Building A Defensible Search And Review Process For ESI

Deborah H. Juhnke
is an e-discovery consultant at Husch Blackwell Sanders LLP, in Kansas City, Missouri. She assists clients with e-discovery planning, early case assessment, data collection, and project management, as well as litigation readiness planning, and data and practice mapping. She has e-discovery experience in a wide array of cases, including products liability, trade secret theft, employment law, contractual disputes, bankruptcy, shareholder class action, and antitrust litigation matters. She is a frequent speaker and has written extensively on ESI topics. Before joining Husch Blackwell Sanders, she was Vice President of Computer Forensics Inc. in Seattle, Washington, where she served as a senior consultant on computer forensic and litigation readiness projects. Ms. Juhnke is a member of the HTCIA (High Tech Crime Investigative Association), The Sedona Conference Working Group I, and ARMA. The author gratefully acknowledges the assistance of Dr. Janet Thornton of the ERS Group in clarifying the statistical approaches suggested in this paper. The author can be reached at deborah.juhnke@huschblackwell.com.

Deborah H. Juhnke

It's time to go beyond the keyword search.

EVOLVING CASE LAW has made the accuracy and efficiency of keyword searching a hot topic. Recent rulings from Magistrate Judge John M. Facciola (United States District Court, Washington, D.C.) and Magistrate Judge Paul W. Grimm (United States District Court, District of Maryland) have called into question whether the current practice of running "best guess" keywords and calling it good serves our clients and the justice system. Though no system (including, especially, full manual review) is perfect, opposing parties and the courts will increasingly demand better and more defensible processes.

Groups such as the Sedona Conference, http://thesedonaconference.org, the TREC Legal Track, http://trec-legal.umiacs.umd.edu, and numerous e-discovery luminaries are wrestling with how best to address this problem and find an optimum solution, especially in light of the mushrooming volume of ESI. In the meantime, however, life goes on and lawyers and their staff personnel must devise good-faith methods to ensure that discovery yields complete and correct results — at least as close to that ideal as is possible. What follows is an interim model designed to address some of the key criticisms of the simple keyword search: that it is both under- and over-inclusive, yielding inaccurate results that are expensive to achieve; and that there is a general lack of quality control in the process.

The model is designed to be straightforward, understandable, and repeatable with little to no expert statistical assistance. There is no silver bullet, and there may be cases that require a more sophisticated approach, but attention to these techniques can move lawyers in the right direction.

THE FIVE I'S • The model assumes five stages of activity required to yield the desired result: intention, investigation, interaction, implementation, and iteration. Careful consideration of each of these phases is a necessary component of a repeatable and scalable process to guide search and review of discovery documents.

Intention

It is important first to understand the goal of search and review. Is the strategy to be restrictive or over-inclusive? Is the review of client data or data from the opposing party? Much will depend on the client, the nature of the dispute, the posture of the opposing side, the volume of raw ESI, any existing court rulings, and the value of the matter. For example, a goal of over-inclusive results suggests that searches will be less specific and the review more cursory. A goal of restrictive results will lead to continued refinement in search strategies, multiple searches, and a more complex review.

Investigation

Investigation takes many forms in this context. It includes a careful review of known information, such as pleadings, key documents, witnesses, data types and sources, and unexpected terms or code words. The goal, of course, is to develop a complete and nuanced semantic understanding of the potential document population. The time and care devoted to this phase will both streamline and improve the ultimate result. As the investigation progresses, lists will evolve: lists of witnesses, lists of keywords, lists of common misspellings, lists of acronyms and/or code words, etc. Taken together, this information will guide the first foray into the raw document population.

Interaction

Interaction with the opposing side is key to ensuring acceptance of the search and review plan. In addition to the obvious benefit of helping to avoid wasteful motion practice, collaboration with opposing counsel can sometimes serve to limit the scope of the search. Even if complete agreement is not reached, any attempt to do so is often a positive and good-faith step in the eyes of the court.

Implementation

The tools available to implement a search strategy are many and varied. Their selection can depend on many of the same issues addressed during the goal-setting phase. Is a winnowed review set desirable, or is the focus on gathering as much as possible? How intensive a review is needed? What is the document population? If mostly email, that may point to a different solution than if the population is primarily mixed MS-Office documents, such as spreadsheets. Although there are dozens of variants, essentially there are three solutions:

- A simple Boolean search tool, such as Summation or Concordance;
- A more sophisticated relational database review tool such as CaseLogistix or Ringtail; and
- One of the concept/analytic/clustering tools, such as Autonomy, Recommind, or Attenex.

Each of these solutions can be the right answer, depending on the circumstance. Understanding the tools available and their relative strengths and weaknesses is fundamental to achieving a defensible process.

Iteration

Feedback loops are a necessary component of quality control in data search and review. Whether through Boolean searches or concept analytics,

computers can be used to create subset collections for manual review. However, the results of these subset-creating searches are rarely examined specifically to determine whether the keywords or assumptions were correct. Iteration serves two purposes: first, to illuminate alternative terms or phrases that can make the collection more accurate, and second, to verify — typically through sampling — that the collection and review have been complete.

SAMPLING IN DATA COLLECTION AND REVIEW • Volumes have been written on the topic of sampling generally, and the statistics behind the types of sampling most appropriate for any given application are equally daunting. Nevertheless, Judge Grimm suggested that attorneys must be held accountable for the accuracy of their data collection and review: "The only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents…in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive." *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D. 251, 257 (D. Md. 2008).

There is, however, a statistical shorthand that we may put to good use. Combining rudimentary statistical sampling with a process-oriented, stepwise approach to search and review will yield a more reliable and defensible result than is typically achieved today.

Deciding How Large A Sample Should Be

Surprisingly, sample sizes need not be as large as one might think. For the purposes of this model, the calculation will rely on the following factors and assumptions:

- The sample will be based on a proportion (responsive), rather than an average;
- The sample will be selected without replacement. (In other words, when a selection is made, it will not be replaced into the population, thereby making it unavailable to be selected again);
- The "degree of accuracy/margin of error" will be five percent;
- The "confidence level" will be 95 percent.

A note about the last two assumptions. Some may extrapolate these numbers to point out that hundreds or thousands of documents will be missed. While true, it is also true that there is a high-cost, yet diminished, benefit to be gained by driving these numbers closer to the ideal of 100 percent with zero error. Sample sizes must be significantly higher, resulting in greater costs and time for review. Further, it has been well-documented that manual search and review is so inferior to any automated process as to make a discussion of 95 percent vs. 99 percent virtually irrelevant. David Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System*, 28 Commc'ns. A.C.M. 289 (Mar. 1985).

Given the above assumptions, a simple chart, such as one found here and excerpted below from http://research-advisors.com/tools/SampleSize.htm, may be used to determine how many documents should be sampled to be representative of the population. For example, for a population of 25,000 documents only 378 would need to be sampled. For a population of 1 million documents a sample size of 384 would suffice. Conversely, however, for a population of 500, the sample size required is 217 documents.

Required Sample Size

Population Size	Confidence = 95% Margin of Error				Confidence = 99% Margin of Error			
	5.0%	3.5%	2.5%	1.0%	5.0%	3.5%	2.5%	1.0%
10	10	10	10	10	10	10	10	10
20	19	20	20	20	19	20	20	20
30	28	29	29	30	29	29	30	30
50	44	47	48	50	47	48	49	50
75	63	69	72	74	67	71	73	75
25,000	378	760	1448	6939	646	1285	2399	9972
50,000	381	772	1491	8056	655	1318	2520	12455
75,000	382	776	1506	8514	658	1330	2563	13583
100,000	383	778	1513	8762	659	1336	2585	14227
250,000	384	782	1527	9248	662	1347	2626	15555
500,000	384	783	1532	9423	663	1350	2640	16055
1,000,000	384	783	1534	9512	663	1352	2647	16317

Generating A Random Sample

The sample for this application is best defined as a "random sample" — one in which every item (document) has an equal chance of selection. (See discussion below, however, regarding stratification.) A random sample may be generated simply through the use of a random number generator, such as the one found at http://www.random.org/integers/. For purposes of litigation documents, which commonly now have an MD5 or SHA1 HASH associated with them, as long as the HASH value does not have an inherent bias, it may be sorted and the top items in the list selected — a poor man's random number set.

Stratification of Sample

One last point: There are situations in which stratification of the sample may be desirable. For example, given a list of 20 custodians, it may be best to stratify them into two or more groups of "key" custodians, so that random samples are generated for each group and each group is thereby adequately represented and tested. This is particularly useful when the likelihood of finding key words, etc. is thought to vary by custodian. It also provides greater precision when estimates are made regarding characteristics of the entire document population, and provides a more defensible process.

SAMPLING FOR QUALITY CONTROL OF SEARCH AND REVIEW: SIX STEPS • Here are the six steps to guide sampling for quality control of the search and review:

- Step 1. Work through the Intention, Investigation, Interaction, and Implementation tasks outlined above to develop lists of custodians, keywords (phrases, abbreviations, concepts), and temporal boundaries;
- Step 2. Apply this information in the search tool of choice to yield a subset of data for review;
- Step 3. From this group, create an appropriately sized sample set (using the Sample Size Chart noted above) to test for precision of the search terms. (Precision refers to whether only the relevant documents were retrieved by the search. Low precision means that a large

number of irrelevant documents have been retrieved.) In other words, are the search terms retrieving what they should retrieve? If there are terms that retrieve a high proportion of irrelevant documents, they should temporarily be removed from the search term list. The examiner of the sample set should also be on the lookout for new or different terms that appear in the documents and that would also be relevant search terms and which should be added to the list;

- Step 4. Apply the newly updated search terms to the original data set a second time;
- Step 5. Repeat Step 3 by creating a sample set to test again for precision. (Although Steps 3 and 4 may be repeated ad infinitum, there will soon come a point when the incremental improvements are too small to make it worthwhile);
- Step 6. The last step in sampling for quality control involves sampling what is left behind to ensure that the recall has been sufficient. "Recall" refers to whether all the relevant documents were retrieved by the search, irrespective of any false hits. Low recall means that regardless of the size of the search result, many relevant documents were not found. Create an appropriately sized sample set from the documents not retrieved by the search. Review these to identify whether there are any relevant documents that were not found using the search terms. If relevant documents are found, note any terms or characteristics that, if used, would have included them in the result set. Adjust the search criteria appropriately and re-run.

Document, Document, Document

There may come a time when the processes and procedures used to select, review, and produce documents are questioned. To prepare for this eventuality, each of the above steps should include some level of documentation regarding the lists used, sampling results, adjustments made, and decisions reached. What is easily remembered and written concurrent with the activity can be nearly impossible to recreate six months or a year down the road.

A STEP IN THE RIGHT DIRECTION • The message is clear: Attorneys must work toward improved processes for search and review. However, it is the rare case or investigation that can cost-justify the support of experts in statistical sampling to achieve the improvement. Further, as much as we would like to believe human review is infallible, it is not, and there are undoubtedly other ways to measure and improve quality control. Using these suggestions and techniques, however, will help attorneys become more informed about their document collections and offer a higher level of confidence than is common today that the expected results are actually achieved.

To purchase the online version of this article, go to www.ali-aba.org **and click on "Publications."**

PRACTICE CHECKLIST FOR
Building A Defensible Search And Review Process For ESI

- Know the five I's of document searching that goes beyond keyword searching: intention, investigation, interaction, implementation, and iteration:

__ Clarify the intention: the goal of search and review. Is the strategy to be restrictive or over-inclusive? Is the review of client data or data from the opposing party? Much will depend on the client, the nature of the dispute, the posture of the opposing side, the volume of raw ESI, any existing court rulings, and the value of the matter;

__ Investigation includes a careful review of known information, such as pleadings, key documents, witnesses, data types and sources, and unexpected terms or code words. The goal is to develop a complete and nuanced semantic understanding of the potential document population. The time and care devoted to this phase will both streamline and improve the ultimate result;

__ Interaction with the opposing side has the obvious benefit of helping to avoid wasteful motion practice, and can sometimes help to limit the scope of the search;

__ In implementing the search strategy, keep in mind that although there are dozens of variants, essentially there are three solutions: a simple Boolean search tool, such as Summation or Concordance; a more sophisticated relational database review tool such as CaseLogistix or Ringtail; and one of the concept/analytic/clustering tools, such as Autonomy, Recommind, or Attenex;

__ Iteration. Feedback loops are a necessary component of quality control in data search and review. Whether through Boolean searches or concept analytics, computers can be used to create subset collections for manual review.

- Sample sizes do not have to be especially large to yield high degrees of accuracy and confidence levels. Be sure to consult a statistical sample size chart.