieval ("show me more like this")	No user history or value-added information in the collection (query logs, click-thru stats)
The end user is familiar with the unique characteristics of the data / documents / subject matter	Little understanding about the nuances of the subject matter
The end user is experienced with the nuances of the specific search engine	Search tools which are complex, difficult to master
Need to find a single "best" document	Need to find "every" relevant document

Thus, the rapidly changing e-discovery environment is difficult to assess using traditional keyword search solutions, leading to frustration for all parties involved in a legal matter. Fortunately there are groups which are dedicated to the legal community and are grappling with these many challenges.

THE SEDONA CONFERENCE ON KEYWORD SEARCH

The Sedona Conference – comprised of federal judges, leading attorneys, and experts in e-discovery - was founded in 1997 with the goal of providing the legal world with practical principles and working guidelines on a wide array of issues surrounding e-discovery and is "dedicated to the advancement of law and policy in the areas of antitrust law, complex litigation and intellectual property rights." This independent think tank meets annually and has published several papers concerning the process and methods used within legal matters – and have written extensively about e-discovery. They readily acknowledge the challenges of keyword search – both in the realms of keyword search and the quality of the process.

These Sedona papers are essential reading for anyone associated with e-discovery, but especially for attorneys working within the confines of keyword search.

In the *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* paper, the Sedona Conference concludes that three problems are inherent to e-discovery:

- *Exponential growth in informational records is a critical challenge to the justice system.*
- *Electronically stored information contains human language, which challenges computer search tools. These challenges lie in the ambiguity inherent in human language and tendency of people within organizations or networks to invent their own words or communicate in code.*
- *The comparative efficacy of the results of manual review versus the results of alternative forms of automated methods of review remains very much an open matter of debate. Moreover, simple keyword searching, while itself a valuable tool, has certain known deficiencies.*[3]

These known deficiencies of keyword searching are root causes of both risk and unwarranted expense in e-discovery and are many.

COMMON PROBLEMS INHERENT IN KEYWORD SEARCH

Despite the best intention of legal teams to create and use keywords for a legal matter, keyword searching is beset with challenges. **The great majority of search solutions on the market were not built with e-discovery in mind, but rather to provide a simple method for finding a single document or a few documents that match a search term.** This "data retrieval over document retrieval" methodology ignores the unique needs of the legal world – namely, returning *all* of the relevant documents. And while the complexity and nuances of human language are troublesome in most search solutions – they are critical issues in e-discovery.

The following issues are common and costly problems when relying exclusively on keyword search:

CREATING KEYWORD LISTS TOO EARLY

Depending on the size and scope of a matter, the legal issues in play encompass a wide variety of topics and, therefore, words. Developing a list of keywords which accurately represent the subject matter of a case can be extremely difficult.

The burden on the lead attorney is to create a list of terms that is neither too narrow (potentially missing key information) nor too broad (producing a document review process which contains an onerous amount of irrelevant material to review) and to also create them before understanding the complete scope of the matter. Too often this requires guess work and conjecture.

While attorneys may create a search keyword list based on their own experience and knowledge of the legal subject matter (e.g. a bankruptcy case and the typical activities that surround it) each case brings with it variations in data and nuances in language. To this end, the actual examination of the documents at issue is critical to the integrity of a list of key words. The implications for choosing poor or incomplete search lists are many, most notably the costs associated with "testing" the terms as well as the costs associated with over or under collection.

The opportunity cost for this is also real and may take time away from other equally important activities such as early case assessment, developing case strategy, and getting document review underway. Instead of making decisions based on documents within the matter, the legal team spends valuable time making decisions about the methods used to identify the documents.

SEARCHER (OVER) CONFIDENCE – THE BLAIR AND MORAN STUDY

In a 1985 study performed by David Blair and M.E. Moran (as reported in the Sedona Conference Commentary on Search and Retrieval), the authors found that searchers are often overly confident that they've found all of the relevant documents, believing their keyword search terms to be exhaustive and search strategies thorough. As an example, the two used a case that had generated more than 40,000 documents and 350,000 unique pages. Their findings demonstrate not just the above mentioned challenges within keyword search – but also, the critical human factor.

> *The attorneys estimated that they had found more than 75% of the relevant documents, but more detailed analysis found that the number was actually only about 20%. The authors found that the different parties in the case used different words, depending on their role. The parties on the BART side of the case referred to "the unfortunate incident," but parties on the victim's side called it a "disaster." Other documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were often not mentioned.*[4]

Far from describing an edge case, this study is a very common, recurring scenario in e-discovery (though, when compared to today's average matter size, the number of documents in this case seems downright quaint). The emphasis on "search" revealing matches while completely hiding those not deemed relevant make discovery of key concepts and documents an iterative, arduous, and expensive process.

OVERLY INCLUSIVE KEYWORDS

Terms which are too general in meaning are deemed overly inclusive. Lacking distinction within the documents, these terms can produce a results set much too large and force the reviewer to unnecessarily review too many documents at the clients expense. For this reason "stop words" (for example: and, or, this, that, at) are often excluded from a results list.

The term "confidential" is often excluded from keyword search lists for this reason. While many important documents and e-mails may include the term confidential to indicate the sensitive nature of the information, many e-mail signature blocks, common as a form of disclaimer at the bottom of nearly every email sent from an active lawyer in a law firm, include this term as well.

OVER INCLUSIVE SEARCH COSTS / RISKS
Key, relevant documents remain hard to find - some may never be found or given appropriate level of importance to case
Review team must review thousands of (likely) irrelevant documents
Number of returned documents can deceive review teams – believing they have found all relevant information - high volume doesn't necessarily equal successful, thorough retrieval
Costs to client are high and can grow exponentially
How much is enough (for an informed, accurate review that obtains the best possible result for a client)? When is collection "complete"?

Running a search on this term will likely produce a set of documents which are indeed important to the matter, but the search will also usually find an overwhelming amount of e-mail which is clearly irrelevant.

This irrelevant material brings with it tremendous costs. A single gigabyte of e-mail can produce 18,000 unique documents which require review. The following table describes some of the costs associated with use of overly inclusive search terms.

To reduce these costs, legal teams can attempt to narrow their search results using more specificity in their queries, as well as Boolean searches.

UNDER INCLUSIVE KEYWORDS AND BOOLEAN SEARCH

Terms too specific in meaning can miss critical documents and are deemed under-inclusive. Use of such terms within keyword search engines can lead to spare results sets – missing the vast majority of the relevant documents. Boolean searches, too, are often employed to work around the overly inclusive problem but have been proven to be a less than ideal method. Boolean searches are difficult to construct and the incorrect use of Boolean operators - AND, OR, BUT – can produce wildly differing results sets. The Text Retrieval Conference (TREC), a working group of scientists and mathematicians who are sponsored by the National Institute of Standards and Technology (NIST) has demonstrated in studies that Boolean searches produce mere fractions of the relevant documents in a corpus.

While less costly up-front, the downstream risks for the use of under inclusive terms are just as perilous as those found with over inclusive search. The following table describes some of the risks found when using under inclusive search terms during e-discovery.

UNDER INCLUSIVE SEARCH COSTS / RISKS
Key, relevant documents remain hard to find - some may never be found or given appropriate level of importance to case
Critical case knowledge may remain undiscovered (and unused)
Belated discovery (worst case, at trial) of critical, relevant material opens legal team to sanctions / malpractice / other potential case-crippling legal rulings
Boolean searches are difficult to construct and can produce wildly differing results sets
How much is enough (for an informed, accurate review that obtains the best possible result for a client)? When is collection "complete"?

SINGLE TERM – MANY MEANINGS

Human language is complex and nuanced – words are building blocks of thought, used in an endless variety of combinations to communicate concepts and meaning – and can frustrate keyword searches. In a study focused on challenges of document retrieval in commercial environments, David Blair, a researcher with the University of Michigan, discussed that as a collection grows so to do the meanings and uses of a single term:

> *. . .words have multiple meanings or uses so the words which represent documents in retrieval systems are, by themselves, inherently ambiguous.*[5]

Blair demonstrates how quickly unique meanings can grow within a collection. Using sets of 1000 and 100,000 documents –we find the unique use of the word "computing" grow from 10 to 84 instances, an 8-fold increase in "uniqueness."

> *Consequently, the increase in the size of the document retrieval system may not be increasing the number of documents relevant to the searchers request by very much, it may only be increasing the size of the retrieved set the searcher must browse through to find what he wants – the haystack in which the searcher must find the needs is simply getting larger.*[6]

BLAIR / ZIPF EXAMPLE:
1000 documents in the collection = 100 "matches" with 10 unique uses of the term "computing"
100,000 documents in the collection = 7,100 "matches" with 84 unique uses of the term "computing"

Keyword search systems only exacerbate these linguistic challenges, on a collection of any size, and turn the search for relevant documents into a slow, risky trial and error process.

MANY WORDS – SINGLE MEANING

Synonyms can make it difficult to properly identify terms up front. Ideas can be expressed in a wide variety of ways, making accurate keyword search term generation impossible. An example can be found in the simple concept of "missing the big picture." Many phrases and euphemisms can relate this concept:

- Can't see the forest for the trees
- Lost in the details
- Cart before the horse
- Tunnel vision
- Down in the weeds

This demonstrates the challenge of accurately and exhaustively searching for a conceptual state. The variety and history of our language make prescription of comprehensive terms up front a time-consuming and error prone endeavor.

MISSPELLINGS, ABBREVIATIONS, AND TEXT MESSAGING

A common occurrence, misspelling a word, can also be a problematic feature of keyword search as misspellings, acronyms, and abbreviations are not found or presented by search engines. And while some popular Web based engines will prompt users with 'did you mean' when encountering a misspelled term – this value added knowledge is often lacking in the ad-hoc, legal search experience. The explosion of text messaging also brings with it an entire new form of language - words and phrases which can be both difficult to define and equally difficult to find.

Company names serve as a useful illustration – often being referred to via use of abbreviations or acronyms in e-mail communications between employees, as well as in company documents. While some terms may be readily apparent to those inside the matter, attorneys representing a client might not know the various terms used to describe companies, divisions or groups. Keyword search only returns "matches" and cannot anticipate or find items which fall outside of a specific query.

Consequently, if key variations and misspellings of a word are at first missed (or if custodians are poor spellers), and then found late in the process, expensive supplemental procedures must occur.

DELIBERATE OBFUSCATION (*OR USE OF CODE WORDS*)

In cases of fraud, perpetrators seldom include terms like "fraud," "illegal," or "stealing" in their documents. Instead unique terms or "code words" are often developed to hide illegal activities. These terms, developed specifically to avoid suspicion or detection, can generally defeat the most thorough efforts to define a comprehensive list of keywords, and whose discovery then relies completely on serendipity or pure luck.

Examples of code words are well known to the legal community. Anecdotal examples include the use of baseball terms - "home run" and "bunt"- to hide insider trader activities and the use of produce – carrots, cabbage, peas – to communicate transfers of money. No amount of preparation or research into case specifics can aid in the discovery of these terms. And the exclusive use of keyword search will simply keep them hidden from the review team.

THE VIEW FROM THE BENCH

The pain and complexity of e-discovery has been topic of much discussion within the judiciary. Several opinions have called out the problems with trying to find the correct information within a matter using keyword searches:

> *Whether search terms of "keywords" will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics. Given this complexity,* ***for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread.*** (JUDGE FACCIOLA – US V O'KEEFE – FEB 2008)

> *While keyword searches have long been recognized as appropriate and helpful for ESI search and retrieval,* ***there are well-known limitations and risks associated with them,*** *and proper selection and implementation obviously involves technical, if not scientific knowledge.* (JUDGE GRIMM, VICTOR STANLEY V CREATIVE PIPE – MAY 2008)

In an attempt to provide greater clarity and guidance on the problem, the courts are increasingly citing the efforts of the Sedona Conference who have published several position papers that encourage practical, pragmatic search methods and transparent communication with opposing counsel.

FINDINGS FROM THE SEDONA CONFERENCE

KEYWORD SEARCH MISSES RELEVANT DOCUMENTS

The Sedona Conference aptly warns attorneys that current search tools have demonstrable shortcomings and must be used with great care and awareness of the faults. Practice Point 5:

> *The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.*[7]

Keyword search then is crude and unpredictable – but it is also common. In order to progress against the back drop of inefficient searching – and the increase in risk and costs for clients, Sedona is recommending more unity within the legal world.

QUALITY ASSURANCE IS CRITICAL TO KEYWORD SEARCH ENVIRONMENTS

To this end, better collaboration and communication is urged among opposing parties. Attempts to measure quality early and often are also encouraged. In the Commentary on Achieving Quality in the E-Discovery Process paper, Sedona provides an excellent general framework for assessing the quality of the collection, review, and production of electronically stored information (ESI).

This includes the development of protocols and processes that have their roots in modern production environments and rely both on statistical analysis and data sampling. When dealing with the "fuzzy" front end Sedona endorses the use of "sampling" to aid in defining the quality of the search terms used:

> *"The selection of keywords as search terms for responding to discovery requests is a special form of judgmental sampling based on many factors, including prior knowledge as well as educated guesses with respect to what a collection of ESI may contain" (emphasis mine).*[8]

Sampling can and should be used within the review phase of e-discovery when completely reliant on keyword search; however, it cannot change the following realities:

1. Counsel must generate a list of keywords – unique terms or phrases deemed critical to the case – at the onset of litigation, when they know the least about the matter at hand and are less able to make effective and accurate decisions about case strategy, tactics and likely outcomes.

2. Keywords are often created early on with little insight into the tools that will be used later to support the review. Unique strengths and weaknesses of search tools are not examined or taken into account.

3. The keyword list must be frequently revised as the matter matures. This creates stresses on the review team, especially if the matter is large and the review team geographically distributed. It can also create chaos and added expense if it requires re-review of documents multiple times or second-guessing of the methods used during an audit, etc.

FINDING BETTER METHODS

The Sedona Conference papers do more than provide recommendations for working within the limits of today's keyword search paradigm. They also encourage the legal community to remain vigilant in the search for newer and better technologies. Practice Point 8 urges that ***"Parties and the courts should be alert to new and evolving search and information retrieval methods."*** They continue:

> *What constitutes a reasonable search and information retrieval method is subject to change, given the rapid evolution of technology. The legal community needs to be vigilant in examining new and emerging techniques and methods which claim to yield better search results. In particular settings, lawyers should endeavor to incorporate evolving technological progress at the earliest opportunity in the planning stages of discovery or other legal setting involving search and retrieval issues.*[9]

Thus, to move beyond these problems, legal teams need a new view into the documents they review. They need a solution which rewards their unique abilities and allows them to use keyword search terms as a complement to their analysis tool set, rather than as the only window into the documents.

WHY SEARCH WHEN YOU CAN SEE WHAT MATTERS?

In the next two parts of this series we will discuss how the combination of visual analytics and concept clustering works within the Sedona Conference recommendations to provide a superior e-discovery experience for all parties by enabling documents to "describe themselves" to quickly overcome the many shortcomings of keyword search.

ABOUT THE AUTHOR

JR Jenkins, MLIS, is senior product marketing manager in FTI Consulting's technology practice, a leading e-discovery software and services provider. He has been active in developing industry thought leadership, such as an e-discovery XML standard, through his participation with the Electronic Discovery Reference Model Project (EDRM). Prior to joining FTI, JR worked on information retrieval projects for 10 years with companies such as Intel Research, Microsoft, and ProQuest Information and Learning. He has been a featured speaker at Information Online Sydney, NISO, and several American Library Association conferences on the topics of Search, OpenURL, and information behavior. He holds a master's degree in library and information science from the University of Washington and a bachelor's degree from Michigan State University.

4, 5, 6 Blair, D. C. (2002). *The challenge of commercial document retrieval, Part I: Major issues, and a framework based on search exhaustivity, determinacy of representation and document collection size.* Information Processing & Management , 273-291.

1 Information Week. (2009). Easing the Pain of E-Discovery. Information Wee, by Andrew Conry-Murray. June 1, 2009.

2 IDC presentation: Information Retrieval for eDiscovery: Beyond Keyword Search , by Vivian Tero.

8 The Sedona Conference. (2009). *Commentary on Achieving Quality in the E-Discovery Process.* The Sedona Conference , 1-23.

3, 7, 9 The Sedona Conference. (Fall 2007). *The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery.* The Sedona Conference Journal , 193-216.