Finding Key Documents: Tech Savvy Counsel Are Changing the Game

Julia L. Brickell and Peter J. Pizzi, New York Law Journal

July 5, 2016

In a recent 60 Minutes interview with Bill Whitaker, Andrew Michaelson, one of the lead prosecutors in the insider trading trial of the Galleon Group hedge fund manager Raj Rajaratnam, recounted his struggle to comb through "hundreds of thousands" of trading and phone records, instant messages and emails to find the source of Rajaratnam's inside information. In an investigation that began in 2006, it took more than six months of painstaking manual analysis before the prosecution came upon the "needle in the haystack in a single, careless instant message from Roomy Khan," one source of that information.[1]

Ten years on, though the document volume would likely be exponentially larger, a team like Michaelson's could have found key information much more easily by calling on advanced tools and expertise to mine the electronically-stored information. Together, tools and know-how can probe data collections directly and strategically. The information that can be mined is more than the words on the page or the image in a picture. It includes hundreds of categories of data about each document (the metadata), allowing multiple methods to cull through the volumes, make important connections, and get answers lawyers need.

As courts and bar associations increasingly challenge lawyers to view the ability to understand technology as an ethical imperative (see, for example, the advisory opinion promulgated in 2015 by the State Bar of California Standing Committee on Professional Responsibility and Conduct regarding counsels' relationship to technology in discovery, and the amended commentary for the ABA Model Rules of Professional Conduct Rule 1.1 Competence[2]), counsel's understanding of advanced search and analytics approaches takes on increased importance. Whether working on an internal investigation or preparing for or engaged in litigation, there are choices available which enable counsel to make use of the data embedded in the information collected, including methods to streamline the process, identify key information, and assess the merits of trial or settlement approaches in a timely manner and at lowered cost. To ignore these potential alternatives is to proceed at some peril.

Some may think that the use of hosting platforms and common search tools are all the technology needed to put legal teams squarely in the digital age. Although hosting platforms are indeed light years ahead of a banker's box review, such technology is primarily used as an organizational tool to expedite a relevance and privilege tagging process, thereby speeding up a review for production. Finding the key documents that make a difference in a litigation or investigation is different matter, however.

Keyword search, if poorly constructed, will also likely fall short. As Judge John Facciola famously stated in *United States v. O'Keefe*: "Whether search terms or 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics. … for 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."[3] And, despite how well or poorly keyword search may do the job, if left as part of a manual process it still relies on a laborious review of the "hits."

To step towards the more sophisticated approaches available for targeting key documents in complex document collections, counsel today can rely on the triumvirate of expertise that Judge Facciola identifies and endeavor to use such expertise to *mine data* instead of *review documents*.

Distilling the data to determine the who, what, where, when, why and how of events when dealing with increasingly large collections—replete as they may be with multiple sources of data, proliferating file types, missing or inconsistent metadata, ambiguous email headers and hidden content—requires a significant expenditure of time, money and human resources. More advanced search and analytics tools coupled with the appropriate expertise in their use can help legal teams expose case-defining information more quickly. Consider the following benefits such tools and expertise can provide:

• **Data preparation:** The first step in any search effort is to reduce and organize the data collection to facilitate its interrogation. For key document analysis, the goal is to isolate data that is unique and more likely to be relevant. Experts can apply various techniques to identify duplicates and near duplicates, analyze email chains (aka "email threading") to locate the master emails that contain an entire thread (thus enabling subsumed content to be set aside), and apply data filters (think complex Boolean searches) that will target certain types of known relevant content and exclude obviously non-relevant content. Not only will these processes refine the data population, but they can help avoid the kind of situation Google faced in its litigation against Oracle over Java,[4] where drafts of a damaging email asserted as privileged slipped into the production. Near-duplicate identification and/or email threading would have increased the chances that the drafts would have been reviewed for privilege at the same time as similar content.

• **Identifying "who":** Identifying individuals directly or indirectly involved in the matter is critical, of course, and the data population may reveal previously unknown relationships. In *U.S. v. Rajaratnam*,[5] the discovery of Rajaratnam's connection to and communication with Roomy Kahn was the critical link in the chain. Expert application of analytics tools can help identify important players, whether to identify (or challenge) witnesses or actually establish culpability. In electronic communications, however, names take many forms and can appear not only in content text (perhaps as a name, perhaps as just "she"), but also in multiple places in metadata and email headers. Identifying name variants, that is, how individual names may appear in corporate email, personal email, instant message IDs, and distribution list associations (e.g., bill smith@abc.com, williamsmith@abcco.com, lakersfan@yahoo.com, bsmith12, Executive Committee, BOD, etc.), increases the chances that all information related to an individual will be found. The location of a name in a communication might even make a difference, depending on the nature of the case (was it a BCC? A name in a comment field?).

• **The "what, why, where, and how":** When it comes to the guts of a matter, it may be tempting to assume language that might have been used and construct keyword searches accordingly. According to 60 Minutes and case materials, the case-breaking text message from Roomy Kahn read "donot buy plcm till I het guidance; want to make sure guidance OK." Would a lawyer-constructed keyword search have revealed such coded jargon to Michaelson's team? Corporations and individuals routinely use shorthand, code words, chat acronyms, text shorthand, abbreviations, or metaphors—which search savvy, not intuition, will reveal.

Through subject matter knowledge acquired in working with counsel and the data, search experts can apply analytic tools and techniques that can segment search topics into their constituent elements to ensure that searches comprehensively cover the issues, entities, and activities that matter. Word and sentence similarity algorithms, for example, can quickly surface documents that contain content based on real or hypothetical phrases and sentences that could be key to the case, pulling up conceptually similar information even if none of the words are the same.

Feature-based searches can uncover discussions of interest that otherwise might escape detection. Such searches include emotional content filters that can surface documents with strong language or that express concern, stress, fear, secrecy, or even attempts to conceal information. Searches can also be informed by analysis of social networks (for example, isolating emails sent to one or very few correspondents which tend to be less formal and unvarnished than larger blasts), and can be refined to include key people, entities, and dates to ensure coverage of the targeted search topics. A variety of analytics tools can identify repeated use of terms or phrases with no apparent relationship to the topics of the litigation or a company's business, which could be code words with another meaning (using sports analogies to disguise illicit trading activities, for example).

• **The "when":** Consideration of time frames is crucial in locating key documents as well as supporting or refuting the chronology of facts. There are various ways to detect both references to time as well as the time information was created or circulated. Metadata is typically important in these analyses. The fidelity of the metadata has an impact on the types of analyses that can be performed and/or the thoroughness of the analysis. For example, with productions from multiple parties, are the metadata fields that are used to search for emails comprehensive and consistent across the collection? Experts typically normalize and group the metadata into consistent fields to ensure that, for instance, all email file types can be easily searched. Equipped with the right tools, experts may be able to find missing information and supplement metadata for search and analytics.

In looking at the collection as a whole, identifying a significant change or unexpected variation in the volume of emails or other documents over time may lead to further investigation on that basis alone. The insight may range from seeing gaps in the production to identifying when key activity was likely to be occurring.

In these and other ways, today's tools together with enough know-how can advance the efforts of counsel to find key documents. Undoubtedly, the complexities and growth of electronic information will continue to challenge litigation and investigation teams. Being aware of what is possible, being able to assess (with appropriate guidance) which possibilities may help meet client goals at reduced cost, and communicating choices to clients (as ABA Model Rule 1.4 contemplates[6]) all fall within the ambit of technological competence for litigation counsel.

Fortunately, advanced analytic tools and search techniques are constantly evolving and becoming more accessible to the layperson. But, the time and skill required to use and apply them strategically should not be underestimated. Ultimately, whether such tools and expertise are hired or acquired, it is certain that counsel who recognize their value will be in a stronger position to provide choices and to represent their client fully and effectively. For Michaelson's team in the *Rajaratnam* case, months of painstaking work may have been averted. Today, counsel who eschew a more technologically savvy approach may be facing an opposing side that doesn't, and it will likely make all the difference—for the other side.

Endnotes:

1. "Inside Edge," 60 Minutes, CBS (May 22, 2016) http://www.cbs.com/shows/60_minutes/video/Vfb5DOv93bX6I2NFqdM5ysMkVJ4ZW5XP/preview-inside-edge/(quoted excerpts available at http://www.cbsnews.com/news/60-minutes-rare-look-at-how-insider-trading-actually-works/).

2. State Bar of California Standing Comm. on Prof'l Responsibility & Conduct, Formal Op. No. 2015-193, Calif. Eth. Op. 2015-193 (2015) (discussing an attorney's ethical duties in the handling of discovery of ESI), available at http://ethics.calbar.ca.gov/Ethics/EthicsTechnologyResources/EthicsOpinionsRelatedtoTechnology.aspx, and Model Rules of Prof'l Conduct r. 1.1 cmt. 8 (Am. Bar Ass'n), available at http://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_1_competence/comment_on_rule_1_1.html.

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

4. *Oracle Am. v. Google*, No. C 10-03561 WHA, 2011 WL 5024457, *2 (N.D. Cal. Oct. 20, 2011).

5. *U.S. v Rajaratnam*, No. 09-cr-1184 (RJH), Doc. 1 (S.D.N.Y filed Oct. 15, 2009); see also *SEC v. Galleon Management*, No. 09-cv-8811, Doc. 1 (S.D.N.Y. filed Oct. 16, 2009) (JSR).

6. Model Rules of Prof'l Conduct r. 1.4 (Am. Bar Ass'n), available at http://www.americanbar.org/groups/professional_responsibility/publications/model_rules_of_professional_conduct/rule_1_4_communications.html.

Julia Brickell is executive managing director and general counsel of H5. Peter J. Pizzi is a founding partner of Walsh Pizzi O'Reilly Falanga.