DrinkerBiddle

Publication - 05/01/2011

Why Document Review is Broken

Client Alert

By Bennett B. Borden

The review of documents for responsiveness and privilege is widely perceived as the most expensive aspect of conducting litigation in the information age. Over the last several years, we have focused on determining why that is and how to fix it. We have found there are several factors that drive the costs of document review, all of which can be addressed with significant results. In this article, we move beyond costs and get to the real heart of the matter: document review is a "necessary evil" in the service of litigation, but its true value is rarely understood or realized in modern litigation.

It was not always so. When the Federal Rules of Civil Procedure were first promulgated in 1938, they established a framework from the common law with respect to which discovery took place. But there was no fundamental change in how one conducted discovery of the comparatively few paper documents that comprised the evidence in most civil cases. There was no Facebook or even email at the time. Only later, when the sheer number of paper documents grew to a point where litigators needed help to get through them, and only later still when the electronic creation of documents became possible and then ubiquitous, did the "problem" of information inflation convert document review into a separate aspect of litigation, and one that accounted for a significant portion of the cost of litigation.

There are three primary factors that drive the cost of document review: the volume of documents to be reviewed, the quality of the documents, and the review process itself. The volume of documents to be reviewed will vary from case to case, but can be reduced significantly by experienced counsel who understands the sources of potentially relevant documents and how to target them narrowly. This requires the technological ability to navigate computer systems and data repositories as well as the legal ability to obtain agreement with opposing counsel, the court or the regulator to establish proportional, targeted, iterative discovery protocols that meet the needs of the case. Because of the important work of The Sedona Conference® and other similar organizations, these techniques are better understood, if not always widely practiced.[1]

At some point, however, a corpus of documents will be identified that requires careful analysis, and how that "review" is conducted is largely an issue of combining skillful technique with powerful technology. In order to take advantage of all of the benefits this technology can provide, the format of the documents, the data and metadata, must be of sufficient quality. When the format of production is "dirty" (*i.e.*, inconsistent, incomplete, etc.), you face a situation of "garbage in/garbage out." For several reasons, "garbage" in this sense no longer suffices.

As we discuss more fully below, the most advanced technology we have found uses all of the aspects of data and metadata to improve the efficiency (and thus reduce the cost) of the review process – and more. This means that the ESI must be obtained, whether from the client for its own review or from opposing counsel for the review of the opposing party's documents, with sufficient metadata in sufficiently structured form to capitalize on the power of the technology. This requires counsel with technological and legal know-how to obtain ESI in the proper format. Many negotiated ESI protocols have become long and complex, but they rarely include sufficiently detailed requirements concerning the format of documents, including sufficiently clean data and metadata such that the most powerful technologies can be properly leveraged. Without this, a great deal of efficiency is sacrificed.

Once a corpus of documents has been identified and obtained in the proper format, the document review commences. This is where we have found the greatest inefficiency, and this is the primary area in which the most significant gains are possible. Our analysis of the typical review process leads us to conclude that the process is broken. By this we mean that, typically, document review is terribly inefficient and has been divorced from its primary purpose, to marshal the facts specific to a matter to prove a party's claims or defenses and to lead to the just, speedy and inexpensive resolution of the matter. This disheartening conclusion led us to question whether document review could be completed efficiently and effectively within days or even hours so that a party could almost immediately know its position with respect to any claim or defense. That kind of document review could become an integral part of the overall litigation as well as the primary driver of its resolution.

But document review has become an end unto itself, largely divorced from the rest of litigation. The typical review is structured so that either contract attorneys or low-level associates conduct a first level review, coding documents as responsive, non-responsive or privileged. Sometimes the responsive documents are further divided and coded into a few categories. But this sub-dividing is usually very basic and provides only the most general outline as to the subjects of the documents. Typically, a second level review is conducted by more senior associates to derive and organize the most important facts. Thus, every responsive document is reviewed at least twice, and usually several more times as the second level reviewers distill facts from the documents to organize them into case themes or deposition outlines that are finally presented to the decision makers (usually partners).

This typical tiered review process is inherently inefficient and requires a great deal of time and effort. The most pressing question that arises in the beginning of a matter, "what happened?", prompts the answer, "We'll tell you in two (or three or six) months." This multiplicitous review process leads to lost information in transfer, lost time, and the attendant increase in cost. The three standard categories (responsive, non-responsive and privileged) result in oversimplification because not all responsive documents are equally responsive. Add to inefficiency, then, simple misinformation. Is this avoidable?

Document review became separated from the litigation process because of the increase in the volume of potentially relevant documents. With thousands or even millions of documents to review, law firms or clients typically threw bodies at the problem, hiring armies of contract attorneys to slog through the documents one by one in an inefficient linear process. The goal was simply to get through the corpus to meet production deadlines. But, if the whole point of document review is to discover, understand, marshal and present facts about what happened and why, then it is the facts derived from document review that drive the resolution of the matter. Thus, the entire discovery process should be tailored to this fundamental purpose. Part of this, as we have noted, must be accomplished through experienced counsel who understands what the case is about, what facts are needed, and how to narrowly and proportionally get at them. The other key is to derive facts from the reviewed documents as quickly and efficiently as possible, and transfer the knowledge distilled from those facts to the decision makers in the most effective and efficient way. In short, document review should be returned to its rightful place as fact development in the service of litigation.

In October 2010, we released an article entitled: *The Demise of Linear Review*, wherein we discussed the use of advanced review tools to create a more efficient review process.[2] There, we showed that by using one of these advanced tools, and employing a non-linear, topic-driven review approach, we were able to get through several reviews between four and six times faster than would be the case with less advanced tools using a typical linear review approach. Since then, we have focused on perfecting both the review application and our review processes. Our results follow below.

The application used in the reviews described in *The Demise of Linear Review* was created by IT.com and is called IT-Discovery (ITD). The non-linear, topic-driven reviews were conducted by a team of attorneys led by Sam Strickland, who has since created a company called Strickland e-Review (SER), and the reviews were overseen by the Williams Mullen Electronic Discovery and Information Governance Section. The ITD application uses advanced machine learning technology to assist our SER reviewers in finding responsive documents quickly and efficiently, as we showed in *The Demise of Linear Review*. But we wanted to show not only that our topic-driven review process was faster, but also that it was qualitatively better than a typical linear review. Here we move into the area of whether humans doing linear review are in fact better than humans using computer-assisted techniques – not only for cost reduction but for improving the quality of results. We tested this and concluded that humans doing linear review produce significantly inferior results compared to computer-assisted review.

To prove this, we obtained a corpus of 20,933 documents from an actual litigation matter. This corpus had been identified from a larger corpus using search terms. The documents were divided into batches and farmed out to attorneys who reviewed them in a linear process. That review took about 180 hours at a rate of about 116 documents per hour. The typical rate of review is about 50 documents per hour, so even this review was more efficient than is typical. Our analysis showed that this was because the corpus was identified by fairly targeted search terms, so the documents were more likely to be responsive. Also, the document requests were very broad, and there were no sub-codes applied to the responsive documents. Both of these factors led to a more efficient linear review.

We then loaded the same 20,933 documents into the ITD application and reviewed them using our topic-driven processes with SER. This review took 18.5 hours at a rate of 1,131 documents per hour, almost ten times faster than the linear review. Obviously it is impossible for a reviewer to have seen every document at that rate of review, so we must question whether this method is defensible. To answer that question, it is important to distinguish between reviewing a document and reading it.

Reviewing a document for responsiveness means, at the most fundamental level, that the document is recognized, through whatever means, as responsive or not. But this does not mean that the document has to be read by a human reviewer, if its responsiveness can otherwise be determined with certainty. The ITD application uses advanced machine learning

technology to group documents into topics based upon content, metadata and the "social aspects" of the documents (who authored them, among whom were they distributed and so forth), as well as the more traditional co-occurrence of various tokens and forms of matrix reductions that constitute modern machine learning techniques to data mine text. Because of the granularity and cohesiveness of the topics created by ITD, the reviewers were able to make coding decisions on groups of documents. But more interestingly, these unsupervised-learning-derived topics aid in intelligent groupings of all sorts, so that a reviewer can "recognize" with certainty a large number of documents as a cohesive group. One can then code them uniformly.

Does this mean that some documents were not "reviewed" in the sense that a reviewer actually viewed them and made an individual decision regarding their responsiveness? No. To understand this by analogy, think of identifying in a corpus all Google news alerts by the sender, "Google alert," from almost any advanced search page in almost any review or ECA platform, in a context where none of these documents could be responsive. Every document was looked at as a group, in this case a group determined by the sender, and was coded as "non-responsive." This technique is perfectly defensible and is done in nearly every review today. What we can do is extend this technique much deeper to apply it to all sorts of such groups and voilá: a non-linear review on steroids.

Isn't there some risk, however, that if every document isn't read, privileged information is missed inadvertently and thus produced? Not if the non-linear review is conducted properly. Privileged communications occur only between or among specific individuals. Work product only can be produced by certain individuals. Part of skillful fact development is knowing who these individuals are and how to identify their data. The same holds true for trade secrets, personal information, or other sensitive data. The key is to craft non-linear review strategies that not only identify responsive information, but also protect sensitive and privileged information.

We showed in our non-linear review that the SER reviewers, using the advanced technology of the ITD application, coded 20,933 documents in $1/10^{th}$ of the time that it took the linear reviewers to do so. The question then becomes, how accurate were those coding decisions? To answer this question, we solicited the assistance of Maura R. Grossman and Gordon V. Cormack, Coordinators of 2010 TREC Legal Track, an international, interdisciplinary research effort aimed at objectively modeling the e-discovery review process to evaluate the efficacy of various search methodologies, sponsored by the National Institute of Standards and Technology. With their input, we designed a comparative analysis of the results of both the linear and non-linear reviews.

First, we compared the coding of the two reviews and identified 2,235 instances where the coding of the documents conflicted between the two reviews. Those documents were then examined by a topic authority to determine what the correct coding should have been, without knowing how the documents were originally coded. Results: The topic authority agreed with the ITD/SER review coding 2,195 times out of 2,235, or 98.2% of the time.

Not only was the ITD/SER review ten times faster, it resulted in the correct coding decision 99.8% of the time. In nearly every instance where there was a dispute between the "read every document" approach of the linear review and our computer-assisted non-linear review, the non-linear review won out. Could this just be coincidence? Could it be that the SER reviewers are just smarter than the "traditional" reviewers? Or perhaps, as we believe, is the fundamental approach of human linear review using the most common review applications of today simply worse? The latter position has been well documented by Maura R. Grossman and Gordon V. Cormack, among others.[3]

The implication of this specific review, as well as those discussed in *The Demise of Linear Review*, is that with our ITD/SER review process we can get through a corpus of documents faster, cheaper and more accurately than with traditional linear review models. But, as we have noted, document review is not an end unto itself. Its purpose is to help identify, marshal and present facts that lead to the resolution of a matter. The following is a real-world example of how our better review process resulted in the resolution of a matter. We should point out that case results depend upon a variety of factors unique to each case, and that case results do not guarantee or predict a similar result in any future case.

We represented a client who was being sued by a former employee in a whistleblower *qui tam* action. The client was a government contractor who, because of the False Claims Act allegations in the complaint, faced a bet-the-company situation. As soon as the complaint was unsealed, we identified about 60 custodians and placed them on litigation hold, along with appropriate non-custodial data sources. We then identified about 10 key custodians and focused on their data first. Time was of the essence because this case was on the Eastern District of Virginia's "Rocket Docket." We loaded the data related to the key custodians into the ITD platform and SER began its review before discovery was issued. We gained efficiency through the advanced technology of the ITD platform. We also gained efficiency by eliminating the need to review documents more than once to distill facts from them. Our review process includes capturing enough information about a document when it is first reviewed so that its facts are evident through the organizing and outlining features in ITD. This eliminated the need for a typical second- and even third-level review.

Within four days, the SER reviewers could answer "what happened?". Soon thereafter, the nine reviewers completed the review of about 675,500 documents at a rate of 185 documents per hour. More importantly, within a very short time we knew precisely what the client's position was with respect to the claims made and had marshaled the facts in such a way as to use them in our negotiations with the opposing party, all before formal document requests had been served.

Knowing our position, we approached opposing counsel and began negotiating a settlement. We made a voluntary production of about 12,500 documents that laid out the parties' positions, and walked opposing counsel through the documents, laying out all the facts. We were able to settle the case. All of this occurred after the production of only a small fraction of all the documents, without a single deposition taken, and at a small fraction of the cost that we had budgeted to take the case through trial.

This real-world example demonstrates the true power of "document review" when understood and executed properly. Fundamentally, nearly every litigation matter comes down to the questions of "what happened?" and "why?". In this information age, the answers to those questions almost invariably reside in a company's ESI, where its employees' actions and decisions are evidenced and by which they are effectuated. The key to finding those answers is knowing how to narrowly target the necessary facts within the ESI. You then can use those facts to drive the resolution of the litigation. This requires the ability to reasonably and proportionally limit discovery to those sources of ESI most likely to contain key facts and the technological know-how to efficiently distill the key facts out of the vast volume of ESI.

The typical linear document review process is broken. It no longer fulfills its key purpose: to identify, marshal and present the facts needed to resolve a matter. Its failure is legacy to the nature of how it came into being as the volume of documents became overwhelming. We believe we have found the right combination of technique and technology to return the process to its roots, resolving litigation.

For more information about this topic, please contact the author or any member of the Williams Mullen E-Discovery Team.

---

[1] See, Bennett B. Borden, Monica McCarroll, Brian C. Vick & Lauren M. Wheeling, *Four Years Later: How the 2006 Amendments to the Federal Rules Have Reshaped the E-Discovery Landscape and are Revitalizing the Civil Justice System*, XVII RICH. J.L. & TECH. 10 (2011), http://jolt.richmond.edu/v17i3/article10.pdf.

[2] See, *The Demise of Linear Review*, October 2010, http://www.williamsmullen.com/the-demise-of-linear-review-10-01-2010/

[3] Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.