Finding a Needle in an Electronic Haystack: The Science of Search and Retrieval

By Ronald J. Levine and Susan L. Swatski-Lebson

Ronald J. Levine

Susan L. Swatski-Lebson

The volume of electronically stored information (ESI) generated by corporations and individuals is growing exponentially; an estimated 100 billion emails are generated daily.[1] This ESI explosion poses new challenges for litigation practitioners engaged in discovery and review in mass tort litigation. The complexities of searching gigabytes of data has rendered litigation practitioners' tried and trusted methods of review unsustainable, leaving them grappling to strike a balance among the quantity of data, the reviewer's capacity, and the clients' budgets. At the cornerstone of this balancing act are automated discovery search tools and methodologies employed to identify, filter, cull, categorize, and ultimately produce responsive ESI.

Most litigators utilize Lexis or Westlaw searches as part of their daily practice and enjoy a certain level of comfort in terms of ease of use and obtaining their desired results. However, those same litigators may find themselves in the midst of a minefield when confronted with the complexities of searching and culling ESI to comply with discovery obligations. Employing the same general thought process and search methodology that one may employ to run a successful search on Lexis is not likely to render comprehensive results in the discovery of ESI.

Two recent U.S. district court opinions address a lawyer's obligations with respect to search methodology in the discovery of ESI: *Victor Stanley, Inc. v. Creative Pipe Inc*.[2] and *United States v. O'Keefe*.[3] Together these opinions offer useful guidelines and best practices for e-discovery to help litigators navigate through search methodology and statistical sampling pitfalls while raising the bar for what courts expect of lawyers in the discovery of ESI. These cases serve as a wake-up call to litigation practitioners who have been relying solely on traditional keyword searches to identify and cull responsive and privileged documents, and further bolster the notion that ESI discovery is a science that requires some degree of technical expertise to produce a search that is capable of sustaining a court challenge to its methodology.

In this article, we first discuss search methodology, including the pros and cons of the most common search tools, as well as initiatives by various governmental bodies and the legal community to develop procedures and principles to help litigation practitioners evaluate the methodologies and to develop best practices. Next, we examine Magistrate Judge Grimm's opinion in *Victor Stanley* and Magistrate Judge Facciola's opinion in *O'Keefe*, and the implications of these decisions for litigators. Last, we provide practice pointers for litigation practitioners engaging in the discovery of ESI.

What Is Search?

Given the limitations of attorney review capacity and client budgets, review of every piece of electronic data is rarely feasible. As a result, litigation practitioners turn to various tools and methodologies to help identify, cull, and categorize data for the purposes of responsiveness, relevance, privilege, and confidentiality. This process is commonly referred to as search.

The problem is that searching is not an exact science; rather, it is a learned skill of some complexity. Context is everything in employing an effective search of ESI; therefore, the most effective method of searching in one case may not be effective in another case. Generally, the more complex the case and the greater the volume of data, the more likely multiple search techniques should be utilized.

Since the Federal Rules of Civil Procedure were amended in 2006 to address the growing presence of ESI in discovery, a cottage industry has sprung up to assist litigators in meeting their e-discovery obligations. This industry, in large part, focuses on helping counsel and their clients implement effective search methodologies by reducing the number of false positives and negatives. Although merely hiring an outside search expert does not insulate the attorney or the client from sanctions if the search methodology is challenged or substantive discovery gaffes are identified, many litigators find it helpful to consult with or to retain an e-discovery expert in complex matters involving a high volume of data. Whether or not a search expert is retained, litigators would be well advised to become familiar with the newest tools and methods of searching as well as the requirements and guidance provided in case law regarding the search of ESI.

Search Tools and Methodologies

The following is a discussion of the most common types of search methodologies.

Keyword searches. Keyword searches employ a broad, natural language search strategy that allows the lawyer to locate data containing a word or a combination of words in designated fields. Even though a keyword search generally retrieves the greatest number of records, employed alone it may be inadequate because by ignoring context entirely, it can result in an unacceptably high percentage of false-positive records.[4]

Boolean searches. Boolean searches are performed by identifying keywords

that appear in a specified relation to one another by employing such terms as "and," "or," "within," and "not" to refine the scope of the search.[5] Results from Boolean searches may only be marginally better than a keyword search because the search will only identify ESI containing the exact specified terms.

A common type of search performed by litigation practitioners is a combination of a keyword and a Boolean search.[6] Even this combination may fail to catch documents using words that are close, but not identical, to the specified search terms, such as nicknames, initials, misspelled words, synonyms, and abbreviations. Moreover, although employing more search terms may reduce the risk of missing relevant ESI, it does so at the price of increasing the number of false positives retrieved. A high percentage of false positives is a potentially serious problem, because practitioners must then manually review the search results to separate the wheat from the chaff as they make responsiveness, privilege, and confidentiality determinations. As a result, practitioners employing only keyword and Boolean searches face the difficult task of striking a balance between being unduly restrictive and missing responsive documents versus overbreadth that drives up review costs.

Taxonomy tools. Taxonomy tools categorize documents containing words that are subsets of relevant topics (e.g., if one of the topics of interest is cats, a taxonomy tool would capture documents that mention Siamese, Himalayan, and Persian).[7] The only relations included in a taxonomy are inclusion relations; lower terms in the taxonomy are subclasses of higher terms in the taxonomy.[8]

Ontology tools. Ontology tools are a more generic species of taxonomy tools. Subsets of relevant topics are searched, but the search is not limited to identifying subset relationships. For example, if one of the topics of interest is cats, then an ontology tool would identify documents that mention kennels or veterinarians.

Statistical clustering. Statistical clustering is "the process of grouping together documents with similar content" based on "the number of words that overlap between each pair of documents."[9] Clustering compares each document in a pool to previously identified, relevant "seed" documents. Clustering may be an effective and economical choice to be utilized as a first pass through data because it requires no human intervention to design.[10]

Practitioners utilizing any combination of search tools and methodologies may still be required by a court to defend their rationale for implementing that search.

Bayesian classifiers. Bayesian classifiers use probability theory to make informed assumptions about the relevance of documents based on the system's prior experience in capturing relevant documents.[11] This is accomplished by assigning a value to words, proximity, and frequency. The resulting values can then be used to rank documents based on their computed "relevancy."[12] Bayesian classifiers, however, assume that every word in a document is independent of every other word; therefore, they cannot detect the interrelationship among words.

Other common searches utilize pattern-matching techniques that enable the identification of naturally occurring patterns in a text based on the usage and frequency of words or terms that correspond to specific concepts. Also, some searches may augment traditional Boolean searching with mathematic algorithms to connect concepts based on the definition or usage of a term.

Sampling. Once a search is complete, practitioners should perform a random sampling of various categories of documents to test the reliability of the search. Absent sampling, a litigator cannot reasonably establish that the categories of documents (responsive, nonresponsive, privileged, or confidential) are over- or under-inclusive.[13]

Whatever search tools or methodologies are utilized, the counsel and client must be actively engaged in designing and implementing the search in order to take into account industry jargon, syntax, and the semantic relationships behind the relevance of the terms. Merely implementing search tools beyond blunt keyword searches does not, in and of itself, give counsel and her client a get-out-of-jail-free card against a subsequent challenge to the search. Practitioners utilizing any combination of search tools and methodologies may still be required by a court to defend their rationale for implementing that search.[14]

In recognition of the challenges that ESI often poses in the discovery review process, the National Institute of Standards and Technology and the Department of Defense are conducting scientific evaluations of the effectiveness of various kinds of ESI search methodologies. This project is known as the Text Retrieval Conference (TREC). The goal of this research effort "is to create industry-specific practices for use in electronic discovery."[15] TREC is expected to identify both cost-effective and reliable search-and-information retrieval methodologies and to make best practices recommendations. In *Victor Stanley*, the court noted that a practitioner's adherence to TREC's practice recommendations "would support an argument that the party employing them performed a reasonable ESI search, whether for privilege review or other purposes."[16]

The Sedona Conference, an educational institute dedicated to moving the law of complex litigation forward, has also been actively involved in helping litigators develop defensible search methodologies. In August 2007, The Sedona Conference released its *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* (the Sedona Conference Commentary). The goal of the Sedona Conference Commentary is to provide "the bench and bar with an educational guide" to increase the accuracy and

efficiency of searches for responsive ESI.[17] The Sedona Conference Commentary found that:

> although basic keyword searching techniques have been widely accepted both by courts and parties as sufficient to define the scope of their obligation to perform a search for responsive documents, the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages).[18]

In *Victor Stanley*, the court cited to the Sedona Conference Commentary, stating adherence to its principles and recommended practices "will go a long way towards convincing the court that the method chosen was reasonable and reliable."[19]

Two Key Recent Cases Discussing Search

Victor Stanley v. Creative Pipe

On May 29, 2008, in a groundbreaking opinion in *Victor Stanley*, U.S. District Court Magistrate Judge Paul W. Grimm found that the defendants waived any privilege or work product protection they may have asserted to 165 electronically stored documents when those documents were inadvertently disclosed, because defense counsel failed to take reasonable precautions while performing their privilege review.[20] Judge Grimm's opinion provides a detailed analysis of the methods of ESI search as well as guidance on what a litigator may be required to prove if the reasonableness of counsel's search methods are challenged.

The parties in *Victor Stanley* agreed to a joint protocol to search for responsive ESI. To search for privileged documents using the joint protocol, the defendants gave their forensic computer expert a list of 70 keyword search terms, which were selected by defense counsel and one of the defendants.[21] The computer expert did not assist in developing the search strategy, which was a linear keyword search, but rather merely ran the search.[22] In conducting their privilege review, defendants relied on the results retrieved from the 70 keyword searches and only reviewed the title page of some of the documents without actually reviewing their content.[23] Shortly after production, counsel for the plaintiff identified 165 potentially privileged and confidential documents. The plaintiff then moved for a ruling that the 165 documents were not protected by any privilege.[24]

To decide whether the defendants waived any potentially applicable privilege, the court employed an "intermediate test," which requires:

> the court to balance the following factors to determine whether inadvertent production of attorney-client privileged materials waives the privilege: (1) the reasonableness of the precautions taken to prevent inadvertent disclosure; (2) the number of inadvertent disclosures; (3) the extent of the disclosures; (4) any delay in measures taken to rectify the disclosure; and (5) overriding interests in justice.[25]

Based on applying the intermediate test, the court found that the defendants waived any attorney-client privilege or work product protection, because the defendants failed to provide the court with an adequate rationale for selecting the 70 keyword search terms and to identify the search terms.[26]

The *Victor Stanley* court reasoned that "[a]ll keyword searches are not created equal; there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search or relying exclusively on such searches for privilege review."[27] In so deciding, the court relied upon recent decisions in *Equity Analytics, LLC v. Lundin*[28] and *In re Seroquel.*[29] In *Equity Analytics*, the U.S. District Court for the District of Columbia found that "determining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a layperson (and a lay lawyer)."[30] In *Seroquel,* the U.S. District Court for the Middle District of Florida criticized the defendant's use of a keyword search in selecting ESI for production where the defendant failed to provide information "as to how it organized its search for relevant material, [or] what steps it took to assure reasonable completeness and quality control."[31] In *Seroquel*, the court explained that "while keyword searching is a recognized method to winnow relevant documents from large repositories . . . [c]ommon sense dictates that sampling and other quality assurance techniques must be employed to meet requirements of completeness."[32]

The court in *Victor Stanley* also found the following facts persuasive in reaching its decision:

- The defendants initially asked for a clawback agreement and then withdrew their request, citing they would be able to do a document-by-document privilege review.[33]
- The plaintiff discovered the privileged documents using a readily available desktop search tool in about one hour, immediately segregated the documents, and notified the defendants.[34]
- There was a large number of allegedly privileged documents that were produced.[35]
- These privileged documents at issue should have been easy to find because many were substantive and comprised of communications between the defendants and their counsel.[36]
- The defendants did not conduct any sampling of the text-searchable ESI documents that were determined not to contain any privileged information.[37]

United States v. O'Keefe

In *O'Keefe*, Magistrate Judge John M. Facciola of the U.S. District Court for the District of Columbia rejected a challenge to the government's production of ESI on the grounds that the government's search terms were inadequate. The court further found that discovery of ESI was not exempt from the rules governing scientific and other expert

evidence.[38] Magistrate Judge Facciola reasoned that "[w]hether 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," and that search-term efficacy is "beyond the ken of a layman and requires that any such conclusion be based on evidence that, for example, meets the criteria of Rule 702 of the Federal Rules of Evidence [FRE]." Therefore, for the defendants to challenge the search terms used by the government, the court found that their challenge had to be based on evidence that met the requirements of FRE 702.[39]

Implications of Victor Stanley *and* O'Keefe

Document searching for counsel is no longer a simple exercise of compiling a list of a few keywords to look for. What is the practical effect of the recent decisions in *Victor Stanley* and *O'Keefe* for litigators? Counsel must treat information search and retrieval as a science. As such, litigators must be prepared to substantiate that reliable tools and methodologies were utilized and implemented in their searches. To accomplish this, using an expert or a person with significant experience in searching and harvesting ESI may be necessary. A person with technical knowledge (e.g., a search retrieval expert, statistician, or computer scientist) would need to be utilized to design and implement the search methodology if counsel wants to build a solid defense against an ESI search challenge. The party mounting a challenge against an adversary's ESI search tools and methodologies likewise should anchor that challenge on expert opinion about the retrieval and statistical science employed as well as the party's rationale for that decision.

Thus, we find courts determining that the use of experts will be more helpful than relying on mere layperson opinions in making "factual determinations involving disputed areas of science, technology, or other specialized information." Consequently, a defensible search methodology is one that utilizes techniques that have been subject to peer review and approval by those with expertise in the science of information retrieval.[40]

Only time will clarify what types of tools and methodologies are in fact deemed "scientific." In the interim, litigators may look to the U.S. Supreme Court's opinion in *Daubert v. Merrell Dow Pharmaceuticals*[41] for guidance. The Court in *Daubert* identified the following four criteria to determine if a method is scientific: (1) falsifiability—can the methodology be tested, (2) peer review—have peers reviewed the methodology and commented on its validity, (3) testing—what is the error rate of the methodology, and (4) scientific acceptance—has the methodology been accepted by the e-discovery community.[42]

> Effective utilization of scientific methodologies and statistical analysis may ultimately reduce the costs of conducting discovery by retrieving fewer false positives.

Given the exorbitant costs of litigation in general and the skyrocketing costs of discovery of ESI in particular, many litigators may be concerned by the prospect of having to hire yet another expert. Indeed, these cost concerns may compel litigants to prematurely settle their case to avoid costly e-discovery. Magistrate Judge Grimm addressed ESI cost concerns as follows:

> For those understandably concerned about keeping discovery costs within reasonable bounds, it is worth repeating that the cost-benefit balancing factors of Fed.R.Civ.P. 26(b)(2)(C) apply to all aspects of discovery, and parties worried about the cost of employing properly designed search and information retrieval methods have an incentive to keep the cost of this phase of discovery as low as possible, including attempting to confer with their opposing party in an effort to identify a mutually agreeable search and retrieval method. This minimizes costs because if the method is approved, there will be no dispute with resolving its sufficiency, and doing it right the first time is always cheaper than doing it over if ordered to do so by the court.[43]

Magistrate Judge Grimm added: "as search and information retrieval methodologies are studied and tested this will result in identifying those that are affected and least expensive to employ for a variety of ESI discovery tasks."[44]

Effective utilization of scientific methodologies and statistical analysis may ultimately reduce the costs of conducting discovery by retrieving fewer false positives and also reducing the risk of responsive information being lost by identifying fewer false negatives. Consequently, although litigators may need to alter their approach to e-discovery, this new approach may ultimately prove to increase efficiency of the discovery process. At least, one can hope.

Practice Pointers

Given the sheer volume of ESI that must be searched, filtered, and reviewed, litigators would be well advised to consider the guidance and practices in the *Victor Stanley* and *O'Keefe* opinions when deciding how to undertake discovery of ESI and how to evaluate an adversary's ESI discovery methods.

The following are practical tips to develop a defensible search of ESI:

- Take time to understand the client's information architecture.
- Develop and train attorneys in e-discovery best practices.
- Recognize limitations in terms of counsel's technical knowledge of search and retrieval tools and methodologies, the client's budget, and the complexity of the underlying subject matter.
- Sole reliance on manual search processes to retrieve and review responsive data is generally infeasible and unwarranted where the

use of automated search methods are available.[45]
- Differing search tools and methods will likely produce varying results. When feasible, consider running a variety of searches.[46]
- The selection of optimal search tools or methodologies is highly dependent upon the specific legal context in which they are sought to be used.[47]
- Consider consulting an information retrieval expert.
- But, "[u]ltimate responsibility for ensuring the preservation, collection, processing, and production of electronically stored information rests with the party and its counsel, not with the nonparty consultant or vendor."[48] In other words, sloppy e-discovery practices may lead to sanctions.[49]
- Exercise due diligence in choosing a particular information retrieval product or service from a vendor.[50]
- Collaborate with opposing counsel on search methodologies.
- Negotiate a formal discovery plan with adverse counsel and get court approval.
- Negotiate a clawback agreement for inadvertent disclosure of privileged materials.
- Absent collaboration with the adversary regarding search methodology, assume that a choice of search tool and/or methodology will be challenged in either a deposition, an evidentiary proceeding, or at trial. Be prepared to explain why such tools and/or methodology were utilized, including citation to credible sources recognizing their use.[51]
- Recognize that utilizing search tools and advanced methodologies does not guarantee retrieval of all responsive data due to language characteristics.[52]
- Conduct sampling or otherwise audit processed data to confirm completeness and accuracy, including spot checking discarded data.
- Document key information, decisions, agreements, and processes in each phase of e-discovery.

Ronald J. Levine and Susan L. Swatski-Lebson are a litigation partner and litigation associate, respectively, with Herrick, Feinstein LLP, with offices in New York, Newark and Princeton, New Jersey.

Endnotes

1. Peter Lyman and Hal R. Varian, *How Much Information?* (2003) www2.sims.berkeley.edu/how-much-info-2003.
2. 250 F.R.D. 251 (D.Md. 2008).
3. 537 F. Supp. 2d 14 (D.D.C. 2008).
4. The Sedona Conference Journal, *The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery*, 8 Sedona Conf. J., 200–202 (Aug. 2007) (hereinafter "The Sedona Commentary at __").
5. *Id*. at 202, 217.
6. *Id.* at 200–02.
7. *Id.* at 221–22.
8. *Id*. at 221.
9. *Id*. at 219.
10. *Id*.
11. *Id.* at 218–19.
12. *Id.* at 218.
13. Victor Stanley, 250 F.R.D. at 257 (finding that "the only prudent way to test the reliability of the keyword search is to perform some appropriate sampling of the documents determined to be privileged and those determined not to be in order to arrive at a comfort level that the categories are neither over-inclusive nor under-inclusive").
14. The Sedona Commentary at 204.
15. Victor Stanley, 250 F.R.D. at 261 n. 10.
16. *Id.*
17. The Sedona Commentary at 191.
18. *Id.* at 201.
19. Victor Stanley, 250 F.R.D. at 262.
20. *Id.* at 253–54.
21. *Id.* at 254–55.
22. *Id.* at 255–56.
23. *Id*. at 256.
24. *Id.* at 253.
25. *Id*. at 259.
26. *Id.* at 259–60.
27. *Id.* at 257.
28. 248 F.R.D. 331, 333 (D.D.C. 2008).
29. 244 F.R.D. 650, 660 n. 6, 662 (M.D. Fla. 2007).
30. *See supra* note 28, at 333.
31. Seroquel, 244 F.R.D. at 660 n.6.
32. *Id.* at 662.
33. Victor Stanley, 250 F.R.D. at 255.
34. *Id.* at 257.
35. *Id.* at 263.
36. *Id.*
37. *Id.* at 257.
38. O'Keefe, 537 F. Supp. 2d at 24.
39. *Id.*
40. Victor Stanley, 250 F.R.D. at 261 n.10.
41. 509 U.S. 579, 113 S. Ct. 2786 (1993).
42. Daubert, 509 U.S. at 594.
43. Victor Stanley, 250 F.R.D. at 261 n.10.
44. *Id.*
45. The Sedona Commentary at 194.
46. *Id*.
47. *Id.*
48. *In re Seroquel Prod. Liab. Litigation*, 244 F.R.D. 650, 664 n.14 (M.D. Fla. 2007) (quoting Sedona Principle 6.d); Zubulake v. UBS, 229 F.R.D. 422, 435 (S.D.N.Y. 2004) (finding that "[c]ounsel is responsible for coordinating her client's discovery efforts"); Cache La Poudre Feeds v. Land O'Lakes, 244 F.R.D. 614, 630 (D. Colo. 2007) (noting that "[c]ounsel retains an on-going responsibility to take appropriate measures to ensure that the client has provided all available information and documents which are responsive to discovery requests") (citation omitted).
49. Zubulake, 229 F.R.D. at 424.
50. *Id.*
51. *Id*. at 195.
52. *Id.* at 194.