Bloomberg Law Reports®

September 28, 2011

Technology Law

# Court Procedure & Technology

## Electronic Discovery

# Finding the Needle in the Electronic Haystack: Moving Beyond Keyword Searches to Find Electronically Stored Information

CARLTON FIELDS
ATTORNEYS AT LAW

Contributed by Rebecca N. Shwayri, Carlton Fields

In searching for electronically stored information ("ESI") that is relevant to litigation, lawyers have traditionally employed keyword searches to assist them in finding relevant electronic data. A leading study conducted by David Blair and M.E. Maron demonstrated that attorneys were only 20 percent effective at thinking of all the ways that the creators of a document could refer to words, ideas, or issues in a case.[1] Federal judges have also started to realize that keyword searches have limitations. At least one court has opined as to the necessity of expert witness testimony in validating the methodology of keyword searches.[2] Another court has found that keyword searches that are not sufficiently supported by a methodology can result in the waiver of attorney-client privilege for documents that are inadvertently produced.[3] A majority of the legal profession still uses keywords searches without realizing some of the drawbacks and risks. In light of some of the risks of keyword searches, lawyers need to be aware of some of the new types of methodologies that could assist them in locating relevant ESI. These new types of methodologies include Bayesian classifiers, fuzzy search models, and concept categorization tools. The utilization of these new methodologies, in combination with keyword searches and active lawyer insight and supervision of a case, can help mitigate against the risks of solely relying on keyword searches to find the needle in the electronic haystack.

### Drawbacks of Relying Solely on Keyword Searches

The Blair and Maron study demonstrates the problems of relying solely on keyword searches to find information. The study case involved a San Francisco Bay Area Rapid Transit ("BART") accident in which the BART train failed to stop at the end of the line.[4] The discovery database contained 40,000 documents and 350,000 pages.[5] In an effort to locate all relevant documents, the attorneys worked with paralegal specialists.[6] While the attorneys believed that they had found 75 percent or more of the relevant documents, a more detailed analysis showed that the number was only about 20 percent.[7] The parties used different terms to describe the same issue.[8] Furthermore, the term used to describe the faulty component that failed part varied depending upon the location in the country where the document was written.[9]

As the Blair and Maron study demonstrates, relying solely on keyword searches to find ESI has drawbacks. The party requesting ESI may not know all of the different terminology that the responding party may use to describe an issue. A keyword search may work well where the requesting party knows the precise terms of art used to describe the issue in the case. Many times, the keyword term may be over-inclusive or under-inclusive. A keyword may be over-inclusive where it identifies all documents relating to the term even though many of those documents have

Originally published by Bloomberg Finance L.P. in the Vol. 5, No. 38 edition of the Bloomberg Law Reports–Technology Law. Reprinted with permission. Bloomberg Law Reports® is a registered trademark and service mark of Bloomberg Finance L.P.

no relationship to the issues in the case. A keyword may also be under-inclusive because it may not include terms of similar meaning that have a great relevance to the case. Keyword searches can also be inadequate due to a failure in imagination. A party thinking up a keyword search may fail to include common misspellings of the term of art.

A keyword search may work well where the requesting party knows the precise terms of art used to describe the issue in the case. Many times, the keyword term may be over-inclusive or under-inclusive. A keyword may be over-inclusive where it identifies all documents relating to the term even though many of those documents have no relationship to the issues in the case. A keyword may also be under-inclusive because it may not include terms of similar meaning that have a great relevance to the case.

## Some Federal Courts Criticize Keyword Searches

Federal courts have started to opine as to the efficacy of keyword searches for locating ESI and some courts have required expert testimony to buttress a keyword search. For example, in *United States v. O'Keefe*, a federal magistrate judge in the District of Columbia found that "[w]hether search terms or 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the science of computer technology, statistics and linguistics."[10] Because of the complexity involved in keyword searches, the court found that a lawyer's opinion as to whether a keyword search would produce information "is to truly go where angels fear to tread."[11] The magistrate judge went so far as to require expert testimony to support the methodology of a keyword search because the topic "is clearly beyond the ken of a layman."[12]

In *Equity Analytics, LLC v. Lundin*, the same magistrate judge required an affidavit from a forensics examiner opining in detail as to how a keyword search would be conducted.[13] In that case, the plaintiff alleged that its former employee gained illegal access to ESI after he was fired.[14] The plaintiff claimed that the proposed keyword searches were inadequate because the former employee loaded a new operating system onto his Macintosh computer.[15] Reiterating his previous ruling from *O'Keefe*, the magistrate judge explained that "determining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a lay person (and a lay lawyer) and requires expert testimony that meets the requirements of Rule 702 of the Federal Rules of Evidence."[16] The magistrate judge required the plaintiff to produce an affidavit from its examiner explaining the impact of loading a new operating system upon the Macintosh computer and describing in detail how the search would be conducted.[17]

While one magistrate judge has went so far as to require expert witness testimony to buttress keyword searches, another judge has found that keyword searches that are not properly supported by the evidence may result in an inadvertent waiver of attorney-client privilege. In *Victor Stanley, Inc. v. Creative Pipe, Inc.*, the plaintiff filed a motion seeking a ruling that five categories of electronically stored documents produced by defendants were not exempt from discovery.[18] The defendants argued that the 165 documents were protected by the attorney-client privilege and work-product doctrine.[19] The parties had previously developed a protocol containing detailed search and information retrieval instructions, including nearly five pages of keyword/phrase search terms.[20] The search terms were designed with the aim of locating responsive ESI and were not aimed at identifying privileged or work-product protected documents within the electronic information.[21] After the ESI was identified, the defendants gave their computer forensics experts a list of keywords to be used to search and retrieve privileged and protected documents.[22] Defendants' counsel was concerned about the possibility of inadvertent disclosure of privileged documents, given the volume of documents to be produced, and requested that the court approve a clawback agreement.[23] The clawback agreement would have allowed counsel to claw back an inadvertently produced privileged document. The defendants' counsel later notified the court that a clawback agreement was unnecessary because defendants would conduct a document-by-document privilege review.[24]

Defendants subsequently conducted a privilege search using 70 different keyword search terms that had been decided upon by one of the parties and several attorneys.[25] The documents that were returned during the keyword searches were segregated for further review.[26] Nontext-searchable files were turned over for manual privilege review.[27] This manual privilege review consisted of reviewing the page titles of documents.[28] If a document's page title indicated that privilege might be applicable, the document was reviewed in its entirety.[29]

In analyzing whether the defendants had waived the attorney-client privilege, the court explained that the defendants had the burden of proving that their conduct was reasonable.[30] The defendants failed to provide the court with information regarding the keywords used, the rationale for their selection, the qualification of the persons involved in formulating the keyword searches, whether the search was a simple keyword search or a more sophisticated search, and the quality of the search's implementation.[31] The court noted that there are

"well-known limitations and risks associated with [keyword searches] and proper selection and implementation obviously involves technical, if not scientific knowledge."[32] "Use of search and information retrieval methodology . . . requires the utmost care in selecting methodology that is appropriate for the task because the consequences of failing to do so, as in this case, may be disclosure of privileged/protected information to an adverse party."[33] A party selecting a particular methodology must be prepared to explain the rationale for the method chosen to the court and show that the methodology was properly implemented.[34] Because the defendants failed to demonstrate that the keyword search was reasonable, the court found that the attorney-client privilege had been waived.[35] While the defendants had the opportunity to protect themselves through a clawback agreement, the defendants voluntarily abandoned their request for a court-approved non-waiver agreement.[36] If the defendants had not abandoned their request for a clawback agreement, they would have been protected from waiver.[37]

### Some Federal Courts Ask Lawyers to Think Beyond Keyword Searches

*O'Keefe, Equity Analytics, and Victor Stanley* demonstrate some of the drawbacks and limitations of keyword searches. In *O'Keefe* and *Equity Analytics*, the court stated that it would not blindly rely on conclusory statements from counsel that keyword searches were sufficient and went so far as to require an expert to opine as to the propriety of the keyword search. The unfortunate consequence of *O'Keefe* and *Equity Analytics* may be to impose additional costs on litigants by forcing them to hire linguistic and scientific experts just to make sure that the keyword terms being selected will meet with court approval. A scientific and linguistics expert may not be justified in all cases. *Victor Stanley* made keyword searches even more problematic where the keyword searches are not supported by a scientifically-validated methodology. In such a case, the keyword search may result in an inadvertent waiver of attorney-client privilege. In order to avoid such a waiver, lawyers should consider discussing and incorporating clawback agreements at an early phase of the litigation.

While keyword searches have their place in locating ESI, over-reliance on keywords has drawbacks as the Blair and Moran study points out. Given these drawbacks, lawyers should try to familiarize themselves with new methodologies that may help them locate ESI that is relevant to the issues in the case. Some of these newer methodologies include Bayesian classifiers, fuzzy search models, clustering, and concept categorization tools.

A Bayesian system sets up a formula that places a value on the relationship, proximity, and frequency of words. In contrast to the Bayesian system, fuzzy search models are an attempt to improve searches by going beyond the word. Through the fuzzy search technique, a word is reduced to its basic element. This technique goes beyond the word by matching all forms of the word. Clustering groups documents together based on similar content. While there are a number of ways to define similarity, a more common method is to count the number of words that overlap in various documents. Because clustering does not require human intervention, it may be an economical approach at initially organizing documents.

While courts have started to recognize drawbacks to keyword searchers, the second generation methodologies are relatively untested in caselaw. Because of the newness of these methodologies to the court system, lawyers must be actively involved in the e-discovery portion of their cases and cannot simply rely on second generation methodologies to identify relevant information for them.

Concept categorization tools can also be utilized to try to find similarly related words or ideas. Concept categorization tools may include thesauri, taxonomies, and ontologies. The structure of the taxonomy or ontology can be used as an organizational paradigm for a document collection. A party seeking ESI can develop rules that specify how documents with certain keywords are related to categories of information. The computer can then organize the documents. Concept search tools should not supplant keyword searches, but may play an important role if metadata has been gathered. When concept search tools are combined with metadata, it may be possible to reveal relationships within the document collection. For example, a party may be able to determine whether a custodian had contact with other custodians to discuss particular issues during a period of time.

While courts have started to recognize drawbacks to keyword searchers, the second generation methodologies are relatively untested in caselaw. Because of the newness of these methodologies to the court system, lawyers must be actively involved in the e-discovery portion of their cases and cannot simply rely on second generation methodologies to identify relevant information for them. Lawyers should make sure that the selection of a methodology is appropriately supported by expert evidence if the case necessitates it. Lawyers should also consider using sampling techniques on a smaller volume of data if the parties are dealing with a large volume of ESI. Utilization of sampling techniques will enable the parties to determine whether the methodology is actually locating relevant ESI. In order to mitigate against the risk of a privilege waiver, consider entering into a clawback arrangement at the inception of the litigation.

Finally, lawyers need to cooperate with their opposing counsel from the beginning of the case in order to identify topics and subjects that are relevant to the litigation. There is no way that the requesting party is going to be able to know all of the terms of art that the responding party may use to describe the issues in the litigation. Only the responding party has access to such information. In the American justice system, parties should be able to obtain all the facts related to the case and should not be able to hide behind cleverly hidden ESI to obfuscate issues. If cooperation and information-sharing are the basis of e-discovery, there is a higher likelihood that employing newer methodologies will actually result in finding more relevant ESI.

*Rebecca Shwayri is an associate at Carlton Fields and is a member of the firm's E-Discovery Team. She assists the firm and its clients in developing policies for the institution of litigation holds and the preservation of electronically stored information. She has also managed teams of attorneys and paralegals in large document productions involving the review and production of electronically stored information. Shwayri can be reached at rshwayri@carltonfields.com*

[1] The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, at 206, August 2007, *available at* http://www.thesedonaconference.org/content/miscFiles/publications_html.

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

[3] *Victor Stanley, Inc. v. Creative Pipe, Inc.,* 250 F.R.D. 251, 263 (D. Md. 2008).

[4] The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery, at 206, August 2007.

[5] *Id.*

[6] *Id.*

[7] *Id.*

[8] *Id.*

[9] *Id.*

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

[11] *Id.*

[12] *Id.*

[13] *Equity Analytics, LLC v. Lundin,* 248 F.R.D. 331, 333 (D.D.C. 2008).

[14] *Id.* at 332.

[15] *Id.*

[16] *Id.* at 333.

[17] *Id.*

[18] 250 F.R.D. 251, 253 (D. Md. 2008)

[19] *Id.*

[20] *Id.* at 254.

[21] *Id.*

[22] *Id.* at 255.

[23] *Id.*

[24] *Id.*

[25] *Id.* at 256.

[26] *Id.*

[27] *Id.*

[28] *Id.*

[29] *Id.*

[30] *Id.* at 259.

[31] *Id.* at 259-260.

[32] *Id.* at 260.

[33] *Id.* at 262.

[34] *Id.*

[35] *Id.* at 263.

[36] *Id.* at 262.

[37] *Id.*