Simple Ways to Make Search-Term Negotiation More Bearable

BY CHRISTINE PAYNE & ADAM NODZENSKI

FOR BETTER OR WORSE, AS A LEGAL COMMUNITY, we are still using search terms to aid in the identification of potentially relevant discovery documents. If you find yourself negotiating search terms with opposing counsel, there are a few very simple things you can do to make life easier.

A threshold note—this short article is *not* meant to address the more sophisticated questions about search terms: whether the application of search terms is actually a useful methodology for identifying potentially relevant documents, whether search terms can or should be used in conjunction with technology-assisted review (TAR), or whether search terms should be tested across an entire collection or across a sample coded for responsiveness.

Our goal is simply to address the situation where you are proceeding with a search-term negotiation (whether you like it or not) and wish to proceed in a reasonable and expeditious manner. For illustrative purposes, we are going to use a hypothetical matter where ownership of a photograph by artist Graciela Iturbide is in dispute. If you are not familiar with her work, go check it out! A moment or two with evocative art is good for the other half of your brain.

1. Make sure your terms will work.

This is a very simple point, but one that is frustratingly elusive in practice. If you are working with a vendor, or an in-house technical support team to conduct search-term analysis, have those professionals look at the terms being negotiated and confirm that the syntax of the terms will indeed operate within the system as intended. If you are working on your own in a licensed platform, read the user guide (which is often available in the Help section of the platform), or contact the software developer to get confirmation regarding their preferred operators, wildcards, and search logic.

Folks who do not take the time to get this confirmation may find themselves utilizing terms that will not operate (or will only operate partially) within their search system. Examples include running asterisk-style expanders when the system requires exclamation points, or including stop words (extremely common words in the English language that review-software developers purposely exclude because their ubiquity typically muddies results) such as "our," "of," or "the" in searching for a specific name like "Our Lady of the Iguanas."

The consequences of running improperly formatted terms are unpredictable—in some search systems, the term will just glitch out and return zero hits. In other platforms, a term might operate partially—maybe returning all hits for "lady" (which could be a lot) without regard to "Iguanas" (the more unique term), as referenced in the example above. In other words, the improper formatting could skew results either low or high.

While you are busy ensuring correct operation on your side of the fence, go ahead and get opposing counsel to confirm the same.

2. Make sure you're on the same page about "hits."

Most folks (and courts) agree that you should test out search terms ahead of time by running them across the document population and seeing how many "hits" you get. The basic theory being that one should know how many "hits" there are so that you know how many documents you'll have to ultimately review. And if you do some sampling you might also learn how many of those documents are likely to be responsive, and how many are likely to be junk. You would be shocked, however, at how frequently parties that are trying to negotiate the use of search terms have completely different understandings of what "hits" are – this could be search term hits or document hits, with either type provided with "unique," "total hit," or "full family" numbers. In that situation, everyone should back up and agree to only discuss de-duplicated document hits, and account for full families if the review will be performed in that manner.

What are de-duplicated document hits? Let's explore with an example. Suppose you have 1,000 documents and you are applying the search term "festival." First, you want to make certain that the 1,000 figure represents de-duplicated documents, reflective of the set that you will later review. Sometimes folks apply search terms to the collection population without going through the de-duplication process. This is problematic because it will artificially inflate the numbers.

After you confirm that the set has been de-duplicated, you would apply the term “festival” to the set. Within the 1,000 document population, there exist 45 documents that contain the term “festival” at least once in the document. So, the number 45 could be one type of “hit” rate—the number of actual documents retrieved by the term. Now imagine that several of those 45 documents are quite lengthy and use the term “festival” hundreds of times in each. If you count the number of times a word appears in a document set as the “hit” rate, that number could exceed 1,000, which would be more than the total number of documents you have in the first place.

But is it really necessary to know how many times the word “festival” appears in the set as a whole? Or, do you focus instead on the total number of documents retrieved using that term (which is 45)? In our practice, we care more about the latter, because we want to know how much work it’s going to be to go through the documents retrieved by a term, as the total number of documents for review ultimately drives the cost of the review. We don’t have an elegant name for that, though others may—we call it the de-duplicated hit rate or sometimes a “document retrieval count.” We hardly ever talk about the other count (the total times a term appears in a document population) except to the extent that frequency might be an indicator of how festival-related certain documents might be.

So, when you request a “hit” report, make sure you understand what data it will yield and what it means. Have the documents been de-duplicated? Are you looking at the total number of documents retrieved (as opposed to the total number of times a term appears)? Are you looking at unique documents retrieved by a particular term? (More on that, below). You may need to talk through this in detail with whomever is running your search-term analysis (as well as opposing counsel), because the concepts are (unfortunately) not something that are universally understood. Indeed, confusion on these points is quite common, even though we’ve been using search terms for years now. And yes, unique hit counts are helpful (see section 4, below), but you still need to know whether you’re looking at actual de-duplicated document retrieval numbers or something else. Finally, if you’re going to be reviewing and producing documents in full family groups, make sure that you also understand the total volume after accounting for families. Having to review 45 documents doesn’t sound too bad, but if each one of those is in a gigantic family group, that could be a problem.

3. Break down long strings.

Often litigants propose “terms” that are something more like long Boolean strings. Consider the following, which is a very typical example of a “term” that might be proposed:

> **(Contract or agree* or document* or paper*) and ((negotiate* or meet* or final*) w/25 (Magnolia or “Powerful Hands” or “Four Small Fish”))**

If you take this search string and simply run it across a document population of 20,000 documents, you might retrieve 15,000 documents. If that seems high to you and worthy of refinement, you’re left to guess at which portion of the string is really driving that number. If you break the string into its constituent parts, however, then you can clearly see what is driving the volume (and you also might realize just how many “terms” you’re actually running). For example: see chart at the top of page 24.

In the chart you can see that the largest volume is coming from just one of the constituent parts: agree* and (negotiate* w/25 Magnolia). This is very helpful. With this knowledge, you can go into the document set with confidence, target exactly the potentially problematic search term, and evaluate potential false positives. The broken-down format also allows you to see what you can easily agree to—maybe you might agree to all the constituent parts except for the term with the outsized number of hits, which could be refined.

4. Don’t continually run agreed terms (i.e., use the lockbox method)

This one is the most difficult to explain, but if you can master the technique, it pays off in spades. The basic idea is that while negotiating parties may consider many different search terms, some are more likely to be used than others. For example, you might decide that there are terms that are so important and unique that those terms may need to be searched regardless, such as the term “Iturbide” itself. (Of course, sampling might reveal that “Iturbide” isn’t as unique as you think it is, but that’s a topic for another article.)

What is likely to happen is that one side will propose a list of search terms to the other; let’s suppose it’s five terms long—and one of the five terms listed will be “Iturbide.” If the party receiving the proposal agrees that “Iturbide” should

Individual component term	Number of documents retrieved (de-duplicated)
Contract AND (negotiate* w/25 Magnolia)	289
Contract AND (negotiate* w/25 "Powerful Hands")	17
Contract AND (negotiate* w/25 "Four Small Fish")	0
Contract AND (meet* w/25 Magnolia)	43
Contract AND (meet* w/25 "Powerful Hands")	0
Contract AND (meet* w/25 "Four Small Fish")	0
Agree* AND (negotiate* w/25 Magnolia)	13,789
Agree* AND (negotiate* w/25 "Powerful Hands")	146
Agree* AND (negotiate* w/25 "Four Small Fish")	5
Agree* AND (meet* w/25 Magnolia)	233
Agree* AND (meet* w/25 "Powerful Hands")	0
Agree* AND (meet* w/25 "Four Small Fish")	109
Document* AND (negotiate* w/25 Magnolia)	70
Document * AND (negotiate* w/25 "Powerful Hands")	0
Document * AND (negotiate* w/25 "Four Small Fish")	0
Document * AND (meet* w/25 Magnolia)	298
Document * AND (meet* w/25 "Powerful Hands")	0
Document * AND (meet* w/25 "Four Small Fish")	0
TOTAL	**15,000**

indeed be a search term (or just judges that there's not a strong basis for opposition), then they should run that term across the document population first, by itself, without the other four terms. Why? Because whatever documents are retrieved by that term are going to have to be reviewed no matter what. We should take them out of the pool and set them aside in a virtual "lockbox." At that point, what remains in the document population are solely documents that *might* have to be reviewed if they are retrieved by a search term that makes it onto the final list. We can run the remaining four terms across the more finite set of documents, creating clearer visibility about what truly remains in dispute.

To see how this works, let's suppose we have 100,000 de-duplicated documents in our collection. The other side proposes five terms. For illustrative purposes, we're going to do it the wrong way first, so we run all five terms across the set. We might get a hit report that looks like this:

ALL TERMS RUN ACROSS 100,000 DE-DUPLICATED DOCUMENTS		
Term	**Total Documents Retrieved**	**Unique Documents Retrieved**
Photograph*	74,124	3,598
Iturbide	64,035	11,238
Purchase* w/5 right*	15,392	10,275
Sonora	61,299	5,897
Iguanas	48,008	13,288

This is a very confusing report. To get oriented, let's start with the definition of "unique" documents. Those are documents retrieved by a specific term *and no other.*

The chart seems to show that there are a lot of documents that would be retrieved by the terms as drafted, but there also seems to be a lot of overlap in the set. In other words, there are likely documents that contain both the term "Iturbide" and also "Sonora." Because the unique document counts *do not* include any documents retrieved by more than one term, the sum total of unique hits will almost always be less than the search population.

One interesting fact is that there are nearly 75,000 documents retrieved by the term "photograph*," but only about 3,500 of those are unique. This probably indicates that the term "photograph*" isn't very helpful in terms of identifying relevant documents that are not also retrieved by some other term on the list. Is that other term "Iturbide"? Is it "Iguanas"? Hard to say.

What if, instead, you used the lockbox method? If, based on your belief that it will return responsive documents, you agree that the term "Iturbide" will be run regardless, then try running it first and seeing if that cleans up the noise: See chart top of page 25.

LOCKBOX RUN		
Term	**Total De-Duplicated Documents Retrieved**	**Unique De-Duplicated Documents Retrieved**
Iturbide	64,035	64,035

Next, take out all the 64,035 documents retrieved by the term "Iturbide." You're going to have to review those documents, so you should stick them in a virtual "lockbox," removing them from the set of documents that's being further evaluated for review. Once you do that, you're left with 34,965. These are documents that you might have to review, depending on how the remainder of the testing and negotiations shake out. Run the remaining terms across this smaller set. You might get something like this:

DISPUTED TERMS RUN ACROSS 34,965 REMAINING DOCUMENTS		
Term	**Total De-Duplicated Documents Retrieved**	**Unique De-Duplicated Documents Retrieved**
Photograph*	10,089	365
Purchase* w/5 right*	15,101	10,199
Sonora	6,139	5,897
Iguanas	13,754	13,288

This second search provides clarity. In this scenario, it shows quite literally that all the documents retrieved by "Iturbide" also had the term "photograph*" in them (because 74,124 minus 64,035 is 10,089). The total hit rate for "photograph*" has significantly decreased and is a lot less concerning given the unique hit count. This second run also shows us that there is little overlap between "Iturbide" and "Purchase* w/5 right*"—so that latter term should likely be refined (perhaps it should just be "purchase rights" or "purchased w/2 rights"). It also demonstrates that the terms "Sonora" and "Iguanas" are pretty good at identifying unique document sets not retrieved by other terms (because their total hits are close to their unique numbers).

The parties may still negotiate these terms further, but there's a lot more clarity about which terms are driving volume. And, furthermore, the lockbox method can be done iteratively. If, after looking at the second run of terms (the four terms across the smaller set), the parties agree that "Sonora" and "Iguanas" should be used, then they become lockbox terms also, and documents retrieved by those terms would be removed from the set and placed in the review queue. You'd be removing another 19,185 documents (5,897 + 13,288) from the disputed set, leaving only 15,780 documents in dispute, and only two search terms to apply, refine, and negotiate further.

For those practitioners who have been at this game for a while, the above example is obviously unrealistic. There are only five terms at issue and they're all pretty unique. The methodology, however, really yields well at scale—when there are millions of documents and the other side proposes dozens of search strings that break down thousands of constituent terms (see section 3, above). In that situation, you want to get clarity as fast as possible—what are we really fighting over? Are there documents we're going to have to review no matter what? If so, let's set them aside and use the lockbox method as many times as necessary to get a clear view of what is truly in dispute.

Christine Payne is of counsel at Gunster and is licensed in Texas and Illinois. Christine is a litigator focused on eDiscovery strategy and lives in Austin, TX.

Adam Nodzenski is of counsel at Gunster and is licensed in Illinois. Adam is a litigator experienced in all things eDiscovery and lives in Chicago, IL.

A huge thanks to Mary Kostick, Dot Wood, Dan Raffle, Michelle Six, and Andrea D'Ambra for their excellent feedback on this article, especially in the form of hearty debate! ★