WINSTON & STRAWN LLP

ARTICLE

Proportionality, Not Perfection, Is What Matters

NOVEMBER 2, 2017

This article originally appeared in Law360*. Reprinted with permission. Any opinions in this article are not those of Winston & Strawn or its clients; the opinions in this article are the authors' opinions only.*

Recently, a handful of jurists and commentators have caused a stir in the e-discovery community by trying to resurrect the argument that litigants should avoid using search terms (aka "keyword searches" or "keyword filtering") to filter or cull a document population before employing a predictive coding engine for responsiveness determinations. The gist of their argument is that using search terms to reduce a document population necessarily results in relevant documents being excluded from the review population, thereby reducing the overall accuracy of the review effort in the interest of efficiency. While there is much to be said about the technical merits of such an argument, the most important point that we can raise is a philosophical one: So what?

Discovery is not a perfect process, and it is not meant to be. Rather, litigants are obligated to make a reasonable search of their files, shouldering a burden that is proportional to the needs of the case, using the tools and processes they deem in good faith to be appropriate. It is only when the requesting party can point to a deficiency that it can ask the court to weigh in on the adequacy of the producing party's efforts. Although the rationale employed by those advocating the prohibition on such filtering practices is superficially attractive—no good-faith litigant wants to exclude relevant documents—upon closer examination, the attempt to prohibit search terms before using predictive coding inappropriately undermines the principle of proportionality at the heart of the recent changes to Federal Rule 26. Worse, it presents an exception that will swallow the whole rule when it comes to keyword search: If it is not reasonable to use search terms to filter a document population prior to using predictive coding, when is it reasonable to use search terms to filter a document population?

Search on Trial

The attempt to ban initial culling via search terms is not an insubstantial issue. It represents not only an attack on the emphasis the 2015 amendments to the Federal Rules of Civil Procedure placed on the need for proportionality, but also an attack on a litigant's ability to use keyword search at all to satisfy burdensome discovery obligations.

Considered in the abstract as a process for analyzing a given population of documents and identifying the relevant

material for production, a workflow relying on predictive coding is the same as a workflow relying on a cadre of attorney reviewers. As pieces of an overall e-discovery process, both are functionally interchangeable, and whether the "reviewer" in question is a group of human reviewers or subject matter experts using a predictive coding workflow, the prior use of search terms to cull the data set informs only the population that will be subjected to review. As such, the "no-cull" rationale is equally applicable, or inapplicable as the case may be, where the predictive coding process fulfills the role of "reviewer" or where the humans do. In either case, using search terms to cull the data set may result in the incidental exclusion of relevant documents from the review set in the interest of reducing the costs and burdens associated with review and production. This precise bargain—sacrificing marginal accuracy in the interests of substantial efficiency—has underpinned e-discovery review practices for the better part of three decades, and most importantly, has been deemed reasonable by countless judges for just as long.

Taken to the logical conclusion, what the "no-cull" proponents are really suggesting is that the use of search terms at all is inherently and unreasonably inaccurate, and should be eschewed in favor of the comprehensive use of predictive coding in every case. They would have the legal community replace the "search term bargain" with the judicial prescription that predictive coding constitutes the only "reasonable inquiry." But to the average litigant, that's a poor bargain. While keyword searches can be done defensibly using, most, if not all, of the e-discovery processing or review platforms on the market, or using an ever-increasing number of client-side on-premises or cloud-based document management and retrieval systems (e.g., Microsoft Office 365), the use of predictive coding often requires engaging specialized vendors, expert consultants and counsel who know their way around the technology. The advent of the species of predictive coding known as "continuous active learning" (CAL or TAR 2.0) is of little comfort, as the number of vendors currently offering true CAL tools can be counted on one hand (and, not coincidentally, also counted among those decrying the use of search terms to cull before using any version of predictive coding).

But why stop there, as the "no-cull" argument rationally extends to a more extreme, yet still logical, conclusion? At bottom, the "no-cull" rationale is just the same old argument—decisively rejected by Rule 26(b)—that parties should "boil the ocean" or "leave no stone unturned" in their pursuit of "all" relevant documents. As such, it applies to any technique that would circumscribe a data population but incidentally exclude relevant documents—from custodian selection to file-type exclusion. Once the door is opened and parties have been banned from using search terms to cull prior to using predictive coding technology, it is easy to imagine the next litigant using the no-cull rationale to support the argument that their opponent should collect its full information technology infrastructure (exchange servers, file storage volumes, document management systems, SharePoint sites—everything), process all of it, and then subject it to review in a predictive coding workflow, in pursuit of an unbiased, "total recall" result—a scorched-Earth process that would drastically increase the costs of data processing and bring most companies grinding to a halt.

Bias-Shmias!

Beyond the argument that the use of search terms is likely to result in exclusion of "relevant" evidence, some argue that the use of search term is likely to bias the predictive coding exercise and the ability of the algorithmic "brain" to accurately predict relevant vs. nonrelevant data. To that argument, we respond: poppycock. There is no empirical evidence to suggest that where search terms are used in a reasonable and defensible manner that they will bias the outcome of the exercise. Indeed, every predictive coding tool used for determining relevance should be used in the context of a reasonable and defensible process that includes statistical validation that the tool is achieving the desired precision and recall, taking into account an acceptable margin of error.

But Isn't It All Just a Wash?

It may be tempting to dismiss the forgoing concerns as just making the old "slippery slope" argument (but see Hyles), and/or as simply overblown. After all, search terms have been employed to reduce document populations from tens of thousands down to thousands, then hundreds of thousands down to tens, then millions down to hundreds of thousands. The savings—in terms of time, expense and burden—were self-evident. But this actually

makes the point for us. Those same concerns regarding time, expense and burden can also be alleviated by using search terms to cull a document set prior to running predictive coding.

Nevertheless, some "no-cull" proponents assume that the shift from predictive-coding-with-search-term-cull to merely predictive-coding-over-the-entire-collection may be done with little or no difference in cost. That simply is a false assumption. For example, it's worth remembering that one hour of so-called "subject matter expert" time (often a partner or senior associate with billing rates in the multiple hundreds of dollars per hour) can easily cost as much as 10, 15, or even more hours of reviewer time. In the context of a many reviews, SMEs are exponential cost multipliers.

Requiring all documents collected to be subject to predictive coding also results in tremendous basic e-discovery costs. For a predictive coding engine to analyze documents, those documents have to be processed and then indexed, and those steps have attendant costs. Predictive coding analysis of every document in a population means that every gigabyte collected must be promoted to the P.C. environment, at per-gigabyte costs that can range from $75 to $250. Effective, defensible use of keywords to cull a population first can reduce the volume promoted for indexing by 80 to 90 percent, meaning that a failure to first cull will increase processing and indexing costs by 400 to 900 percent. Since the vast majority of that data still will not be relevant or useful, for purposes of the efficient and inexpensive determination of the action, most of that money is as good as thrown away.

Importantly, one of the purposes of using search terms to cull the data set prior to running predictive coding is to increase the richness of the population to be sampled, so that adequate training (whether active or passive) can be accomplished in fewer rounds. In a low-richness dataset, training to achieve only 80 percent recall can go on for thousands upon thousands of documents. It isn't hard to conceive of a scenario where the added expense associated with expensive SME training in such a low-richness environment swiftly dwarfs the marginal utility of the relevant documents added to the review population by skipping a preliminary search-term cull—especially considering that a target recall of anything less than 100 percent necessarily means that at least some portion of relevant documents in the data set will be missed in any event.

No "One-Size-Fits-All" Proportionality

That is why this controversy implicates concerns about proportionality. Rule 26 defines the scope of discovery to encompass that which is both relevant to the claims and defenses in the action and proportional to the needs of the case; documents that can only be attained by disproportionate effort, even if relevant, are beyond the scope of discovery. That scope will always serve as a thorn for suggested blanket rules that a particular process, method or solution should "always" or "never" be adopted.

Counsel on both sides of the "v," as well as the court, necessarily make proportional decisions at every step in the litigation to define the scope of the body of evidence that will be discovered in a case. Examples that impact the scope of discovery include decisions on custodians, date ranges, file types and document specifications. Potentially "relevant" documents will always be excluded as a result of these decisions, but that expectation is at the very heart of the "proportionality" inquiry. Use of search terms in front of the use of a predictive coding exercise is only one more example of proportional decision-making.

Proponents of the "no-cull" rationale seek a blanket rule treating a potentially more accurate but less efficient choice as the presumptive "best" choice. In so doing, they ignore the foreseeable likelihood that the burdens associated with an unknown and speculative increase in accuracy by overinclusion may be disproportionate to the needs of a particular case. They also ignore the possibility that "culling" a document set using search terms may involve formulating exclusion criteria to identify clearly nonrelevant documents prior to training.

The reality is that no two cases require precisely the same approach, requiring litigants and judges to be flexible in prescribing solutions that are tailored to the matter and data in question. This is precisely why Rule 26(b) includes six nondispositive criteria for assessing proportionality in individual matters. A "no-cull" blanket rule will invariably

lead some parties to unnecessarily expend time and resources on inefficient and disproportional pursuit of a “more accurate” (read: more perfect) production. This is simply not supported by the Federal Rules or applicable case law.

No Shortcuts to Proportionality

So what would the blanket protect? It seems to be an unstated premise of the proponents of the “increased accuracy” of the “no-cull” approach that it is prima facie unreasonable to leave relevant documents undiscovered in the unfiltered set, without any idea of how prevalent those documents may be or how critical they may be to the case. A blanket rule would simply reflect the presumption that the otherwise undiscovered “relevant” material is both prevalent and important. These presumptions would relieve the parties and the court of having to assess whether it is reasonable in a given case to deprive the producing party of its choice of discovery methods and force them to shoulder a potentially heavy burden—in other words, of having to make the same sorts of showings and arguments that the parties made in the Biomet case.

That shouldn’t be how proportionality is applied under Rule 26(b). If a requesting party wants to attack a producing party’s use of search terms to cull before applying predictive coding, the requesting party should be required to show a likelihood that relevant and proportional material has been missed, then walk hand in hand with the producing party and the court in assessing both the prevalence of that material (through sampling) and its relative importance (through assessment of the sample set). We have no illusions that in some cases, the requesting party may have the better argument—such as where an inordinately large amount of relevant information may be excluded by the search terms, or where analysis reveals that some of that information is important to the resolution of the issues in the case. But in others, the producing party will be able to carry its burden and demonstrate that the effort demanded by the requesting party’s proposal is disproportionate to the needs of the case.

At the end of the day, predictive coding is a useful tool (among many) that can be used to promote the just, speedy and efficient outcome of the case when used with the right people and process. To date, however, there have been many stumbling blocks created, preventing the broader adoption of the tool in civil litigation. Prohibiting the use of search terms in advance of running a predictive coding exercise is just one more obstacle that will inhibit broader adoption—one without logical, legal, or empirical support.

10 Min Read

Related Locations

Washington, DC

Related Topics

Law360

eDiscovery

Related Capabilities

eDiscovery & Information Governance

Related Regions

North America

Related Professionals

John Rosenthal

Jason Moore