By Linda Beaudoin and Sheila Mackay

Right-size Your Review Leveraging Data Analytics

Your digital footprint is bigger than you think. Most of the electronic communications you have had in the last ten years may exist somewhere: on servers, backup tapes, hard drives or other removable media. It is not exactly breaking news that the amount of electronically stored information (ESI) has exploded over the last 15 years, but there are considerable expenses associated with the prevalence of digital information. The costs to store, manage, organize, preserve, collect, process, filter, review, produce and analyze this data have grown proportionately. With so much more ESI, the challenge faced by corporations and their law firms is to identify, preserve, review and designate documents for production in a cost effective manner. In other words, you need to right-size your review to be proportionate to the value of your case.

Using Search Criteria

Imagine being tasked with developing a strategy to locate relevant information in a database of text messages, rife with constantly changing slang and acronyms. Now imagine that this collection of texts also contains several languages other than English (LOTE).

The globalization of the business world has only increased the challenges we face in e-discovery. As we create more documents than ever before, the level of precision and formality of our written words seems to erode. When you consider the slang, abbreviations, vagueness, misspellings, acronyms, code words and foreign languages in a document collection, developing search criteria that will select the most responsive documents for attorney review is a very difficult task.

In e-discovery, search results with good precision return documents with a high responsiveness rate, while search results with good recall have fewer items left out of review

These challenges have placed a spotlight on the effectiveness of utilizing keyword or search criteria to reduce document populations that need to be reviewed, and there have been numerous court opinions recently that agree. For example, Judge Grimm in *Victor Stanley Inc. v. Creative Pipe, Inc.*[1], opined: "all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search."

Judge Grimm pointed out that searching is important and requires skill. He stressed:

> Selection of the appropriate search and information retrieval technique requires careful advance planning by persons qualified to design effective search methodology; the implementation of the methodology selected should be tested for quality assurances; and the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that is it appropriate for the task, and show that it was properly implemented.

In *William A. Gross Construction Associates, Inc. v. American Manufacturers Mutual Insurance Co.*[2], Judge Peck suggested that:

> where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of "false positives." It is time that the Bar—even those lawyers who did not come of age in the computer era—understand this.

Litigants are faced with ballooning volumes of ESI along with judicial mandates advocating defensible data reduction processes. For the increasingly cost conscious legal teams, the remedies are currently limited. Technology simply allowing reviewers to review faster is not the answer. On the other hand, the earlier use of a comprehensive data targeting methodology leveraging advanced technological tools and human judgment to build, test, validate and execute sophisticated search criteria to target the most relevant set of documents for review results in a more efficient, defensible and cost effective review.

Search Criteria Development and the Data Analytics Model

Many e-discovery service providers routinely tout the speed with which their technology platform allows an army of contract attorneys to speed through a document review making document designations on large volumes of documents. Many of these companies cite statistics that display the number of documents per hour (or day) that an attorney can review (e.g. 2,500/docs/day/reviewer). While these appear to be useful statistics, in reality they focus on the wrong numbers. Such high rates of review necessarily depend upon the review collection containing a high percentage of non-responsive documents. To focus on speed alone, without regard for the ultimate responsive rate, is a misguided approach. The more documents that are designated as non-responsive, the less value you get for each review dollar spent.

> "Since most of the cost of e-discovery lies in review expenses, estimates range from 50% to 80%, our efforts should be focused on searches that reduce the amount of ESI to be reviewed. Obviously, the better the search, the more chaff is separated from the wheat."[3]

Is placing an emphasis on the speed of review the appropriate tactic and correct use of technology to tackle the ever growing amount of ESI and the need to reduce the cost of review? Or should the focus be on the defensible reduction of data using search technology prior to review?

In the context of data targeting and volume reduction, one cannot discuss search strategy without addressing *precision* and *recall*. In e-discovery, search results with good precision return documents with a high responsiveness rate, while search results with good recall have fewer items left out of review. These are critical factors and the goal should be to achieve a *balance* of precision and recall. The result is a review set that contains more responsive documents with fewer responsive items missed, as well as a reduced volume of non-responsive items that need to be re-

viewed. Achieving this requires a transparent process built around technology, teamwork and validation.

The data analytics process for developing effective search criteria typically begins with the terms created by the legal team. These terms are tested against a sample of documents from the collected corpus, and research is conducted using publicly available data and existing case materials. Automated analytic tools may also be used to identify potential modifications or additions to the search criteria. Statistical analysis of the search results along with the review determinations made on the sample data sets are then used to iterate through multiple versions of search criteria until the legal team is comfortable with the results, and a high percentage of the items returned by the criteria are responsive. By participating in this iterative criteria development process, the case team has the opportunity to learn about documents that relate to their case before review even begins.

This is not the end of the process. As Judges Grimm and Peck noted, a defensible process for reduction of data through searching also needs to include validation steps for quality assurance. There are different types of validation that may be used throughout the data analytics process. While the search criteria are being developed and the iterative approach is generating new search criteria sets, samples of items that were responsive to an earlier search criteria set but have dropped from the current search results are reviewed. This "dropped item" validation ensures that responsive documents are not being left behind by revisions made to the search criteria.

Items not returned by search criteria are also sampled and reviewed. If responsive items are found during this "non hit" validation, the search criteria must be refined in such a way that these items will be returned by the searches and moved into review. If no responsive items are found, then several iterations of non-hit analysis will provide quality assurance that the search criteria are appropriately selecting items for review.

In addition to ensuring that the right documents are moving forward into review, steps need to be taken to avoid overly broad search results and review sets. There needs to be a method for obtaining "feedback" from the document review. If certain search criteria are yielding large volumes of non-responsive documents, then modifications will be made to the search criteria to limit or exclude non-responsive items from review.

Conclusion

The Sedona Conference advises:

> [M]erely adopting sophisticated automated search tools, alone, will not necessarily lead to successful results. Lawyers must recognize that, just as important as utilizing the automated tools, is tuning the process in and by which a legal team uses such tools, including a close involvement of lead counsel. This may require an iterative process which importantly utilizes feedback and learning as tools, and allows for measurement of results. The time and effort spent on the front end designing a sophisticated discovery process that targets the real needs of the client must be viewed as a condition precedent to deploying automated methods of search and retrieval.[4]

Rather than relying solely on technology that aids in review efficiency but may still be cost prohibitive by the sheer volume of items that require review, the data analytics process targets a higher percentage of responsive documents while taking steps to ensure that responsive items are not missed and, at the same time, reduces the volume of non-responsive items selected for review. This results in cost and time savings for document review as well as increased defensibility surrounding the search criteria and data reduction process.

1. 250 F.R.D. 251 (D. Md. 2008).
2. 256 F.R.D. 134 (S.D.N.Y. 2009).
3. Hundredth Blog: *Thoughts on SEARCH and Victor Stanley*, Inc. v. Creative Pipe, Inc., in http://ralphlosey.wordpress.com/2008/06/08/hundredth-blog-thoughts-on-search-and-victor-stanley-inc-v-creative-pipe-inc.
4. *The Sedona Conference's Best Practices Commentary on Search and Retrieval Methods*, 8 THE SEDONA CONFERENCE 189, 199 (2007). http://www.thesedonaconference.org/dltForm?did=Best_Practices_Retrieval_Methods___revised_cover_and_preface.pdf

Linda Beaudoin is a member of the Daegis consulting team specializing in Data Analytics. Ms. Beaudoin is responsible for the design and implementation of the Data Analytics process, including consultations with clients, execution of Data Analytics strategies, improving and developing the process, quality control and related documentation. In her capacity as Vice President, Sheila McKay is responsible for the success of Daegis' Professional Services group, which includes Consulting, Client Services and Project Management. Ms. Mackay also plays a critical role in the development of DocHunter and general business strategies such as marketing and strategic planning. Sheila's extensive background in the legal industry has provided her with keen insight into the E-Discovery challenges faced by large law firms as well as solutions providers.

a defensible process for reduction of data through searching also needs to include validation steps for quality assurance