# USING KEYWORD SEARCH TERMS IN E-DISCOVERY AND HOW THEY RELATE TO ISSUES OF RESPONSIVENESS, PRIVILEGE, EVIDENCE STANDARDS AND RUBE GOLDBERG

By: Gregory L. Fordham[*]

Cite as: Gregory L. Fordham, *Using Keyword Search Terms in E-Discovery and How They Relate to Issues of Responsiveness, Privilege, Evidence Standards, and Rube Goldberg*, 15 RICH. J.L. & TECH. 8, http://law.richmond.edu/jolt/v15i3/article8.pdf.

## I. INTRODUCTION

[1] The emergence of digital evidence and the widespread implementation of e-discovery has brought both benefit and repercussion. In many respects, digital evidence has proven to be a better truth detector than its paper counterpart. At the same time, the volumes in which digital evidence exists make time-tested discovery techniques impractical. In fact, so significant are the technological differences between paper and digital evidence that even the handling procedures require considerable overhaul.

[2] One area undergoing an overhaul is the field of information retrieval. In a litigation environment, information retrieval has many applications, such as finding responsive documents, relevant documents, privileged documents, and documents related to particular events, issues or people that are of significance to the case. While initially this capability may have been limited to various indexing techniques, digital searching

[*] Gregory Fordham is a founder of K&F Consulting in Atlanta, Georgia. He regularly advises clients on how to structure their e-discovery plans in order to minimize cost and maximize return. He has been an expert witness in state and federal cases involving e-discovery and computer forensics. He can be reached at greg@knfcon.com.

capabilities have dramatically improved so that it is now possible to electronically search very large repositories of data. Furthermore, these searches are not just of indexed attributes of the documents. Rather, the searches can be performed against the entire contents of the document, including, in the case of native format documents, embedded data that is not otherwise available to the normal user.[1]

[3] Although word search capability has existed for many years, the technology has greatly improved and advanced features and capabilities like Boolean connectors, proximity locators, fuzzy logic, and stemming are now available in many keyword search tools.[2]

[4] Interestingly, as the prevalence of digital evidence and the practice of e-discovery have permeated the legal profession, the use and need for keyword search capabilities has increased. With many modern litigations producing and relying on volumes of digital evidence, it is simply not practical to take a "boots on the ground" approach to document review and analysis. Certainly, the size and extent of the data make it commercially impractical to use anything other than computerized techniques for keyword searches. Moreover, many other fields of human activity have demonstrated that the weak link in the chain is often the human element. For example, statistical sampling techniques are often used not only for economic purposes but for increased accuracy as well.[3]

[5] In fact, the weakness in the once preferred human element of document review and analysis is even reflected in Practice Point 1 of the Sedona Conference's *Best Practice Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*.[4] Practice Point 1 states that, "In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding

[1] *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 210 (2007) [hereinafter *Sedona Conference Best Practices Commentary*].
[2] *See infra* Part III.B. for description.
[3] B.J. MANDEL, STATISTICS FOR MANAGEMENT: A SIMPLIFIED INTRODUCTION TO STATISTICS 174 (Angela Murray & George H. Trafton eds., 1977) ("We can cite many cases where sampling has served the purpose better than a complete enumeration of the population . . . better in terms of accuracy and better in terms of cost and time.").
[4] *See Sedona Conference Best Practices Commentary*, *supra* note 1, at 208.

responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary."[5]

[6] For the above reasons, computerized keyword search techniques have become widespread. Furthermore, their use will likely continue to become more prevalent. Practitioners, who have not used these searches in the past, will be forced to implement these technologies and techniques as digital evidence and e-discovery force them to forego the traditional "boots on the ground" approach.

[7] The transition will not be painless. Practitioners are now having digital evidence and e-discovery challenges thrown at them en masse.[6] The result is an increasing number of decisions involving the integration of digital evidence and e-discovery into the profession.[7] The use of keyword search techniques is no exception.

[8] In the past year, there have been numerous decisions involving the use of keyword search techniques in responding to production requests and performing privilege reviews.[8] In resolving those disputes, judges have realized that the complexity of the subject warrants more than a lawyer's

[5] *Id.*

[6] A search of Federal Court decisions in Westlaw for the words "electronically stored information" over the years 2005 through 2008 produced the counts of 7, 27, 143 and 228 for each respective year. Similarly, a search of Federal Court decisions in Westlaw for the words "electronic discovery" over the same four years produced counts of 13, 25, 72 and 99 for each respective year. *See* Westlaw Legal Research, http://westlaw.com.

[7] *Sedona Conference Best Practices Commentary*, *supra* note 1, at 197. *See generally* Covad Commc'ns Co. v. Revonet, Inc., 254 F.R.D. 147 (D.D.C. 2008) (holding that a customer lead services provider should produce its e-mails in their native format—electronic—rather than in hard copy); *In re* Seroquel Prods. Liab. Litig., 244 F.R.D. 650 (M.D. Fla. 2007) (holding that discovery sanctions were proper for the manufacturer's failure to meet electronic discovery commitments); Mosaid Techs. Inc. v. Samsung Elects. Co., 348 F. Supp. 2d 332, 339 (D. N.J. 2004) (stating that parties have an affirmative obligation to preserve "potentially relevant digital information").

[8] *See Sedona Conference Best Practices Commentary*, *supra* note 1, at 197. *See generally* South Yuba River Citizens League v. Nat'l Marine Fisheries Serv., No. CIV. S -06-2845 LKK/JFM, 2008 WL 2523819 (E.D. Cal. June 20, 2008) (stating that the scope and methods of keyword search performed by defendants was adequate); Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251 (D. Md. 2008) (holding that the keyword search performed by defendants had not been reasonable).

representation on which to base a decision, since keyword search techniques are a part of information retrieval science that lies at the intersection of linguistics, statistics, and computer technology.[9]

[9] The following sections examine the significance of these three subjects and how they have affected issues involving responsiveness, privilege, and evidence standards. Finally, the lessons learned are used to formulate recommendations for practitioners when designing keyword search plans in their cases.

## II. LINGUISTICS, STATISTICS, & COMPUTER TECHNOLOGY

[10] Computerized keyword search techniques are not just a technology issue involving how to mechanically read digital data in order to find a needle in a haystack. Indeed, after the technological problem is solved, the resulting outcome can be categorized into four different groups:

> (1) exact matches, where the keyword search terms are in documents containing the matters of interest;
>
> (2) false positives, where the search terms are in documents not related to the matters of interest;
>
> (3) false negatives, where the documents of interest do not contain any of the search terms; and
>
> (4) complete rejections, where the documents do not contain the search terms and do not contain any of the matters of interest.

Ideally, the searcher would like to have the documents separated into only two of these groups—exact matches and complete rejections. In other words, the researcher would like to prevent documents from falling in the other two categories of false positives and false negatives. The searcher would like to avoid false positives because reviewing such documents wastes resources. Avoiding false negatives is equally important because their exclusion may result in an incorrect outcome, which could be disastrous.

[9] *See Sedona Conference Best Practices Commentary*, *supra* note 1, at 197.

[11] Because it is so important to avoid the potential for false negatives, the search terms are often overly inclusive. By increasing the inclusiveness of search terms, however, the chance of false positives as well as the costs of the search is increased. Thus, it is the researcher's goal (through linguistics, statistics, and technology) to reduce the chance of false positives and false negatives. The following sections examine the linguistics, statistics, and technology and how they can be used to reduce false negatives and false positives.

### A. LINGUISTICS

[12] The science of linguistics involves how people use language to yield meaning.[10] In electronic discovery, this is important because different terms can be used to describe the same issue. Although this is to be expected when going from case to case, it can also occur within the same case. Furthermore, seemingly common terms can be given unique meanings by the parties of interest in a particular case. While linguistics would seem to have as its greatest goal the reduction of false negatives, linguistics can also be used to reduce false positives. After all, if the correct selection of keyword search terms is accomplished, then both situations can be eliminated.

[13] A frequently cited example involving linguistics in the formation of keyword search terms is the Blair and Maron Study.[11] The Blair and Maron Study is best known for its examination of a case from 1985 where a Bay Area Rapid Transit (BART) System vehicle failed to stop at the end of the line.[12] In that case, attorneys working with experienced paralegals were able only to find about 20% of the relevant documents despite their belief that they had found more than 75% of the relevant documents.[13]

[10] Merriam-Webster Online Dictionary, http://www.merriam-webster.com/dictionary/linguistics ("linguistics: the study of human speech including the units, nature, structure, and modification of language") (last visited Mar. 4, 2009).
[11] David C. Blair & M.E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMUNC'NS. OF THE ACM 289, 295 (1985).
[12] *Sedona Conference Best Practices Commentary, supra* note 1, at 208.
[13] *Id.*

[14] The Blair and Maron Study is referenced by many commentators, including the Sedona Conference.[14] As explained in a Sedona Conference publication on information retrieval, Blair and Maron found that the words used by the two parties to refer to the relevant issues were entirely different.[15] For example, BART and related parties referred to "the unfortunate situation," while the victim and related parties referred to it as a "disaster."[16] In other places, terms like the "event," "incident," "situation," "problem," or "difficulty" were used.[17] In the end, the linguistic differences were far greater than realized by the legal team, underestimating the disparity adversely affected their work.

[15] Clearly, linguistics poses challenges for those using keyword search terms to find documents of interest. Searchers can overcome these challenges through several techniques. First, they could use thesauri in order to develop synonym lists. In fact, such a process could even be incorporated by various electronic search tools. In addition, searchers could perform interviews and other investigative techniques to learn the terminology commonly used by the creators of documents within a population. Finally, searchers can perform test runs of search terms against document populations in order to test their effectiveness and to potentially identify search terms that would be better predictors of documents of interest.

### B. STATISTICS

[16] The science of statistics is another way to battle the problem of false positives. Under this methodology, probability theory is used to make decisions about the relevancy of documents.[18] The application of the statistical analyses may be based on sets of "model" documents.[19] Then based on the model documents, comparisons are made against documents

[14] *See id.* at 206.
[15] *Id.*
[16] Blair & Maron, *supra* note 11, at 295; *see Sedona Conference Best Practices Commentary*, *supra* note 1, at 206.
[17] Blair & Maron, *supra* note 11, at 295; *see Sedona Conference Best Practices Commentary*, *supra* note 1, at 206.
[18] *See Sedona Conference Best Practices Commentary*, *supra* note 1, at 218.
[19] *Id.*

in the population.[20] The actual comparisons that are made could be things like the number of times that the keyword appears in the document, the location of the keyword in the document, or the association of keywords with other words in the document.[21]

[17] Once these comparisons are made and the metrics tabulated, probability theory is applied in order to "score" the likelihood that the document involves the issues of interest and to categorize the document accordingly.[22] Interestingly, this rather simple approach seems to work well in complex, real world situations.[23]

### C. COMPUTER TECHNOLOGY

[18] Computer technology is the third leg of the keyword search triangle. It involves more than just the computerized implementation of linguistics, statistics, and computer technology. Indeed, there are a considerable number of variables that enter the equation and must be considered when planning a keyword search.[24] In general, the three main variables this article will focus on are: approach, features, and limitations.

## III. APPROACH, FEATURES, AND LIMITATIONS

### A. APPROACH

[19] The approach considers how the search engine performs its functions. Essentially, there are two basic types of approaches. One is an

[20] *Id.* at 219.
[21] *Id.* at 218.
[22] *Id.* at 219.
[23] *See, e.g.*, Ehud Guttel & Alon Harel, *Uncertainty Revisited: Legal Prediction and Legal Postdiction*, 107 MICH. L. REV. 467, 467, 477-79, 498-99 (2008) (discussing the use of probability theory in making laws that will be more effective than others for encouraging certain behaviors); Lawrence Joseph & Caroline Reinhold, *Fundamentals of Clinical Research for Radiologists: Introduction to Probability Theory and Sampling Distributions*, 180 AM. J. ROENTGENOLOGY 917, 917 (2003) (discussing the use of probability theory to determine the effectiveness of certain radiological diagnostic tools over others).
[24] Kenneth H. Ryesky, *From Pens to Pixels: Text-Media Issues in Promulgating, Archiving, and Using Judicial Opinions*, 4 J. APP. PRAC. & PROCESS 353, 383 (2002).

indexed search, and the other is an un-indexed or single pass method.[25] The indexed search method first produces an index, which has the advantage of providing the searcher with iterative capability. [26] The timeliness of search results can be important when trying to develop the best predictors for locating the desired documents because the selection can require an interactive approach of firing off terms and reviewing the results. Therefore, the indexed method is much better suited to sampling and testing the adequacy of search terms.

[20] Unlike the indexed method, which reads the documents in advance, creates an index, and then performs a search against the contents of the index, the un-indexed or single pass method essentially starts at the beginning and proceeds through the population of documents to be searched.[27] Both techniques have advantages and disadvantages. Since the indexed method must first create the index, it has a longer setup time before the first search results are ever realized than that of the single pass method. Accordingly, the non-indexed method can provide its initial results quicker than the indexed method. Each iteration under the non-index method, however, must traverse the entire document population while the index method does not. Indeed, the index method simply queries the index. As a result of its faster iterative speed, the indexed search method is a better way to refine the search terms to not only limit false positives, but also to test the adequacy of synonyms in finding documents of interest.

## B. FEATURES

[21] In addition to the indexed or un-indexed choice there are several other features for which the users of search engines should look.[28] All of these features provide the ability to overcome the linguistic difficulties

[25] *See generally* Curt Franklin, *How Internet Search Engines Work*, HOW STUFF WORKS, http://computer.howstuffworks.com/search-engine.htm/printable (last visited Mar. 4, 2009) (explaining how search engines develop indexes for the purposes of conducting searches).
[26] *Id.*
[27] *Id.*
[28] *See Sedona Conference Best Practices Commentary*, *supra* note 1, at 192.

mentioned above. These are Boolean connectors, proximity locators, fuzzy logic, and stemming, to name a few.[29]

[22] First, Boolean connectors permit more complex searches than a single term word search. The typical connectors are terms such as AND, OR, and NOT.[30] Their use can reduce false positives by increasing the filter criteria on the document selection process.[31] Second, a subset of a Boolean connector searches are proximity locators. Using proximity locators refines the search criteria in order to more accurately pinpoint the documents of interest and avoid false positives.[32] Proximity locators provide capability similar to Boolean connectors in that the terms are within or not within certain distances of each other.[33]

[23] Third, fuzzy logic recognizes that the search terms could be misspelled or spelled differently within the document.[34] Fuzzy logic places the equivalent of wild cards within the spelling of the search term in order to permit alternative spellings.[35] The use of fuzzy logic is often implemented by allowing users to specify the placement of the wild card characters or by specifying the degree of fuzziness. Lastly, stemming, also referred to as "wildcard operators," recognizes that the search term can be part of the basic word as in the case of plural terms or words with alternate endings like "ed", "ing", "ly," or "ion."[36] The use of stemming is not limited to suffixes and can include prefixes as well.[37]

### C. LIMITATIONS

[24] Another element with which users of keyword search technology should be familiar, is the wide array of keyword search limitations. These

[29] *Id.* at 192, 197.
[30] *Sedona Conference Best Practices Commentary*, *supra* note 1, at 197.
[31] *See id.*
[32] *See id.* at 217.
[33] *Id.*
[34] *See id.* at 219.
[35] *Id.* at 202, 219.
[36] *See id.* at 218.
[37] *See id.*

limitations come in two general forms: limitations of the specific search engine and limitations of keyword searches in general.[38]

[25] The limitations of the search engine are generally restricted to the data being searched. For example, a text-based search engine cannot interpret non-textual data.[39] The classic example of such a condition is a graphic image. Even though the graphic image may display textual data that data is not stored in a textual format that can be searched. Rather, the data must first be converted into a textual format, before it can be searched.[40] One example of this kind of situation is an engineering drawing, which contains installation instructions, part numbers, and part descriptions. Other examples include fax pages stored on a fax server, brochures and marketing literature, and imaged documents stored as part of a document retrieval system.[41]

[26] The conversion of the graphic image to text-based data is typically accomplished through other software tools that perform Optical Character Recognition (OCR).[42] Although the process is not perfect, it is one means for converting imaged documents into text-based documents that can then be searched. Since documents that have been processed through this type of conversion are more likely to contain spelling errors, they are good candidates for fuzzy logic searches.

[27] In the world of electronic discovery, not every document fits in either the text-based or image-based category. There are still places in between that the search engine must be able to handle. A classic example of the

[38] Why Catalogue the Internet? The Limitations of Search Engines, http://www.vuw.ac.nz/staff/alastair_smith/catint/srcheng.htm (last visited Mar. 31, 2009); *see* Autonomy Corp.: Keyword and Boolean Searches, http://www.autonomy.com/content/Technology/autonomys-technology-limitations-of-other-approaches-keyword-boolean/index.en.html (last visited Mar. 31, 2009).
[39] *Id.*
[40] *Id.*
[41] *Id.*
[42] Carol Schlein, *Time for a Scanner: Prime Essential for Your "Paperless" Office*, 21 LAW. PC 1 (2004); *see* Sami Lais, *Quick Study: Optical Character Recognition*, COMPUTERWORLD, July 29, 2002, http//www.computerworld.com; Microsoft Office Online, About Optical Character Recognition (OCR), http://office.microsoft.com/en-us/help/HP030812551033.aspx?pid=CH010000951033 (last visited Mar. 31, 2009).

middle category is a compressed file type such as a zip file.[43] In the case of a zip file, the entire contents could be text-based documents but as a result of the compression algorithm, the characters no longer present recognizable words.[44] In the case of compressed files, the user has two choices: either manually uncompress the zip file archives so that the documents can be searched by the search engine, or use a search engine that can handle compressed document formats.[45] In fact, compressed archives are not the only place where this kind of situation is encountered. Compound documents of all types, including e-mail, provide similar obstacles to overcome.

[28] The searcher must also be sensitive to whether the document has been encrypted, such as with a password protected file.[46] In this case, the file may be an otherwise recognizable file type that the search engine can handle, but if it has been password protected or otherwise encrypted its contents could be scrambled.[47] Thus, the searcher needs a method for identifying encrypted or password protected files in a data population so that any encryption can be removed. Typically such detection is based on an entropy test which is not likely part of the search engine's capabilities. Nonetheless, in these days of heightened information security awareness, the detection of encrypted files prior to processing by a search engine is essential.[48]

[29] Finally, the searcher must consider the status of the information being sought. For instance, is the target an active file or a deleted file? If it is a deleted file, then the search engine will need to be capable of

[43] Microsoft Office Online, Zip or Unzip a File, http://office.microsoft.com/en-us/help/HA011276901033.aspx (last visited Mar. 31, 2009).
[44] *See id.*
[45] *See id.*
[46] *See, e.g.*, SETH SCHOEN, ELECTRONIC FRONTIER FOUNDATION, TRUSTED COMPUTING: PROMISE & RISK 4 (Oct. 2003), http://www.eff.org/files/20031001_tc.pdf.
[47] INFORMATION INFRASTRUCTURE TASK FORCE, INTELLECTUAL PROPERTY AND THE NATIONAL INFORMATION INFRASTRUCTURE: THE REPORT OF THE WORKING GROUP ON INTELLECTUAL PROPERTY RIGHTS 185-86 (1995), *available at* http://www.uspto.gov/web/offices/com/doc/ipnii/front.pdf (last visited Mar. 31, 2009).
[48] *See, e.g.*, National Cyber Security Alliance, Stay Safe Online, http://www.staysafeonline.org/ (last visited Mar. 31, 2009) ("[T]he mission of the NCSA is to create a culture of cyber security and safety awareness by providing the knowledge and tools necessary to prevent cyber crime and attacks.").

searching media at a lower level than the filing system.[49] If the search engine does not have this capability, then the searcher will need to expend some minimal effort to recover deleted files so that they can be searched.[50]

[30] The search can be even trickier if distinctions must be made between a media's "free space" and a file's "slack space."[51] The "free space" label applies to those areas on the media that are available for new data as it is saved to the media.[52] "Slack space" refers to a section of memory that is already allocated to storing an active file, but which is not completely filled by the file it is storing.[53] If the search target includes free space or even slack space in files, then the search engine must be capable of searching the media below the operating system level.

[31] The lesson for the search engine user is to know the contents of the document population in terms of the data formats that it holds, including encrypted files, and then to ensure that the search engine can handle those file types. For those that it cannot handle, the searcher will have to devise an alternate approach, such as first converting the documents to a format that the search engine can examine.

[32] Not every kind of computer search is suitable for keyword search terms. For example, finding spoliation can involve the identification of artifacts pointing to files that no longer exist on the media but had existed at a point in time when there was a duty to preserve.[54] After all, the mere fact that documents of interest are not returned by a keyword search is not absolute proof that they have been spoliated, particularly when the opposing party's claim is that they were never there. A keyword search plan that yields no results is exactly what one would expect if the target was a hard drive whose files of interest had been deleted and the free

[49] *See* Kenneth J. Withers, *Electronically Stored Information: The December 2006 Amendments to the Federal Rules of Civil Procedure,* 4 NW. J. TECH. & INTELL. PROP. 171, 174 (2006).
[50] *See id.*
[51] United States v. Criminal Triumph Capital Group, Inc., 211 F.R.D. 31, 46 (D. Conn. 2002).
[52] *Id.* at 46 n.6.
[53] *Id.* at 46 n.7.
[54] *See, e.g., id.* at 47.

space of the drive overwritten.[55] Furthermore, the fact that this has occurred will not be revealed by the keyword search effort**,** but rather, only through visual examination of the free space areas of the drive as well as a review of other system metadata artifacts.[56]

[33] Similarly, the storage location for documents is not limited to a litigant's computer's hard drive. Indeed, files could also have resided on other storage devices that were attached to the litigant's computer or on devices to which the litigant had access over a network.[57] Thus, whether or not the hard drive surrendered is the only place where the keyword search effort should be performed is a question that cannot be answered by keyword search terms alone.

[34] Therefore, there are clear limitations to keyword search techniques. Those limitations are not restricted to linguistics, statistics, or economics. Rather, they include search engine technology as well as the kinds of questions to be answered. Keyword search techniques are useful in answering the question, "Where is the needle in the haystack?" But if the question involves which haystack and how it got there, keyword search tools are not efficient in answering that question—particularly if the farmer is playing "hide the haystack." It is important for the searcher to understand all of these limitations and how they apply to keyword search tools when designing and implementing a keyword search plan.

## IV. RESPONSIVENESS

[35] Some of the most common uses of keyword search terms are in locating responsive documents for a Rule 34 of the Federal Rules of Civil Procedure production request, a protocol developed under Rule 26, or other meetings and negotiations. Disputes can arise when the search results are not as expected. In the event where the producing party

[55] *See, e.g., id.* at 47 & n.14.
[56] *See generally id.* at 47 (discussing how the litigant acted reasonably by "opening, screening and manually reviewing data" in the hard drive to ensure a thorough search effort).
[57] *See generally* Benjamin D. Silbert, Comment, *The 2006 Amendments to the Rules of Civil Procedure: Accessible and Inaccessible Electronic Information Storage Devices, Why Parties Should Store Electronic Information in Accessible Formats*, 13 RICH. J.L. & TECH. 14, ¶¶ 39-43 (2007) (discussing the ever-changing technologies for data back-up).

controls the search and the selection of keywords, the dispute is often about the adequacy of the terms. Two decisions published during 2008 exemplify this situation: *United States v. O'Keefe*[58] and *Equity Analytics, LLC v. Lundin*.[59] Interestingly enough, both cases were decided by Judge Facciola of the United States District Court of the District of Columbia. The following sections examine the decisions in these two cases.

### A. UNITED STATES V. O'KEEFE[60]

[36] The case of *United States v. O'Keefe*[61] was a criminal matter involving the bribing of State Department officials to expedite the issuance of visas.[62] As part of their defense, the defendants sought discovery showing that there was no established policy or procedure regarding expediting visa applications, or that it was the routine practice to violate or disregard the policy.[63]

[37] The defendants' claim was that nothing unorthodox happened in this case, because visa applications were routinely expedited in consulates in Toronto, in other parts of Canada, and Mexico.[64] In fact, they claimed, expediting was so routine that the decision whether to expedite was delegated even to non-official, clerical personnel.[65] Thus, obtaining documents through discovery to show these facts would undercut any evidence offered by the government concerning the formality of the process.[66]

[38] In answering the defendants' request, the government used the following search terms: "early or expedite* or appointment or early & interview or expedite* & interview."[67] Clearly, these terms demonstrate some sophistication in the approach as well as the search engine, since

[58] 537 F. Supp. 2d 14 (D.D.C. 2008).
[59] 248 F.R.D. 331 (D.D.C. 2008).
[60] 537 F. Supp. 2d 14 (D.D.C. 2008).
[61] *Id.*
[62] *Id.* at 15-16.
[63] *Id.* at 16.
[64] *Id.*
[65] *Id.*
[66] *Id.*
[67] *Id.* at 18.

there are Boolean connectors and stemming attributes to the terms. In addition, the use of "early" or "expedite" demonstrates that at least some efforts were made to overcome linguistic issues. The results of the search produced numerous documents including those that were unrelated to the issue, such as early departures of employees for various personal reasons (e.g., dental appointments).[68] Nonetheless, the result did include some responsive e-mails, some standard operating procedures, and the Non-Immigrant Visa (NIV) Schedule Calendar.[69]

[39] What was not clear from these results and the search in general was the information relating to the type of search engine, what data formats it could handle, and whether the recovered e-mails were still within an e-mail server post office or personal mailbox, or were simply copies that individuals had extracted and saved in plain text. Also, there was no indication whether any efforts were made to prepare non-text-based files for searching. The question arose: were relevant forms and applications subsequently scanned and stored in some kind of electronic document retention system, and had those images been converted to text prior to performance of the search? Similarly, there was no discussion about the steps taken by the government to determine how such expediting tasks were commonly described by employees, and where or in which data format the information about those activities would be stored. For example, there was no indication if this information would be captured in a database application, if the application could be searched by the text search engine, or whether it was ever searched.

[40] After the search, the defendants were disappointed in the results and protested both the search terms used as well as the process.[70] More specifically, the defendants faulted the government for not interviewing the employees to ascertain how often they had used electronic means to create any electronic documents regarding expedited interviews.[71] The defendants also questioned whether the search engine was capable of

[68] *Id.*
[69] *Id.*
[70] *Id.* at 21-22.
[71] *Id.* at 22.

searching for e-mail within a .pst file, which is a Microsoft Outlook personal mailbox container.[72]

[41] In terms of keyword search techniques, *O'Keefe* illustrates numerous failures in the process of search term design and planning as well as technology planning. There was inadequate analysis of the visa-granting environment and the phraseology commonly used by government employees in order to determine the appropriate keywords. There was also inadequate testing of the terms to determine whether they were actually good predictors of responsive documents in the population. Finally, there was no apparent examination conducted to determine whether the particular search technology was adequate for the data types of potentially responsive documents or whether responsive documents even existed in the locations searched.

### B. EQUITY ANALYTICS, L.L.C. V. LUNDIN[73]

[42] *Equity Analytics, LLC v. Lundin*[74] was a trade secret case that illustrated the limitations of keyword search terms in locating responsive documents.[75] Specifically, Equity had sought to have its forensic expert examine Lundin's computer to ascertain: "(1) whether Lundin accessed Equity's confidential customer data and/or trade secrets; (2) whether the data ha[d] been forwarded to Lundin's new employer an Equity competitor; and (3) whether the data was purged or overwritten."[76]

[43] Since Lundin's computer and its hard drives contained very personal information including attorney-client communications, business records, medical records, tax and banking records, and data (including photographic images) created for Lundin's professional photography business, the plaintiff's counsel proposed that the forensic computer examiner use search terms to restrict the search to data relevant to the case.[77]

---

[72] *Id.*
[73] 248 F.R.D. 331 (D.D.C. 2008).
[74] *Id.*
[75] *Id.* at 332.
[76] *Id.*
[77] *Id.*

[44] Equity, however, recognized that keyword search techniques would be inadequate because Lundin had loaded a new operating system onto his computer that could have compromised the integrity of the files that were previously on the computer.[78] A new operation system installation or computer usage that led to the deletion or partial overwriting of files could potentially result in the remainder of only fragments of information rather than complete files.

[45] In addition, Equity also questioned Lundin's restriction of the keyword search terms to certain document types like Microsoft Word, Excel, PowerPoint, and Adobe Acrobat.[79] Equity argued that confidential files could have been downloaded and saved in a phony format or with a different extension in order to "disguise their identity."[80] In that case, even though their contents could betray them, these types of documents would never be selected to be searched in the first place.

[46] *Equity* illustrates the limitation of keyword search terms in litigation. Unlike *O'Keefe*, however, the limitations in *Equity* are not limitations of process, but rather, limitations of technology and the kind of tasks for which it is suited. For example, even when the documents are present on the media and found through the use of keyword search terms, their existence on Lundin's computer still does not answer questions such as whether they have been forwarded to someone else or shared with Lundin's new employer. On the other hand, even if documents are not on the computer and not found by keyword search terms, questions such as whether Lundin accessed Equity's confidential data and whether that data had been purged or overwritten are not answered. If the data had been saved to another storage device, for example a thumbdrive or external hard drive, instead of Lundin's computer, the keyword search terms would not confirm the data's existence or that Lundin had accessed it. Furthermore, unless the other storage device had been included as part of the keyword search, it would have been impossible to reveal the documents' existence in Lundin's possession through that search.

[78] *Id.*
[79] *Id.* at 332-33.
[80] *Id.* at 333.

[47] Similarly, if the documents had been purged or overwritten this fact would not be confirmed simply by the failure of a keyword search term to return any results. Only if the documents had been deleted from the media and some kind of fragment remained, could the keyword search effort prove fruitful.[81] If the file had been wiped or entirely overwritten, then the keyword search effort would likely be useless.

[48] *Equity* also illustrates the limitation of keyword search terms in locating responsive documents. In *Equity*, this limitation involved questions about how a document was used.[82] Certainly it is possible that finding a particular document on more than one computer can indicate that it has been shared and seen by more than one person. A document's mere existence on certain media, however, is less than conclusive proof of its usage. Evidence of usage can only be found by examining other artifacts of system metadata. Of course, even knowing what terms to use in finding such artifacts with keyword search terms would not be possible until after the document is found and its file system name determined.

[49] All of this presupposes that the document still exists on media or was not hidden elsewhere. If the document was hidden on other media or removed from the media under examination, then keyword search terms will never find the document. The only hope for unearthing evidence of a deleted file, is by uncovering system metadata artifacts referencing the file by its file system name. Finding such artifacts would at least confirm its earlier existence.[83] If the document was hidden on another media, keyword search terms are unlikely to reveal on what media it might reside, even though that information is likely stored in the hardware keys of the Windows registry. Thus, physical examination of the Windows registry for this kind of case is essential, as are other system metadata artifacts.

[50] The decision in *Equity* has similarities to the decision in another case, *Calyon v. Mizuho Securities USA Inc.*[84] In *Calyon*, the plaintiff was

[81] *See* Withers, *supra* note 49.
[82] *Equity Analytics*, 248 F.R.D. at 332-33.
[83] System Metadata is information automatically generated when a file is created and may contain information such as authorship and time and date of creation. W. Lawrence Wescott II, *The Increasing Importance of Metadata in Electronic Discovery*, 14 RICH. J.L. & TECH. 10, ¶ 3 (2008).
[84] No. 07CIV02241RODF, 2007 WL 1468889 (S.D.N.Y May 18, 2007).

denied access to forensic images of the defendants' hard drives because the court found no compelling justification for why the defendants' experts did not find a thorough and responsive search result.[85] Specifically, the court noted the following: there was no argument that the defendants had failed to provide responsive documents, that there were discrepancies or inconsistencies, that data had been lost, or that there was any information on the image of the hard drive that the defendants would have been unwilling or unable to produce.[86]

[51] Like *Equity*, the *Calyon* case involved trade secret data. As explained above, while keyword search terms may, in the right circumstances, reveal the existence of certain data, these searches will not answer other questions such as how the data was obtained and what happened to it. Both decisions resulted in denying the plaintiff's access to the hard drives in order to answer those questions. Thus, both *Equity* and *Calyon* underscore the limitations of keyword search terms, particularly in instances of data hiding, yet also illustrate an emerging trend in defense tactics.

[52] In the early part of the decade when the answer to e-discovery was cost-shifting, defense tactics adapted to case precedent by using cost-shifting rules to frustrate discovery and conceal evidence.[87] The use of keyword search terms in poorly matched situations may be a similar tactic.

[53] While judges are sensitive to the privacy rights of defendants, the answer may not be limiting the analysis to keyword search terms. Rather, it may be a combination of keyword search terms and limited analysis. For example, the kinds of analysis that computer forensic examiners would want to perform (other than recovery of deleted files), when

[85] *Id.* at *5-6.
[86] *Id.* at *5.
[87] *See, e.g.*, Oppenheimer Fund, Inc. v. Sanders, 437 U.S. 340, 358-59 (1978); Hagemeyer N. Am., Inc. v. Gateway Data Scis. Corp., 222 F.R.D. 594, 601-02 (E.D. Wis. 2004); OPENTV v. Liberate Techs., 219 F.R.D. 474, 476-77 (N.D. Cal. 2003); Rowe Entm't, Inc. v. The William Morris Agency, Inc., 205 F.R.D. 421, 423, 428 (S.D.N.Y. 2002); McPeek v. Ashcroft, 202 F.R.D. 31, 34 (D.D.C. 2001); *see also* Rebecca Rockwood, Comment, *Shifting Burdens and Concealing Electronic Evidence: Discovery in the Digital Era*, 12 RICH. J.L. & TECH. 16, ¶¶ 15-21 (2006) (detailing the early development of cost shifting rules).

answering questions about document access and data hiding are largely available through the examination of system metadata artifacts such as the registry, system logs, software application logs, file pointers, and related file system information. Thus, perhaps any examination should include these artifacts as well as keyword searches.

## V. PRIVILEGE

[54] Privilege issues are another area where keyword search terms have been used and disputes have arisen. Generally, the disputed issues involve situations where privileged documents have been produced inadvertently. Two such cases, *Victor Stanley, Inc. v. Creative Pipe, Inc.*[88] and *Rhoads v. Building Materials Corp. of America*,[89] were decided in 2008 and are discussed in the sections below.

### A. VICTOR STANLEY, INC. V. CREATIVE PIPE, INC.[90]

[55] *Victor Stanley, Inc. v. Creative Pipe, Inc.*[91] was a case where 165 privileged documents were inadvertently produced by Creative Pipe, Inc. (Creative Pipe).[92] Once the disclosure was discovered by Victor Stanley, Inc. (Victor Stanley) it argued that the privilege had been waived, while Creative Pipe argued that the disclosures were inadvertent and that the privilege had not been waived.[93]

[56] *Victor Stanley* is instructive in that it demonstrates the importance of the keyword search process in preserving privilege in the event of an inadvertent disclosure. After all, in order to preserve privilege in the event of an inadvertent disclosure, one must demonstrate that the efforts taken were reasonable.[94] In addition, the case exemplifies the importance of Practice Point 7 in the Sedona Conference's *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* which states: "Parties should expect that their choice of search

[88] 250 F.R.D. 251 (D. Md. 2008).
[89] 254 F.R.D. 216 (E.D. Pa. 2008).
[90] 250 F.R.D. 251 (D. Md. 2008).
[91] *Id.*
[92] *Victor Stanley, Inc.*, 250 F.R.D. at 253.
[93] *Id.*
[94] FED. R. EVID. 502(b)(2).

methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials)."[95]

[57] The decision does not explain the terms used by Creative Pipe in its search, because none were provided as evidence.[96] Also, it does not explain whether testing or sampling of the search terms was performed to assess their adequacy as good predictors, because none were provided by Creative Pipe.[97] The case does reveal that volumes of data were split into approximately 4.9 gigabytes of searchable text and 33.7 gigabytes of non-searchable text.[98] Relevant facts reveal that keyword search technology was used on the searchable text data, however, the non-searchable text files received only a file name evaluation.[99] In other words, the non-searchable text files were not converted into a searchable format.[100] Instead, their names were used to determine whether the file would likely contain privileged information and warrant subsequent review.[101]

[58] According to Victor Stanley, however, the privileged data that was produced was completely contained within the text-based files which could have been electronically searched.[102] Therefore, Creative Pipe's failures cannot be attributed to its failure to convert non-text files into something searchable. Consequently, one can surmise that whatever terms and processes were used by Creative Pipe, they were ill-conceived and not properly validated.

[95] *Sedona Conference Best Practices Commentary*, *supra* note 1, at 212.
[96] *See Victor Stanley, Inc.*, 250 F.R.D. at 262.
[97] *See id.*
[98] *Id.* at 256.
[99] *Id.*
[100] *Id.*
[101] *See id.*
[102] *Id.* at 257.

### B. RHOADS INDUSTRIES, INC. V. BUILDING MATERIALS CORP. OF AMERICA[103]

[59] *Rhoads Industries, Inc. v. Building Materials Corp. of America*[104] was another case where privileged material was inadvertently produced.[105] As in *Victor Stanley*, the case is instructive in that it demonstrates the importance of process in preserving privilege in the event of an inadvertent disclosure. *Rhoads* was a breach of contract and negligent representation case.[106] After Rhoads Industries, Inc. (Rhoads) inadvertently produced about 800 privileged electronic documents, Building Materials Corporation of America (BMC) moved to deem a number of Rhoads' privilege claims waived.[107]

[60] *Rhoads* is one of the first cases to review this issue in light of the recently enacted Rule 502 of the Federal Rules of Evidence.[108] Under Rule 502, an inadvertent disclosure does not waive the privilege if the holder took reasonable steps to prevent disclosure and to rectify the error.[109] Therefore, the focus of the decision was to review the procedure used by Rhoads in performing its privilege review.[110] In reaching its decision, the court found numerous failures, yet ultimately found in favor of Rhoads with respect to the 800 inadvertently produced documents.[111]

[61] With regard to Rhoads' failures, the court found that it should have used additional search terms to weed out potentially privileged documents.[112] In particular, Rhoads should have used the names of all of its attorneys. Also, its search for privileged documents was limited to e-mail address lines, and did not include the e-mail body.[113] Therefore, any potentially privileged e-mails (as defined by its search terms) that were

---

[103] 254 F.R.D. 216 (E.D. Pa. 2008).
[104] *Id.*
[105] *Id.* at 218.
[106] *Id.*
[107] *Id.*
[108] *See id.* at 218 n.1 (stating that President Bush signed Rule 502 into law on September 19, 2008).
[109] FED. R. EVID. 502(b).
[110] *Rhoads Indus., Inc.*, 254 F.R.D. at 224.
[111] *Id.*
[112] *Id.*
[113] *Id.*

subsequently forwarded outside of Rhoads' infrastructure would not have been captured by its search.[114] In order to perform the search, Rhoads purchased new software.[115] Although using the new software, the system produced documents that its limited search should have caught.[116] Therefore, Rhoads failed to craft the right searches, and the searches it ran failed to identify documents meeting the search criteria.

## VI. EVIDENCE STANDARDS

[62] As evidenced by the preceding cases, the subjects of keyword search terms, techniques, and methods are likely subjects for dispute. Since keyword search efforts lie at the intersection of linguistics, statistics, and computer technology, it is only natural that the resolution of those disputes is sometimes beyond the knowledge of a layman and can require more than the representations of an attorney. In many instances, expert advice may be required.

[63] *U.S. v. O'Keefe*,[117] discussed above,[118] was one of the first cases to recognize that deciding such a dispute was beyond the knowledge of the layman and that expert advice could be needed. *O'Keefe* created a firestorm of debate about the impact of such requirements on the cost of litigation and whether it was indeed warranted.[119] The answer to such a problem came only a few months later in the case of *Victor Stanley, Inc.* v. *Creative Pipe, Inc.*, [120] also discussed above.[121] As a result, it seems well settled that expert testimony is a needed element of discovery motions involving the use of keyword search terms in electronic discovery matters, as well as potentially in other discovery rulings involving scientific and technical matters. The following sections review the debate on this issue as reflected in *O'Keefe* and *Victor Stanley*.

---

[114] *See id.*
[115] *Id.*
[116] *See id.* at 222.
[117] 537 F. Supp. 2d 14 (D.D.C. 2008).
[118] *See supra* Part IV.A.
[119] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 261 n.10 (D. Md. 2008).
[120] 250 F.R.D. 251 (D. Md. 2008);
[121] *See supra* Part V.A.

### A. U.S. v. O'KEEFE[122]

[64] As discussed previously, *U.S. v. O'Keefe*[123] was a criminal matter in which the defendants sought discovery showing that there was no established policy or procedure regarding expediting visa applications, or that it was the routine practice to violate or disregard the policy.[124] The defendants maintained that obtaining documents through discovery to show these facts would undercut any evidence offered by the government concerning the formality of the process as suggested by the indictment.[125]

[65] Although the government used keyword searches to answer the defendant's discovery requests, the defendants claimed that the government's search efforts were inadequate.[126] In deciding this case, the court recognized the technological nature of the issue.[127] While the court was sympathetic to the defendant's claims, it also acknowledged its own limitations in deciding such a dispute without the benefit of expert advice:

> Given this complexity, 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. This topic is clearly 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. Accordingly, if defendants are going to contend that the search terms used by the government were insufficient, they will have to specifically so contend in a motion to compel and their contention must be based on evidence that meets the requirements of Rule 702 of the Federal Rules of Evidence.[128]

---

[122] 537 F. Supp. 2d 14 (D.D.C. 2008).
[123] *Id.*
[124] *Id.* at 15-16.
[125] *See* O'Keefe, 537 F. Supp. 2d at 18-19.
[126] *Id.* at 22.
[127] *Id.* at 24 ("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.").
[128] *Id.*

Judge Facciola's decision in *O'Keefe* and his subsequent decision in *Equity* stirred up concerns by commentators that Rule 702 of the Federal Rules of Evidence would now govern discovery issues as well as admissibility issues.[129] After all, Rule 702 addresses the admissibility of evidence, which is not the purpose of discovery rules like Rule 26(b), which facilitate the search and collection of evidence.[130]

[66] Part of the concern by commentators in applying Rule 702 is the resulting increase on the cost of discovery.[131] Litigants would find it necessary to engage experts and expend financial resources much earlier in the process. Additionally, such a requirement could also have unintended consequences for litigants with more meager financial resources than deep pocketed corporations, for example.[132] It could make access to the legal system simply unattainable for all but the wealthy. Aside from the increased cost of experts, Rule 702 also involves the application of other legal principles such as the *Daubert* test and its requisite motions and hearings.[133]

[67] Although the issue in *O'Keefe* was related to the adequacy of keyword search terms and techniques,[134] the issue of expert testimony has broader application to all discovery matters involving scientific or technical subjects. For example, Principle 6 of the *Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production* suggests that the producing party is best positioned

[129] *See* Ronald J. Hedges, *Rule 702 and Discovery of Electronically Stored Information*, Digital Discovery and E-Evidence (BNA) at 121, 122 ( May 1, 2008).
[130] *Compare* FED. R. EVID. 702 (describing admissibility requirements), *with* FED. R. CIV. P. 26(b) (discussing the standards for searching and collecting evidence).
[131] *See, e.g.*, Derek L. Mogck, *Are We There Yet?: Refining the Test for Expert Testimony Through* Daubert, Kumho Tire *and Proposed Federal Rule of Evidence 702*, 33 CONN. L. REV. 303, 316 (2000).
[132] Hedges, *supra* note 129, at 122.
[133] *See* FED. R. EVID. 702 amendments advisory committee's note; *see also* Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 597 (1993) ("'General acceptance' is not a necessary precondition to the admissibility of scientific evidence under the Federal Rules of Evidence, but the Rules of Evidence-especially 702-do assign to the trial judge the task of ensuring that an expert's testimony both rests on a reliable foundation and is relevant to the task at hand.").
[134] *See* U.S. v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008).

to locate and produce their own electronically stored information (ESI).[135] Both *O'Keefe* and *Equity*, however, illustrate that the requisite knowledge of how best to locate and produce ESI does not automatically follow the parties with the best position. Furthermore, the two trade secret cases discussed above in Part IV.B., *Equity* and *Calyon*, demonstrate that the parties in the best position are not always the most incentivized to locate and produce ESI. As a result, Judge Facciola, who decided *O'Keefe* and *Equity*, found a similar need for expert opinions regarding the scientific and technical claims in those cases.[136]

### B. VICTOR STANLEY, INC. V. CREATIVE PIPE, INC.[137]

[68] As discussed above in Part V.A., *Victor Stanley, Inc. v. Creative Pipe, Inc*.[138] addressed a situation where privileged material was inadvertently produced.[139] The challenge was to determine whether privilege had been waived.[140] Making this determination required the

[135] THE SEDONA CONFERENCE, THE SEDONA PRINCIPLES: BEST PRACTICES RECOMMENDATIONS & PRINCIPLES FOR ADDRESSING ELECTRONIC DOCUMENT PRODUCTION ii, 38 (2d ed. 2007) ("Responding parties are best situated to evaluate the procedures, methodologies, and technologies appropriate for preserving and producing their own electronically stored information.").

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

> As I explained in that case, 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. Obviously, determining the significance of the loading of a new operating system upon file structure and retention and why the contemplated forensic search will yield information that will not be yielded by a search limited by file types or keywords are beyond any experience or knowledge I can claim.

*Id.* at 333.

[137] 250 F.R.D. 251 (D. Md. 2008).

[138] *Id.*

[139] *Id.* at 253.

[140] *See id.*

court to assess the processes the party used to perform its document review in response to the opposing party's discovery request.[141]

[69] In reaching his decision in *Victor Stanley*, Judge Grimm faced similar hurdles to Judge Facciola's in *O'Keefe* and *Equity*.[142] While commentators had questioned Judge Facciola's decision to intermingle Federal Rule of Evidence 702 with discovery rules, Judge Grimm found it appropriate:

> Judge Facciola made the entirely self-evident observation that challenges to the sufficiency of keyword search methodology unavoidably involve scientific, technical and scientific subjects, and *ipse dixit* pronouncements from lawyers unsupported by an affidavit or other showing that the search methodology was effective for its intended purpose are of little value to a trial judge who must decide a discovery motion aimed at either compelling a more comprehensive search or preventing one. . . . Indeed, it is risky for a trial judge to attempt to resolve issues involving technical areas without the aid of expert assistance.[143]

[70] In fact, to Judge Grimm, no artificial barrier exists between rules of evidence and discovery rules.[144] On the contrary, when the issues involve scientific or technical information, it is only reasonable that the information considered is from the kind of source contemplated by Rule 702.[145] Furthermore, he explained, this interplay between rules of evidence and other pretrial determinations is already common practice:

---

[141] *See id.* at 259 ("The intermediate test 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 to prevent inadvertent disclosure . . . .").

[142] *See Equity Analytics, LLC*, 248 F.R.D. at 333 ("[D]etermining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond a 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.").

[143] *Victor Stanley, Inc.*, 250 F.R.D. at 261 n.10.

[144] *See id.* ("That these common sense criteria are found in the rules of evidence does not render them off-limits for consideration during discovery.").

[145] *See id.* ("The rule is one of common sense . . . .").

> The goal of Federal Rule of Evidence 702 is to set standards to determine whether information is "helpful" to those who must make factual determinations involving disputed areas of science, technology or other specialized information. The rule is one of common sense, and reason-opinions regarding specialized, scientific or technical matters are not "helpful" unless someone with proper qualifications and adequate supporting facts provided such an opinion after following reliable methodology. That these common sense criteria are found in the rules of evidence does not render them off-limits for consideration during discovery. It is not unusual for pretrial factual determinations in civil cases to look to the Federal Rules of Evidence for assistance in resolving fact disputes. Indeed, in summary judgment practice, [Rule 56(e) of the Federal Rules of Civil Procedure] requires that the parties support their motions with "such facts as would be admissible in evidence."[146]

[71] *O'Keefe*, *Equity*, and *Victor Stanley* have blazed a trail, exemplifying that, at least in scientific and technical matters involving discovery, the claims of pretrial motions should be buttressed with expert assistance. While these precedent-setting cases involved issues about keyword search terms and techniques, the entire subject matter of electronic discovery is a fertile area for their application.[147] For example, the determination of accessible versus inaccessible data provides ample opportunities for expert assistance as technology and discovery tools make more and more data sources easily accessible.[148] Additionally, the best methods of preservation and whether they are overly burdensome or disruptive is another suitable subject area.

[72] For those concerned about the increased cost of discovery in light of *O'Keefe*, *Equity*, and *Victor Stanley*, Judge Grimm advised greater

[146] *Id.*
[147] *See* MANUAL FOR COMPLEX LITIGATION, FOURTH § 11.446 (2004).
[148] *See* Philip Beatty, *The Genesis of the Information Technologist-Attorney in the Era of Electronic Discovery*, 13 J. TECH. L. & POL'Y 261, 276-77 (2008).

cooperation in discovery planning.[149] According to Judge Grimm, the increased planning aspect is an underutilized feature of the discovery rules.[150] Although some 2008 surveys attributed the rising costs of discovery and litigation to the increased practice of electronic discovery, Judge Grimm expressed other ideas in his decision in *Mancia v. Mayflower Textile Services, Co.*[151]

[73] In *Mancia*, Judge Grimm reviewed the survey results, advisory committee notes, and studies and conclusions of numerous groups and legal scholars over the last fifty years.[152] The long history of discovery problems is evidence that the real problem is neither e-discovery nor the recent changes to the federal rules.[153] In general, Judge Grimm's analysis attributes the cause for escalating discovery costs to those not following the rules that have been in place for years.[154] Although Judge Grimm identified several failings, one of the most significant involved cooperation.[155]

[74] For example, in keeping with Judge Grimm's analysis in *Mancia*, Rule 26(g) of the Federal Rules of Civil Procedure imposes "an affirmative duty to engage in pretrial discovery in a responsible manner."[156] Some argue that the American adversarial system, which does not lend itself to the cooperation required by the rule, has undermined the requirement.[157] Yet, according to Judge Grimm, the

[149] Victor Stanley, Inc. v. Creative Pipe Inc., 250 F.R.D. 251, 261 n.10 (D. Md. 2008); *see also* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, ¶ 3 (2007) ("Litigators must collaborate far more than they have in the past, particularly concerning the discovery of information systems.").
[150] *See* Mancia v. Mayflower Textile Servs. Co., 253 F.R.D. 354, 361 n.3 (D. Md. 2008) ("Courts repeatedly have noted the need for attorneys to work cooperatively to conduct discovery, and sanctioned lawyers and parties for failing to do so.").
[151] *See id.* at 360 ("Discovery abuse is a principal cause of high litigation transaction costs.").
[152] *Id.* at 360-62.
[153] *Id.* at 360 ("Comparing these recent lamentations about the costs of civil litigation to those voiced eighteen years ago when the Civil Justice Reform Act of 1990 . . . was passed, and comprehensive changes to the discovery rules enacted, reflects that little has changed, despite concerted efforts to do so . . . .").
[154] *Id.*
[155] *Id.* at 361 n.3.
[156] FED. R. CIV. P. 26(g) advisory committee's notes; *Mancia*, 253 F.R.D. at 357.
[157] *Mancia*, 253 F.R.D. at 360-61.

cooperation required by the rule does not undermine the advocacy system.[158] Under Judge Grimm's analysis, advocacy is a form of public service.[159] Advocacy, however, ceases to be helpful when it hinders the process and "misleads, distorts and obfuscates," thus making the decision process more difficult.[160] According to Judge Grimm, there is ample justification for embracing the assistance of experts in discovery matters of a scientific and technical matter, and in mitigating any increased costs through increased cooperation by the parties.[161]

## VII. RUBE GOLDBERG

[75] Rube Goldberg was a 20th century cartoonist who is perhaps best known for his depictions of complex devices doing simple tasks in convoluted ways.[162] After reviewing the preceding cases and realizing that a properly designed keyword search methodology could include features like iterative testing, sampling, Boolean logic, proximity locators, stemming, fuzzy logic, thesauri, synonyms, statistical analysis, etc., the litigator may well feel like a cog in one of Goldberg's devices. Interestingly enough, a common characteristic found in many of Goldberg's contraptions is an animal performing some element of the convoluted process.[163]

[76] If that is not enough, consider that Practice Point 5 from the Sedona Conference, advising the use of an electronic search and retrieval method, does not guarantee that all responsive documents will be found and that results will be uniform.[164] Thus, even after performing these tedious and

[158] *Id.* at 361.
[159] *Id.*
[160] *Id.*
[161] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 261 n.10 (D. Md. 2008).
[162] Rube Goldberg: Biography, http://www.rubegoldberg.com (follow "About Rube" hyperlink) (last visited Mar. 31, 2009).
[163] AbsoluteAstronomy.com, Rube Goldberg, http://www.absoluteastronomy.com/topics/Rube_Goldberg (last visited Mar. 31, 2009).
[164] *Sedona Conference Best Practices Commentary*, *supra* note 1, at 211.

> The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover,

complicated tasks, the results could still be imperfect. Take heart, however, that, as stated in Practice Point 5, there is no requirement for perfect searches.[165] The only requirement is that the "parties act reasonably in the good faith performance of their discovery and legal obligations."[166]

[77] Surely, practitioners would prefer that the perfect technology could be invented so that a term, topic, or concept could be entered and all of the relevant documents within a population related to the search criteria could be found. But this is not likely to happen. Even if a cap could be invented that would fit on a person's head and formulate the perfect search plan, without being versed in the linguistic nuances of the case, knowledge about the make-up of the ESI population to be searched, as well as technical knowledge about the workings of the particular search engine to be used, the likely answer would be a blank sheet of paper.

[78] If there is a lesson to be learned from the aforementioned cases, it is that the real problem is not the technology. It is not a failure of the technology to find the documents, but rather, it is a failure in the process of designing the search. Interestingly enough, there is quite a variety of failures in the process. It ranges from simply not properly identifying the population in which to search, as was the case in *Rhoads*, where only sender and recipient addresses were considered,[167] to the kind of analysis performed being ill-suited to keyword searches in the first place, as was the case in *Victor Stanley*.[168] In between, there are a range of situations where the process requirements of keyword search efforts were not understood and included in that process—things like synonyms, iterative testing, and sampling.

---

> differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval.

*Id.*

[165] *See id.* ("Just as with past practice involving manual searches through traditional paper document collections, there is no requirement that 'perfect' searches will occur . . . .").

[166] *Id.*

[167] *See* Rhoads Indus., Inc. v. Bldg. Materials Corp. of Am., 254 F.R.D. 216, 221-22 (E.D. Pa. 2008).

[168] *See* Victor Stanley, Inc. v. Creative Pipe, Inc. 250 F.R.D. 251, 262 (D. Md. 2008).

[79] Certainly the technology can result in capabilities like Boolean logic, proximity locators, stemming, and fuzzy logic. There could be ways for the technology to link into thesauri in order to better find relevant documents. Technology can even facilitate iterative testing and sampling with an index-based search engine. It is unlikely, however, that technology can replace disciplined testing and sampling of the methodology in order to validate its adequacy.

[80] There are still further complications when it comes to formalizing the search plan into some kind of negotiated protocol. The difficulty is not only how to perform the initial linguistic analysis, but how to assess the data population and test the adequacy of the search terms prior to formalization. Perhaps the answer is an iterative approach to developing the protocol where it identifies the various process stages and the manner in which the separate baselines will be formalized. In other words, the initial protocol would classify the identification of each baseline in the process. The ultimate goal would be the development of the final baseline that would then be formalized into the actual search protocol. If properly implemented, the development cycle would likely have five baselines:

> (1) Identify the various subject matters of interest to the case and for which discovery was needed;
>
> (2) Identify linguistic analysis and preliminary search term formulation;
>
> (3) ESI evaluation where the data types and likely stores of relevant ESI are determined and matched to an appropriate search engine;
>
> (4) Search term testing and validation; and
>
> (5) Final search plan formalization followed by the execution of the actual plan.

Even the development of these baselines could require some iteration. In other words, the results obtained during the fourth baseline effort, search term testing and validation, could require a return to an earlier baseline like linguistic analysis for further synthesis and revision.

## VIII. CONCLUSION

[81] The use of keyword search terms in litigation has received careful attention over the years. As the prevalence of digital data increases in litigation, along with the practice of electronic discovery, the use of keyword search terms will be relied upon more and more to winnow the wheat from the chaff. Although such an effort is seemingly simple, in order to successfully implement and sustain a keyword search plan, there are actually very complex issues and numerous criteria that must be navigated. It is not a pure technology problem. Rather, it is a problem of process, analogous to tire performance.

[82] The processes of balancing and alignment are essential to proper tire performance, and hence good driving. To improve driving, one does not need a better tire, just reliable processes of improving tire performance, like balancing and alignment. And rather than performing these processes on their own, car drivers utilize experts trained and equipped with the right tools to perform the task, so that once completed, the drivers can then be off to the races. Perhaps there is a lot to be learned from this analogy, which not only applies to keyword search terms, but to all of cyber litigation.