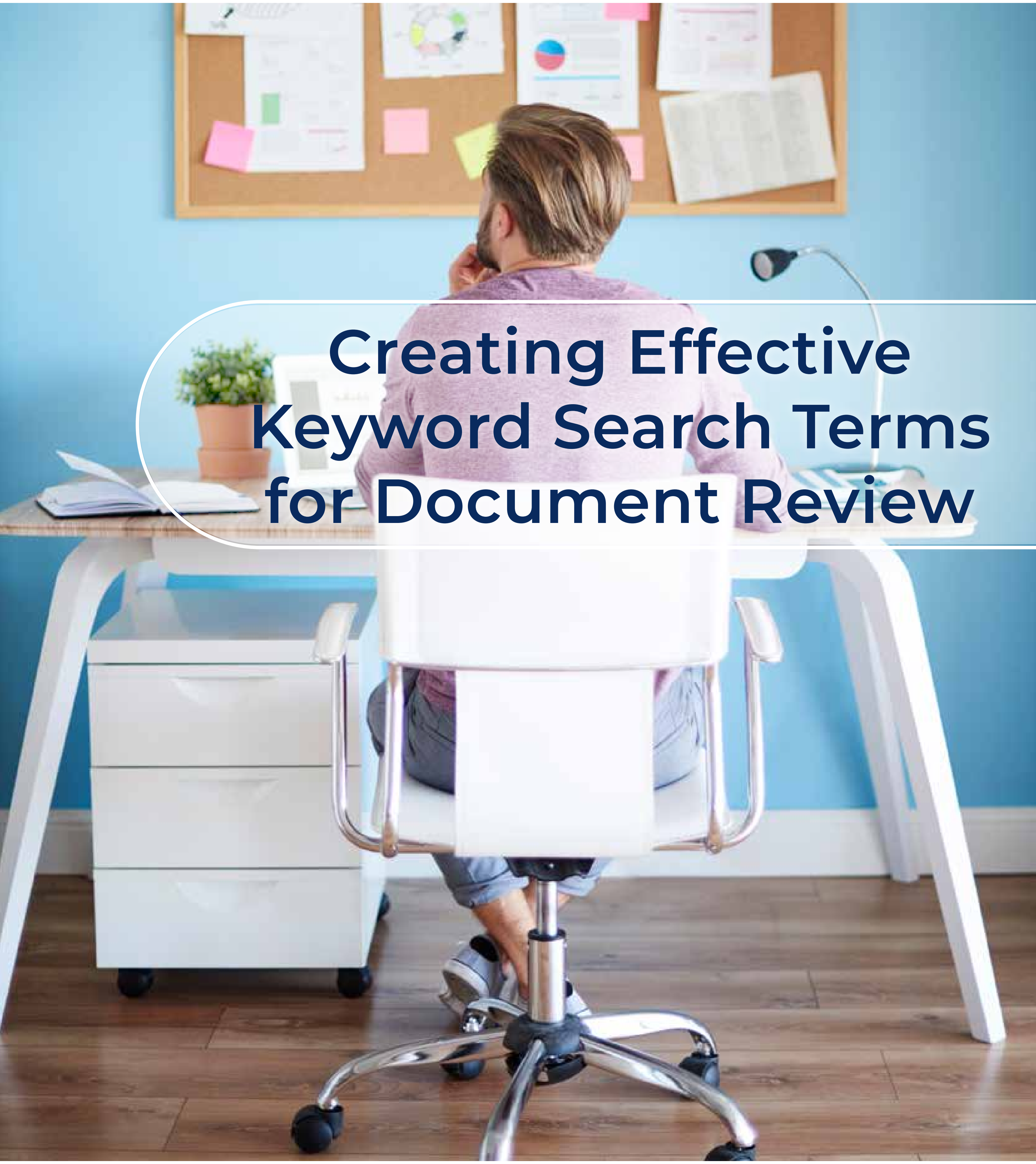
INNOVATIVE DRIVEN



Creating Effective Keyword Search Terms for Document Review

Creating Effective Keyword Search Terms for Document Review

INTRODUCTION

Keyword search remains the most used form of search in document review. In fact, even in workflows that include advanced analytics, keyword search is often used at least as a component of quality control. Therefore, it is important to not lose sight of best practices in this area, as its impact on the quality of document review can potentially alter the outcome of a case.

This whitepaper will discuss the following topics:

- Background of Keyword Search in Discovery
- The Legal Standards for Conducting an Adequate Search
- Challenges in Creating an Effective Keyword Search
- Best Practices in Conducting a Reasonable Keyword Search

BACKGROUND OF KEYWORD SEARCH IN DISCOVERY

Keyword search is a methodology for locating responsive electronically stored information ("ESI") by searching a database for specific words or combination of words relevant to your case, including terms relevant to the subject matter of the case, names of key personnel, dates, and other factors.[1] In general terms, its purpose is to attempt to "guess" the content of the text in the database in order to bring back documents relevant to the review. Therefore, it is imperative to actually check the results of the keyword search before relying on them, so that this guess becomes an educated one rather than a blind leap of faith.

The first look at a dataset can be intimidating because datasets typically consist of many unknown documents jumbled together. The first business of early data interrogation, then, is to put as much order to the chaos as possible. These efforts include de-NISTING (a process that removes certain system files as identified by the National Institute of Standards and Technology), de-duping (removal of actual duplicate documents as identified by their hash tag values), removal of junk records (spam and other nonrelevant email as identified by domain) and organization by file type.

Creating Effective Keyword Search Terms for Document Review

Along with (or prior to) early data interrogation, keyword searching is mainly used for two purposes: to cull a dataset prior to review or to prioritize data at the review stage. Search in e-Discovery can be conducted in virtually any stage of litigation: the first level is at the collection & preservation stage; the next level is at ingestion or during processing; and the final stage is during early case assessment (or early data interrogation) or preparation for review.

Some records information management technology allows parties to conduct specific searches to see communication patterns that can be used to enact targeted litigation holds on key data. Technology also exists at the collection stage, where data can be searched prior to extraction. One important caveat to searching at different levels is that a search performed before processing (i.e., at collection) is extremely risky because it limits the universe of data forever to those search results. Therefore, limiting keyword searches should not be performed at beginning stages unless both parties have agreed to the terms beforehand or it is otherwise certain what is to be determined relevant. Otherwise, the client runs the risk not only of not preserving all relevant data, but of not having access to data useful to the defense or prosecution of a case.

We will focus here on the document review stage, which requires traditional attorney analysis of ESI for discovery and motions practice.

Keyword searching is initially an educated guess regarding which terms and jargon were actually utilized by the custodians whose ESI is being reviewed; therefore, much care should be taken when constructing those terms. This requires that search terms be tested against the actual data, not just the expected data. As Magistrate Judge John Facciola remarked:[2]

...[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.

THE LEGAL STANDARDS FOR CONDUCTING AN ADEQUATE SEARCH

Keyword search methodology does not come with hard and fast rules in case law because it can be utilized in so many various ways. As in much of discovery, the linchpin appears to be one of reasonability. In one recent case, a party used keyword searching to narrow their dataset from 19.5 million documents to 3.9 million records, comprising 1.5 terabytes of data. Ultimately, 2.5 million documents were produced using a combination of keyword searches followed by predictive coding. The requesting party challenged this methodology, claiming the production should have been close to 10 million documents if predictive coding had been utilized on the entire 19.5 million documents. The Court refused to find the keyword searches to be unreasonable in first limiting the data prior to the use of predictive coding, partly due to proportionality concerns.[3]

Creating Effective Keyword Search Terms for Document Review

Producing parties are often challenged on whether their discovery production was the result of a reasonable search. A producing party can demonstrate a search was adequate with the following:[4]

> ***...[An] affidavit [that is] reasonably detailed, setting forth the search terms and the type of search performed, and averring that all files likely to contain responsive materials (if such records exist) were searched so as to give the requesting party an opportunity to challenge the adequacy of the search.***

Judges and receiving parties want confirmation that the person who conducted the search was qualified to do so when a search is challenged:[5]

> ***I must insist that the person performing the search have the competence and skill to do so comprehensively. An evidentiary hearing will then be held, at which I expect the person who made the attestation to testify and explain how he or she conducted the search, his or her qualifications to conduct the search, and why I should find the search was adequate.***

Judges take these requirements seriously. One Court denied a motion to compel on the allegation that a producing party failed to conduct adequate searches, because the moving party failed to present any expert testimony by affidavit, which would have allowed the Court to conclude that the producing party's search was inadequate.[6] Magistrate Judge James C. Frances IV took this a step further in refusing to rule on a search term dispute without any expert testimony, stating, "A court should be hesitant to resolve issues that demand technical expertise."[7]

In another case, the requesting party challenged the production of only 25 email messages as inadequate. The Court stated that it could not "compel the production of information that does not exist" and crafted the following remedy: the producing party was ordered to "disclose the sources it has searched or intends to search and, for each source, the search terms used."[8]

CHALLENGES IN CREATING AN EFFECTIVE KEYWORD SEARCH

The Importance of the Meet & Confer Conference Regarding Search Terms

Attorneys best serve their clients' interests by properly utilizing the Rule 26(f) conference to discuss proposed keyword search terms in advance, if applicable.[9] It is better to test the preliminary results of a keyword search before finalizing agreed-upon search terms. This can cut down on contention later, when terms are actually run and results are analyzed. At the time of the meet and confer conference, it is important for both parties to be familiar with the types and sources of potentially relevant data in their control, any data that may not be in their possession (for example, in cloud storage), date range limitations, key custodians, common terminology or industry jargon used by the custodians, and general knowledge of the technology used by the custodians. Cooperation in discovery is not just an ethical requirement for attorneys, but is also extremely practical, especially when it comes to the creation of preliminary search terms. Obviously, the best reason to agree on what needs to be searched is to avoid disputes about inadequate searches. It is just as important to cooperate on final search terms, not just by examining hit lists but also by testing against actual document results. It is not enough to delete a search term simply because it brings back a large number of data.[10] The true question is whether or not the data brought back is relevant. If not, that search term should be limited by additional syntax or edited to ensure the extraneous documents are not returned. Cooperation on the final results will help limit discovery disputes later.[11]

Creating Effective Keyword Search Terms for Document Review

Limitations of Technology : Indexing and Syntax

The first step in keyword search is creation of an index. An index is created from individual words taken from the extracted text of the dataset, minus certain "noise" or "stop" words such as "a" or "an." A keyword search can be used to cull a dataset by identifying relevant documents, nonrelevant documents, hot documents, privileged documents, or any other subset of documents.

A keyword search can be individual words or search strings. A search string contains proximity logic and connectors like OR, AND, and NOT. This search is then run against the index in an attempt to locate documents containing certain language.

It is important to ensure that the syntax utilized by the review tool itself is taken into consideration when drafting keyword searches. For example, "wildcard" characters such as "!" or "*" can be used to find different versions of a word – but you need to be sure which character "means" wildcard to a particular tool. Depending on the wildcard character recognized by the tool, a wildcard search of "earn*" or "earn!" can find every version of "earn," from "earnings" to "earnest." In fact, some search technology allows users to see the different versions from the wildcard and then select the search terms to use on the dataset.

As mentioned above, many tools remove "noise" or "stop" words from the index in order to speed up the searching process. However, this failure to index certain terms needs to be known to the creator of the search in order to make sure the search terms do not rely on a term that will not be indexed. One solution to this issue is to index the entire dataset. While this idea seemed drastic several years ago, technological improvements have increased the speed of search to the point that it is now a viable way to eliminate the possibility that portions of search terms will not be indexed.

BEST PRACTICES IN CONDUCTING A REASONABLE KEYWORD SEARCH

Search term creation requires understanding the language used by the parties communicating. For example, one Court ordered misspellings of a party's name to be included as search terms.[12] Consider the following factors when developing a search strategy:

Slang: Make sure the keywords include common slang, especially when searching informal types of communications such as instant messages, social media, or texts.

Terms of Art: Certain industries use their own jargon. Make sure you enlist the aid of someone familiar with an industry to complete your search terms if necessary. Acronyms: Many industries use common abbreviations and acronyms. Even individuals use modern Internet abbreviations and acronyms. Be sure to include these in your terms.

Alternative Spellings: Some terms lend themselves to multiple spellings, such as theater vs. theatre, color and colour, etc.

Common Footer Language: Some terms, such as "confidential," tend to appear in most email footers. Attempt to limit these terms by additional syntax.

Developing search terms is often a process of trial and error. Search term efficiency reports can be analyzed to determine which search terms are helpful against those that are false positives, or documents that are returned due to the term that are not relevant. For example, it is unlikely the word "earnest" will be a valid hit if searching for ESI pertaining to "interest earnings." The accuracy of search terms should be validated by manually checking a random sample of hit results.

Creating Effective Keyword Search Terms for Document Review

CONCLUSION

Lawyers will always have to search for information that proves their case and to find the evidence that tells their clients' stories. This universal constant requires lawyers to determine which ESI will ultimately result in evidence to which a judge, mediator, or jury will positively respond. The use of effective keyword searches can help reach this goal.

About Innovative Driven

Innovative Driven, developer of the ONE integrated e-Discovery platform, is a leading provider of customizable e-Discovery solutions and services across the Electronic Discovery Reference Model, as well as comprehensive computer forensics and expert consulting services.

Resources

1 Arkfeld on Electronic Discovery and Evidence, 5-31, citing Zakre v. Norddeutsche Landesbank Girozentrale, No. 03-257, 2004 U.S. LEXIS 6026 at 1-2 (D.N.Y. April 9, 2004).

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

3 In re Biomet M2a Magnum Hip Implant Prods. Liab. Litig., 2013 U.S. Dist. LEXIS 84440 (N.D. Ind. Apr. 18, 2013).

*4 Mullen v. United States Army Crim. Investigation Command, 2012 U.S. Dist. LEXIS 93977, at *13 (E.D. Va. July 6, 2012), citing Rein v. United States PTO, 553 F.3d 353, 362-63 (4th Cir. Va. 2009).*

5 Peskoff v Faber, 240 F.R.D. 26, 31 (D.D.C. 2007).

*6 Culler v. Shinseki, 2011 U.S. Dist. LEXIS 96043, at *25-26 (M.D. Pa. Aug. 26, 2011).*

*7 Assured Guar. Mun. Corp. v. UBS Real Estate Sec. Inc., 2012 U.S. Dist. LEXIS 167981, at *11 (S.D.N.Y. Nov. 21, 2012).*

*8 Am. Home Assur. Co. v. Greater Omaha Packing Co., 2013 U.S. Dist. LEXIS 129638, at *17 (D. Neb. Sept. 11, 2013).*

*9 One judge resolved an e-Discovery dispute by ordering disputing parties to have the attorneys responsible for ESI productions to hold an in-person Rule 26(f) conference to resolve outstanding production issues. Procongps, Inc. v. Skypatrol, LLC, 2013 U.S. Dist. LEXIS 47133, at *11-12 (N.D. Cal. Apr. 1, 2013).*

10 There is a potentially dangerous idea in e-Discovery: that you can save costs by forcing parties to use less search terms. This model order limits the number of email custodians to five and limits each requesting party to a total of five search terms per email custodian per party. http://www.cafc.uscourts.gov/images/stories/announcements/Ediscovery_Model_Order.pdf

*11 "Common practice governing the discovery of electronically stored information requires the use of search terms to make an extraordinarily burdensome search comply with the tenets of Fed.R.Civ.Proc. 26(b) (2)(C). If the producing party generates the search terms on its own, the inevitable result will be complaints that the search terms were inadequate." EEOC v. McCormick & Schmick's Seafood Rests., Inc., 2012 U.S. Dist. LEXIS 13134, at *14 (D. Md. Feb. 3, 2012).*

*12 Northington v. H&M Int'l, 2011 U.S. Dist. LEXIS 14378, at *2-3 (N.D. Ill. Feb. 14, 2011).*