

Understanding the Limitations of Keyword Search

Conrad J. Jacoby, Esq.

Attorneys have used the same basic technique to find potential relevant evidentiary documents in litigation for generations: (1) find a document of significance to a case; then (2) try to find other documents like it to complete the story. For attorneys and paralegals skimming through boxes of documents, specific words and phrases have been useful indicators of a document's potential value, permitting them to use concrete criteria to exclude documents in a discovery document collection and to focus on a smaller group of documents much more likely to have genuine relevance to the legal dispute. In particular, keywords have provided clear and simple direction to the young attorneys, paralegals, and other members of a legal team not fully familiar with the issues in the case who are often drafted to assist with document review projects.

Today, keyword search continues to be a common, if not the most common, technique used to search discovery document collections for potentially relevant material. Attorneys often use their own knowledge of a case and its underlying subject matter to develop a list of terms likely to occur in documents of interest, but some also use subject matter and even linguistic experts to develop extensive lists of words and phrases likely to occur in relevant documents. Because search term lists are developed through brainstorming, not all search terms will actually occur in the target document population. However, it's difficult, if not impossible, to predict in advance which search terms will best identify documents, so queries are typically refined through testing and sampling.

Computer technology is essential to applying complex keyword searches to large document populations. Computers have made it possible to quickly search thousands, if not millions, of documents with a single query. Computer programs and the Boolean logic used to control them have also made it possible to create complex searches that can include some terms, exclude others, and check whether key terms are within a specified proximity to one another. None of these complex searches could be carried out with any degree of accuracy using only human reviewers, but automated keyword search has been accepted by attorneys, judges, and litigants as an objective method for locating documents.

And yet, for all its power, the limitations inherent in keyword search routinely generate unsatisfactory results when this method is used to search discovery document collections. Keyword search possesses the seemingly contradictory weaknesses of finding too few documents (under-inclusion) and finding too many documents (over-inclusion). Of late, these limitations have led to a small but growing judicial voice questioning whether keyword search alone meets the legal standards for reasonably and defensibly looking for potentially relevant documents and information.

Keyword Search Is Under-inclusive

The greatest challenge in organizing a document collection of any size is properly classifying its documents. For litigation document reviews, improper categorization may have extremely significant consequences, such as when documents protected by attorney-client privilege are inadvertently disclosed to an opponent. Keyword search has been used for many years to help legal teams quickly identify documents that appear to have relevance to the immediate legal dispute.

Unfortunately, keyword search offers only an imperfect solution to this problem because it misses many relevant documents. By its fundamental nature, keyword search is sharply focused to find the exact terms specified in the query. Search for documents containing the word “car,” and you’ll find exactly what you’ve requested. However, you’ll miss documents containing potentially relevant words like “automobile,” “Ford,” “GM,” and “Toyota.” One can compensate by increasing the number of search terms and by adding stem searching to find plurals of words, but the results will still likely overlook some relevant materials.

It’s seemingly always possible to add more search terms to enhance a keyword search query, but even with hundreds of search terms (one recent reported case, *In re Fannie Mae Securities Litigation*, __ F.3d __, 2009 WL 215282009, U.S. App. LEXIS 9 (D.C. App. Jan. 6, 2009), involved 400 search terms), questions can still remain as to whether the search authors included enough relevant terms and concepts in their query. Many organizations and professions develop an internal vocabulary for common objects or activities that take place. Within the medical field, for example, “PDR” refers to the “Physicians’ Desk Reference,” a collection of Food and Drug Administration-approved drug labels and warnings. However, at NASA, “PDR” refers to a “Preliminary Design Review,” a project management milestone required in all engineering projects. Outsiders routinely overlook one or more specialized terms and fail to include them in a search query. As a consequence, relevant documents are not caught by the search query.

Keyword searches may also overlook key documents due to human error. Many litigation document collections contain voluminous amounts of e-mail messages, many of which are sent without being spell-checked and from mobile e-mail

devices with thumb keyboards that encourage innovative abbreviation. Keyword search will overlook messages where key terms have been misspelled or otherwise fail to match the search criteria. Some of the most important documents in a case—raw, unbiased, timely commentary written as events are taking place—may be overlooked for this reason.

Compelling evidence of the limitations of keyword search appears in the most recent results of a multi-year study conducted by the National Institute of Standards and Technology. The Text Retrieval Conference (“TREC”) Legal Track study, which is designed and managed by a mixed group of academics and legal practitioners, seeks to evaluate the ability of different search technologies, including keyword search and pure human review, to identify relevant documents in a standardized collection in which the actual mix of documents is already known. In 2008, as in prior years, participants were organized into mock legal teams and asked to develop keyword search queries that would identify documents relevant to legal claims raised in mock complaints.

Results of the 2008 TREC “consensus” keyword searches were compelling. While some of the topic-specific Boolean queries found more documents than others, on average, Boolean keyword search found only 24% of the total number of responsive documents in the target data set. *Overview of the TREC 2008 Legal Track* at pp. 8-10 (available at <http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf>). These findings are consistent with the 2007 TREC results (22% vs. 24%), and they strongly support the finding that many if not most relevant documents are overlooked when using keyword search as the sole means of identifying responsive materials.

Keyword Search Is Over-inclusive

At the other end of the spectrum, keyword search queries routinely vacuum up many wholly irrelevant documents whose review wastes time and money. The mere mention of a term is no indication that it is being used in a relevant context. In a products liability lawsuit, searching for documents containing the word “car,” for example, will likely also capture documents about car pools, car seats, car sickness, and car washes. If stem searching is used in conjunction with keyword search, the same query might also return documents discussing carrots and cartons. In almost all cases, such unexpected search results are simply not relevant to the intended purpose of the query and are unintended byproducts of an unduly-broad search query.

Increasing the number of search terms to reduce under-inclusion further exacerbates the problem of capturing irrelevant documents. In the *In re Fannie Mae Securities Litigation* case previously cited, the requesting party’s 400 keyword search terms flagged 80% of a government agency’s entire e-mail archive as being potentially relevant to an underlying legal dispute in which it was

not even a named party. It's extremely likely that most of the hundreds of thousands of documents so identified would ultimately be deemed irrelevant, but that determination will only be the voluminous search results are individually reviewed. Such a task can consume enormous resources (9% of a Federal Agency's entire annual budget in the *Fannie Mae Securities* case) while providing only modest benefits.

Keyword Search Alone Does Not Provide Legally Defensible Results

In light of well-publicized cases where a litigant failed to locate key documents until long after fact discovery had closed, many federal and state courts are taking a harder look at the reasonableness of a producing party's efforts to identify relevant documents for production in litigation. A growing number of courts have become increasingly skeptical about the adequacy of keyword search as the sole means of identifying potentially relevant documents. For example, in the widely-cited case of *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D. 251 (D. Md. 2008), presiding Judge Paul Grimm wrote:

[W]hile it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying exclusively on such searches for privilege review.

Victor Stanley, 250 F.R.D. at 256-57.

Another court has even called into question whether attorneys are competent to develop keyword searches in the first place:

“[F]or 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. This topic is clearly beyond the ken of a layman and requires that any such conclusion be based on evidence that, for example, meets the criteria of Rule 702 of the Federal Rules of Evidence.

United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008).

Taken together, these opinions strongly hint that while keyword search has not been completely discredited, new search and review technologies, particularly affordable methods for grouping and clustering documents based on their content, are changing the expectations that judges have for litigants who use automation to search document repositories and collections for relevant evidence.

Keyword Search's Inflexibility Can Cause Timing And Budget Issues

Finally, although it's not a legal issue *per se*, relying exclusively on keyword search to define a review document population can strain both discovery budgets and timelines when a case's underlying variables shift. For example, parties may negotiate a series of keyword search terms to define the documents that will be reviewed for relevance. One factor of the negotiation is the size of the search results—if a query flags too many documents, the parties agree that the search is overbroad. However, the legal claims in a case may shift after the negotiation is complete, due to an amended complaint or answer—or the joinder of additional litigants. Adding new search terms to the existing queries already approved by the parties may cause an unexpectedly large increase in search query results, generating a much larger (and much more expensive) document review for the producing party than anticipated. Depending on how far the case has developed before changes in its legal underpinnings, it may not be possible for the producing party to re-negotiate the initial search terms—or for the litigant to apply to the court for relief from a suddenly onerous document review.

Keyword search results can also exceed expectations when keyword search terms are based on the results of searching only a part of the target document population. Under the Federal Rules of Civil Procedure and an increasing number of state court analogs, parties must reach agreement on the scope of discovery early in a case—often before document collection is complete. As a result, keyword searches are tested against the documents that a litigating party has on hand at that stage in the dispute. However, as the document collection process continues to bring in materials from a broader range of repositories, initial estimates based on searches of the most easily accessible documents may be materially incorrect. For example, e-mail messages are often some of the first material collected and searched early in a case. However, these messages, many of which are fairly short or completely unrelated to the grounds of the dispute, may not contain many uses of terms of interest that are included in a negotiated keyword search. However, when the same search query is later applied to a more balanced document population that contains corporate memoranda, reports, and other more verbose documents, the search results will skew sharply higher than initial projections. And again, the producing party is left with a far larger number of documents to review for possible production—and relatively few ways to seek relief.

Conclusion

Keyword search remains an important starting point for most document review projects and for legal professionals who must quickly get an approximate sense of the potentially relevant documents within a collection. Increasingly, however, keyword search must be viewed as only one of several tools for identifying relevant documents. Relying exclusively on keyword search, especially in light of new tools and increasingly educated judges and opponents, runs the risk of mismanaging a key part of litigation fact discovery to the grave detriment of the client.

Copyright 2009 by Conrad J. Jacoby. All rights reserved.

Conrad Jacoby is the founder of efficientEDD, a consultancy specializing in electronic discovery and litigation information management issues. A seasoned litigator as well as a technology consultant, Mr. Jacoby writes and lectures extensively on electronic discovery issues. He is a long-standing member of the Sedona Conference Working Group on Electronic Document Retention and Production and is also the co-founder and Chair of the District of Columbia Bar Litigation Section's E-Discovery Committee. Mr. Jacoby can be reached at conrad@efficientEDD.com.