June 2015

Electronic Data, Electronic Searching, Inadvertent Production of Privileged Data: A Perfect Storm

Donald Wochna

ELECTRONIC DATA, ELECTRONIC SEARCHING, INADVERTENT PRODUCTION OF PRIVILEGED DATA: A PERFECT STORM

WHY ATTORNEYS ARE BEING FORCED TO RECOGNIZE THAT SEARCHING ELECTRONICALLY STORED INFORMATION IS AN EXPERT FUNCTION

Donald Wochna

I. INTRODUCTION

Recent case law,[1] changes in civil procedural rules,[2] and the dramatic increase in the volume of electronically stored information[3]

1. *See* United States v. Gainer, 468 F.3d 920 (6th Cir. 2006); United States v. O'Keefe, 252 F.D.R. 26 (D.D.C. 2008); Stanley v. Creative Pipe, Inc., 250 F.D.R. 251 (D. Md. 2008).

2. *See* Fed. R. Civ. P. 26(f) ("Under Rule 26(f), parties must sit down together at an early 'meet and confer' conference to discuss a range of issues involving electronically stored

have combined to form a "perfect storm" in which to trap unwary attorneys into potentially committing malpractice. Faced with enormous volumes of client data that must be reviewed for privilege and unacceptably high costs of manual review, many attorneys are relying upon electronic searches to identify privileged documents within large client data sets.[4] Recent case law discussed herein analyzes this type of electronic searching and concludes that it is an expert function.[5] Attorneys who fail to treat electronic searching as an expert function may be unable to defend their electronic search protocols when challenged and, as a consequence thereof, may incur sanctions, including loss of attorney-client privilege protection for client documents inadvertently produced in litigation.[6] This article examines the case law analyzing electronic searching of client data and concludes that treating electronic searching as an expert function is consistent with the

---

information. Such a conference is intended to be broad in scope and to cover the gamut of preservation, scope, formatting, and accessibility issues.").

3. George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, *12-13 (2007).

> Probably close to 100 billion e-mails are sent daily, with approximately 30 billion e-mails created or received by federal government agencies each year. The amount of stored information continues to grow exponentially. Perhaps more easily grasped, the amount of information in business has increased by thousands, if not tens of thousands of times in the last few years.

Id.

4. The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 192 (2007) [hereinafter *Best Practices*] ("Discovery of relevant information gathered about a topic in dispute is at the core of the litigation process. However, the advent of 'e-discovery' is causing a rapid transformation in how that information is gathered. While discovery disputes are not new, the huge volume of available electronically stored information poses unique challenges.").

5. *See* O'Keefe, 537 F. Supp. 2d 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. . . . 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 the 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.

Id.

6. Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 261 n.10 (D. Md. 2008).

> [J]udge 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.

Id.

requirements of Evidence Rule 702. This article suggests that the practical impact of treating electronic searching as an expert function is to permit attorneys to focus and strategize on the process of electronic searching rather than on the completeness of document production. In effect, electronic searching permits attorneys to quit focusing on finding documents and begin focusing on identifying electronic sources of information on which reside relevant documents that can be extracted by means of electronic searching protocols.

II. THE CHALLENGE OF MANUAL REVIEW: VOLUME

There appears to be no serious case law discussion of the minimum competency standards required to search paper documents or physical file cabinets for data relevant to a matter or privileged as attorney-client communication.[7] It is beyond cavil that, in a world of paper documents, it has been standard procedure for attorneys to manually review data prior to production in litigation and determine whether the data was privileged, relevant, confidential, etc.[8] As clients migrated from typewriters to word processors to computers, not only has the media on which data resides changed, but the volume of client data has exploded:

> The shift of information storage to a digital realm has, for a variety of reasons, caused an explosion in the amount of information that resides in any enterprise profoundly affecting litigation. This massive amount of electronically stored information is distributed broadly among different storage devices, from large mainframe computers, to tiny machines capable of storing information equivalent to several warehouses of documents each, all of which are or can be integrated into other systems. These systems are complex, interdependent, and evolve spontaneously, like ecosystems. It is often impossible to find one person, or even one discrete group of people, who completely understand the working of this new form of "information ecosystem.[9]

---

7. *Best Practices*, *supra* note 4, at 193.

> Just a few years ago all information was stored on physical records such as paper. . . . It was reasonable, and indeed relatively easy in all but the exceptional case, for the legal profession to gather and then manually review all the individual items collected as part of the discovery process prior to their production.

Id.

8. *Id.* ("Discovery has changed. In just a few years, the review process needed to identify and produce information has evolved from one largely involving the manual review of paper documents to one involving vastly greater volumes of electronically stored information.").

9. *Id.*

Although client data has undergone a radical transformation from discrete pieces of paper to an "information ecosystem," attorneys have generally continued to manually review client electronic data for privilege, treating the electronic data in the same manner as they have reviewed paper documents for generations.[10] The impact of treating electronic information as if it were the same as paper documents is most significant in the manual review of data for privilege.

Much of the manual review of client data occurs as part of the general discovery process.[11] The United States Supreme Court has long held that discovery of data relevant to a matter and in the possession, custody, or control of a litigant was a necessary part of litigation in order to ensure open, efficient, and fair dealings within the federal court system.[12] Under the *Hickman* view of litigation, "every party to a civil action is entitled to the disclosure of all relevant information in the possession of any person, unless the information is privileged."[13] The goal of liberalized discovery was to avoid surprise and to "make a trial less a game of blind man's bluff and more a fair contest with the basic issues and facts disclosed to the fullest practicable extent."[14] Discovery of relevant information has become the way for litigants to obtain the fullest possible knowledge of the issues before trial, while permitting attorneys and clients to preserve privileged communications.[15]

---

10. *See Best Practices*, *supra* note 4. *See also supra* note 7 and accompanying text.
11. *See Best Practices*, *supra* note 4. *See also supra* note 8 and accompanying text.
12. Hickman v. Taylor, 329 U.S. 495, 501 (1947).

> The various instruments of discovery now serve (1) as a device, along with the pre-trial hearing under Rule 16, to narrow and clarify the basic issues between the parties, and (2) as a device for ascertaining the facts, or information as to the existence or whereabouts of facts, relative to those issues. Thus civil trials in the federal courts no longer need to be carried on in the dark. The way is now clear, consistent, with recognized privileges, for the parties to obtain the fullest possible knowledge of the issues and facts before trial.

Id.

13. *Id.* at 507-08.

> [T]he deposition-discovery rules are to be accorded a broad and liberal treatment. No longer can the time-honored cry of "fishing expedition" serve to preclude a party from inquiring into the facts underlying his opponent's case. Mutual knowledge of all the relevant facts gathered by both parties is essential to proper litigation. To that end, either party may compel the other to disgorge whatever facts he has in his possession. The deposition-discovery procedure simply advances the stage at which the disclosure can be compelled from the time of trial to the period preceding it, thus reducing the possibility of surprise. . . . And as Rule 26(b) provides, further limitations, come into existence when the inquiry touches upon the irrelevant or encroaches upon the recognized domains of privilege.

Id.

14. United States v. Proctor & Gamble Co., 356 U.S. 677, 682 (1958).
15. *See Hickman*, *329 U.S. at 501*.

As the world transitioned from paper documents to electronically stored information, the Federal Rules of Civil Procedure, generally, were interpreted to accommodate that change as part of the discovery process.[16] Rule 34 of the Federal Rules of Civil Procedure and its state-law counterparts were generally interpreted to include electronic information within the definition of "data compilation."[17] As a result, "data compilations" were deemed to be documents just like traditional paper documents and subject to discovery and production.[18] Just like paper documents, "data compilations" needed to be reviewed and privileged client data identified and excluded from production to a party opponent in litigation.[19]

It was not long before the unique features of electronic data began to interfere with the review and production of data. The volume of electronic information compared to paper documents, the redundancy of multiple electronic copies of the same information, the lack of a coherent filing system in which electronic information may be stored, and the unique cost issues associated with electronic information storage systems and media that have become obsolete were the primary reasons that electronically stored information was difficult to review for privilege and produce.[20] Although these features of electronic data

16. Roland Bernieri, *Avoiding an E-Discovery Odyssey*, 36 N. KY. L. REV. 491, 495 (2009).

> [T]he legal community has attempted to address the effect of technology on discovery issues. In August 2004, an advisory committee published a proposed set of amendments for the Federal Rules of Civil Procedure designed to guide courts and attorneys on issues associated with electronic discovery. The committee passed a revised set, and ultimately these were adopted by the U.S. Supreme Court without a substantive modification.

Id.

17. The Sedona Conference, *Foreword to Second Edition of The Sedona Principles: Best Practices Recommendations & Principles for Addressing Electronic Document Production*, The Sedona Conference Working Group Series, June, 2007, at iv [hereinafter *Sedona Principles Second Edition*]. ("When the Working group began its deliberations, the starting point was that under Rule 34 and many of its state counterparts, all 'data compilations' were deemed documents just like traditional paper documents and subject to discovery.").

18. *See Best Practices, supra* note 4. *See also* supra note 8 and accompanying text.

19. *Sedona Principles Second Edition*, *supra* note 17, at page iv ("This equal treatment suggested that electronic information should be searched for, processed, and produced like paper.").

20. Byers v. Ill. State Police, 2002 WL 1264004, at *10 (N.D.Ill. May 31, 2002).

> [T]he Court is not persuaded by the plaintiffs' attempt to equate traditional paper-based discovery with the discovery of e-mail files. Several commentators have noted important differences between the two. . . . Chief among these differences is the sheer volume of electronic information. . . . Additionally, computers have the ability to capture several copies (or drafts) of the same email, thus multiplying the volume of documents. . . . Also, unlike most paper-based discovery, archived e-mails typically lack a coherent filing system. Moreover, dated archival systems commonly store information on magnetic tapes which have become obsolete. Thus, parties incur additional costs in translating the data from the tapes into useable form.

dramatically increased the cost of privilege review, the consequences of failing to adequately review client data continued to threaten attorneys and clients with the draconian results of privilege waiver.[21]

As technical challenges to the production of electronically stored information were encountered, a body of research and law began to be created giving some guidance to attorneys regarding the choices and decisions necessary to produce electronically stored information in discovery.[22] Privilege review, however, has only recently been addressed by case law.

Whether electronically stored information can be reviewed by attorneys for privilege or relevancy in a manner identical to the review of paper documents has not been the subject of much research, and still less case law.[23] This may be because, until recently, attorneys were manually reviewing all electronically stored information prior to production.[24] It is highly doubtful, however, whether manual review of documents for privilege can survive the increase in volume of data that has occurred as the result of the ubiquitous use of computers and electronic communication networks that form the new "information ecosystem."[25] Indeed the manual review for privilege of ever-

---

Id.

21. Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 267-8 (D. MD 2008) "[T]he court finds that the Defendants waived any privilege or work product protection for the 165 documents at issue by disclosing them to the Plaintiff."); *see also* FED. R. CIV. P. 26(b)(5) Advisory Committee Note ("The Committee [on the Rules of Practice and Procedure] has repeatedly been advised that the risk of privilege waiver and the work necessary to avoid it, add to the costs and delay of discovery. When the review is of electronically stored Information, the risk of waiver, and the time and effort required to avoid it, can increase substantially because of the volume of electronically stored Information and the difficulty in ensuring that all information to be produced has in fact been reviewed.").

22. *Sedona Principles Second Edition*, *supra* note 17, at page iv. ("Far from supplanting *The Sedona Principles*, the new Federal Rules have highlighted the many areas of electronic discovery in which there is continued and growing need for guidance.").

23. *See* Roland Bernier, *Avoiding an E-Discovery Odyssey*, 36 N. KY. L. REV. 491 (2009); George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH J.L. & TECH. 10 (2007); Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.D.R. 251 (D. Md. 2008); United States v. Ganier III, 468 F.3d 920 (6th Cir. 2006); U.S. v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008).

24. *See Best Practices*, *supra* note 4. *See also supra* notes 7-8 and accompanying text.

25. *Best Practices*, *supra* note 4, at 193.

> [W]ith the digital revolution there has also been a paradigm shift in the review process which is feasible. The shift of information storage to a digital realm has, for a variety of reasons, caused an explosion in the amount of information that resides in any enterprise-profoundly affecting litigation. This massive amount of electronically stored information is distributed among different storage devices. . . . These systems are complex, interdependent, and evolve spontaneously, like ecosystems.

Id.

increasingly larger volumes of electronic information has become the single most costly step in the production of electronically stored information.[26]

III. THE COST OF MANUAL REVIEW OF ELECTRONICALLY STORED INFORMATION

The cost of manual review is driven initially by the sheer volume of data that can now be stored on very small devices.[27] For example, manually reviewing one gigabyte of electronic documents can be estimated to cost a client about $32,000 of attorney time and labor.[28] This estimate is based upon a common assumption that one gigabyte of data constitutes 80,000 to 100,000 pages of data.[29] A single attorney ought to be able to review 500 pages of data per hour with acceptable accuracy.[30] One gigabyte of data, therefore, will require one attorney to spend 160 to 200 hours reviewing the data and identifying whether it is privileged.[31] At an average billable rate of $200 per hour, one attorney can review one gigabyte of data at a cost of between $32,000 to $40,000.[32]

Given the cost of manual review, it is not surprising that the continued use of this procedure is becoming (and in many cases, has already become) and unacceptable cost of litigation. In one recent case, for example, a litigant spent eighteen months and $11.4 million to hire contract attorneys to review electronic documents culled from 127

---

26. Paul & Baron, *supra* note 3, at 4 ("Litigators can no longer depend on manual review alone. It is too time-consuming and expensive – with costs often exceeding the amount in dispute.").

27. *Best Practices*, *supra* note 4, at 198 ("In many organizations, the average works maintains several gigabytes of stored data. At the same time, the costs of storage have plummeted from $20,000 per gigabyte in 1990 to less than $1 per gigabyte today.").

28. Paul & Baron, *supra* note 3, at 20.

> Take then, for example, litigation in which the universe subject to search stands at one billion e-mail records, at least 25% of which have one or more attachments of varying length. Generously assuming a model reviewer is able to review an average of fifty e-mails, including attachments per hour. Without employing any automated computer process to generate potentially responsive documents, the review effort for this litigation would take 100 people, working ten hours a day, seven days a week, fifty two weeks a year, over fifty-four years to complete. And the cost of such review, at an assumed average billing rate of $100/hour, would be $2 billion.

Id.

29. *See id.*

30. *See id.*

31. *See id.*

32. See *id.*

document custodians for privilege prior to production.[33] Cases involving terabytes of data (one terabyte = 1000 gigabytes) will require tens of millions of dollars to manually review.[34] It has become obvious to anyone that is familiar with these changes and costs that the litigation system cannot continue to operate under these strictures.[35] Cost has gotten so significant that manual review of large datasets for privilege, relevance, or work product has been characterized by at least one group as "indefensible."[36]

In response to the cost of manual review, attorneys are being forced to leverage technology and "use computers and not just associates, contract lawyers, or outsourced offshore workers to search [client data]."[37] Generally, attorneys use computers to search client data by running keyword searching software programs to identify documents responsive to requests for production of documents.[38] The most common form of electronic search tool is a software program that accepts "keywords" or phrases and identifies instances of those words or phrases in the client data.[39] The keywords and phrases can be either simple words, word combinations, or may contain Boolean and related operators.[40] While the use of this type of keyword searching has long been used to search for relevant case law in computerized legal libraries, its use to identify privileged and work product documents within the client data set is relatively new. Some commentators have duly noted that keyword searching case law libraries is significantly different than

33. Oracle v. SAP AG, 2009 WL 3009059, at *15 (N.D. Cal. 2009).

34. *Id.* at *2 (discussing the "huge" production of data in the case).

35. Paul & Baron, *supra* note 3, at 20 ("The numbers add up to more than a burden than any party should assume, no matter how rich in resources, without changes being made in the way cases are litigated and to techniques used in discovery.").

36. *Best Practices*, *supra* note 4, at 199 ("Although the continued use of manual search and review methods may be indefensible in discovery involving significant amounts of electronically stored information, merely adopting sophisticated automated search tools, alone, will not necessarily lead to successful results.").

37. Paul & Baron, *supra* note 3, at 36.

38. *Id.* at 37 ("The legal profession has adopted keyword searching in light of its longtime familiarity with its use in connection with the offerings of the major online legal retrieval services, which allow for searches to be made of structured databases containing case precedent and statutory authority.").

39. *Best Practices*, *supra* note 4, at 200

> By far the most commonly used search methodology today is the use of "keyword searches" of full text and metadata as a means of filtering data for producing responsive documents in civil discovery. . . . [T]he use of the term 'keyword searches' refers to set-based searching using simple words or word combinations, with or without Boolean and related operators.

Id.

40. *Id.* at 21.

keyword searching client data, primarily because the language in case law is much more structured and predictable than the language used in communications and documents created by employees in a workplace environment.[41]

Relying upon the results of a keyword search tool or any other form of electronic search protocol to identify documents that a litigant claims are privileged or work product necessarily exposes that search tool or electronic protocol to analysis when its results are challenged.[42] This analysis has only recently been the subject of a few cases in which courts have begun to define the nature of electronic searching and the minimum competency necessary to defend search results.[43] A detailed discussion of each of these cases reveals a common thread: Configuring legally defensible electronic search strategies is an expert function, significantly different than the expertise needed to review paper documents. The failure to recognize the expert nature of electronic searching may lead attorneys to construct search strategies that cannot be defended when challenged.

IV. NATURE OF ELECTRONIC SEARCHING: EXPERT LANGUAGE AND EXPERT FUNCTION

The first case to define some characteristics of electronic search tools and protocols was a criminal matter in which the defendant challenged the testimony of a prosecution witness.[44] In the *Ganier* case, the prosecution sought to elicit the testimony of Special Agent Wallace Drueck regarding the electronic searches that defendant Ganier had run on the defendant's computer and the deletion of certain data relevant to a grand jury investigation.[45] Basically, the prosecution sought to introduce

41. Rich & Baron, *supra* note 3, at 38 ("First, and most importantly, there are profound issues of ambiguity and indeterminacy in human language, and thus it all texts in large, heterogeneous databases subject to discovery. . . . Furthermore, people make up words on the fly, including new codes that function as language.").

42. Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.D.R. 251, 262 (D. Md. 2008).

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

Id.

43. *See id.*; United States v. Ganier, 468 F.3d 920 (6th Cir. 2006); United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C 2008).

44. Ganier,468 F.3d 920.

45. *Id.* at 924.

Agent Drueck's testimony to link, in time and by subject, the search activity on Ganier's computer with the grand jury deliberations.[46] The defense objected to Drueck's testimony on the ground that the testimony was admissible only under Evidence Rule 702, as expert testimony, and the prosecution had neither properly identified Drueck as an expert witness, nor properly tendered an expert report as required by Rule 18 of the Federal Rules of Criminal Procedure.[47] The trial court agreed with the defense and dismissed the case.[48] On appeal, the Sixth Circuit Appellate Court examined the issue whether Agent Drueck's testimony was of such a character that it required he be admitted as an expert under Evidence Rule 702.[49]

The prosecution argued that Agent Drueck was a fact witness, not an expert witness, because Agent Druek merely launched certain special software to run over the defendant's computer and then observed the results:

> The government argues that Drueck's proposed testimony is not based on scientific, technical, or other specialized knowledge, but is simply lay testimony available by "running commercially-available software, obtaining results, and reciting them." The government contends that this testimony is of the same type as "facts . . . that could be observed by any person reasonably proficient in the use of commonly used computer software, such as Microsoft Word and Microsoft Outlook (such as the existence and location of multiple copies of documents that are identical or virtually identical to the allegedly 'deleted' documents)," which Ganier previously indicated he did not consider to be expert testimony.[50]

The Sixth Circuit Appellate Court analyzed the issue whether the proposed testimony of Agent Drueck was expert testimony by reviewing the type of knowledge that Agent Drueck would necessarily apply to the

---

46. *Id.* at 924-25.
47. *Id.* at 924.
48. *Id.*
49. United States v. Ganier, 468 F.3d 920, 925 (6th Cir. 2006).
> We must first determine whether the district court erred by concluding that Drueck's proposed testimony fell within the scope of Federal Rule of Criminal Procedure 16(a)(1)(G). Rule 16(a)(1)(G) requires, in part, that "[a]t the defendant's request, the government must give to the defendant a written summary of any testimony that the government intends to use under Rules 702, 703, or 705 of the Federal Rules of Evidence during its case-in-chief at trial."

Id.
50. *Id.* at 925-26.

output/report generated by the "commercially-available software" used by Agent Drueck.[51]

> The reports generated by the forensic software display a heading, a string of words and symbols, a date and time, and a list of words. The government asserts that these reports reveal three different types of searches performed with particular search terms at particular times, but such an interpretation would require Drueck to apply knowledge and familiarity with computers and the particular forensic software well beyond that of the average layperson. This constitutes "scientific, technical, or other specialized knowledge" within the scope of Rule 702.[52]

In addition to the special knowledge and familiarity with computers that the court determined was an integral part of Agent Drueck's testimony, the Sixth Circuit Appellate Court also analyzed the language that Agent Drueck would necessarily use in his testimony to explain the actions taken to search and destroy information on defendant's computer.[53] It is useful to note that the court analogized the language used by Agent Drueck to describe computer-related facts with the specialized language used by police officers to explain drug arrests.[54] The Appellate Court affirmed the designation made by the trial court that Agent Drueck's testimony was properly suppressed pursuant to Evidence Rule 702, requiring that he be identified as an expert.[55]

Applying the *Ganier* analytic paradigm to an explanation of the use of electronic search tools by attorneys to identify privileged, relevant, or work product documents, requires that we identify the nature of the "knowledge" and "vernacular" that would need to be used to explain the use of such tools.[56] The most common type of electronic search

---

51. *Id.* at 926.
52. *Id*.
53. *Id.*
54. United States v. Ganier, 468 F.3d 920, 926 (6th Cir. 2006).

> Because the categorization of computer-related testimony is a relatively new question, comparisons with other areas of expert testimony are instructive. Software programs such as Microsoft Word and Outlook may be as commonly used as home medical thermometers, but the forensic tests Drueck ran are more akin to specialized medical tests run by physicians. The average layperson today may be able to interpret the outputs of popular software programs as easily as he or she interprets everyday vernacular, but the interpretation Drueck needed to apply to make sense of the software reports is more similar to the specialized knowledge police officers use to interpret slang and code words used by drug dealers.

Id.

55. *Id.* at 927.
56. *Id.* at 926.

methodology used today is "keyword" searches.[57] Keywords can be used by themselves or combined with "operators" to construct search engines which are then applied to data sets.[58] Many courts embrace keyword searching as an electronic search protocol and many attorneys attempt to agree upon the keywords that will be used for purposes of production in discovery.[59]

Defending the use of specific keywords requires an analysis of the metrics by which keyword searching is measured. Keyword search measurements include specialized concepts, such as basic information retrieval metrics of "recall" and "precision." Recall is a measure of completeness—namely, how well an electronic search protocol has identified all the potentially responsive documents from the client data set.[60] It is derived by dividing the number of responsive documents retrieved by the total number of responsive documents.[61] "Precision" is a measure of efficiency—namely how well an electronic search protocol has identified responsive documents as a percentage of the total number of documents retrieved, including all false positives.[62] It is derived by dividing the number of responsive documents retrieved by the total number of documents retrieved.[63]

Explaining the manner in which search terms were chosen may require use of specialized language. "Ambiguity" and "variation" are common characteristics of language that need to be incorporated into electronic search protocols to render them effective for particular client

57. *See Best Practices*, *supra* note 4. *See also supra* note 38 and accompanying text.

58. *Best Practices*, *supra* note 4, at 207 ("First, there are keyword based methods, ranging from the simple use of keywords alone, to the use of strings of keywords with what are known as 'Boolean operators' (including AND, OR, 'AND NOT' or 'BUT NOT').").

59. *Id.* at 200 ("Courts have not only accepted, but in some cases ordered, the use of keyword searching to define discovery parameters and resolve discovery disputes.); *see also* Balboa Threadworks v. Stucky, 2006 WL 763668, at *5 (D. Kan 2006).

> As to the formulation of a search protocol, whether one using keyword searches and/or other search procedures, the parties are directed to meet and confer in an attempt to agree on an appropriate protocol, and should lean heavily on their respective computer experts in designing such a protocol. Numerous types and varieties of search protocols have been discussed and adopted by courts and these may guide the parties in designing a search protocol to be used in this case.

Id. *See e.g.*, Rowe Entm't v. William Morris Agency, 205 F.R.D. 421, 432-22 (S.D.N.Y.2002); Simon Prop. Group L.P. v. MySimon, Inc., 194 F.R.D. 639, 641-44 (S.D.Ind.2000); Playboy Enters. v. Welles, 60 F.Supp.2d 1050, 1053-55 (S.D.Cal.1999); Antioch Co. v. Scrapbook Borders, Inc., 210 F.R.D. 645, 653-54 (D. Minn. 2002)).

60. Paul & Baron, *supra* note 3, at 41.

61. *Best Practices*, *supra* note 4, at 205.

62. Paul & Baron, *supra* note 3, at 41.

63. *Best Practices*, *supra* note 4, at 205.

data sets.[64] "Ambiguity" refers to the tendency of words and expressions to have different meanings when in different contexts.[65] Each context is a "variation."[66] One of the compelling characteristics of language is the ability to use many different words and expressions to convey content.[67] Configuring electronic search protocols to effectively identify privileged, relevant, or work product documents in the client data set requires that the search protocols reflect the ambiguity and variation of language used in the client data set.[68]

Specialized language may be necessary to explain the manner in which ambiguity and variation were recognized in the client data set and incorporated into the search. "Taxonomies" and "ontologies" are essentially synonyms of words and relevant classes of words that are developed and included in electronic search strategies to refine the search by maximizing recall and precision.[69]

Specialized statistical concepts may be necessary to explain the basis upon which the size of statistically significant random samples of client data sets were calculated.[70] Increasing reliance upon sampling of electronically stored information was expressly incorporated into amendments to Rule 34:

> Rule 34(a)(1) is also amended to make clear that parties may request an opportunity to test or sample materials sought under the rule in addition to inspecting and copying them. That opportunity may be important for both electronically stored information and hard-copy materials. The current rule is not clear that such testing or sampling is authorized; the amendment expressly permits it. As with any other form of discovery, issues of burden and intrusiveness raised by requests to test or sample can be addressed under Rule 26(b)(2) and 26(c).[71]

---

64. *Id.* at 206 ("The richness of human language causes a severe challenge in identifying informational records.").

65. *Id.*

66. *Id*.

67. *Id.* at 207 ("But as the Blair and Maron study demonstrates, human language is highly ambiguous and full of variation.").

68. *Id.* ("In the years since Blair and Maron, the IR community has been engaged in research and development methods, tools, and techniques that compensate for endemic ambiguity and variation in human language, and thus maximize the recall and precision of searches.").

69. Paul & Baron, *supra* note 3, at 43.

70. *Best Practices*, *supra* note 4, at 207 ("[T]here are a variety of statistical techniques, which analyze word counts.").

71. FED. R. CIV. P. 34, Advisory Committee's Note.

Explaining the manner in which a client data set was randomly sampled or electronic search results were applied to the client data set almost surely require specialized language of the type that the *Ganier* court labeled as "expert testimony."[72]

Finally, explaining the type of electronic search that was used in a particular case and comparing that type of search with other searching methods will require specialized language.

> Even before the emergence of the Web, information retrieval science has constituted a vast and growing field . . . . However, broadly speaking, information retrieval methods fall into three broad classes: set-theoretic (Boolean strings, supplemented by fuzzy search capabilities), algebraic (premised on the mathematical idea that the meaning of a document can be derived from the constituent terms in a document, and thus weighting retrieval by the proximity of a document's terms in the form of two or higher dimensional maps, as in vector space modeling), and probabilistic (using language models and Bayesian belief networks, the latter of which involves make educated inferences about the relevance of future documents based on prior experience in reviewing documents in a given collection.[73]

The *Ganier* court's "specialized language" analysis, when applied to the testimony, language, and vernacular required to defend the use of electronic search tools, including keyword searching, by attorneys for purposes of identifying privileged, work product, or relevant data, appears to characterize such testimony as Rule 702 expert testimony.[74] Court decisions subsequent to *Ganier* have analyzed the specific types of expert testimony that will be required in order to successfully defend particular keyword searching protocols.

V. ANALYSIS OF NATURE OF KEYWORD SEARCHING: EXPERT FUNCTION COMBINING LINGUISTICS, STATISTICS, AND COMPUTER TECHNOLOGY

In *US v O'Keefe*, Judge Facciola analyzed a defendant's challenge to the electronic search protocols used by the Department of State to locate all information in its possession custody or control related to O'Keefe's indictment charging he expedited visa requests in exchange for gifts.[75] In his analysis, Judge Facciola set forth the scope of the

72. *See supra* note 52 and accompanying text.
73. Paul & Baron, *supra* note 3, at 42
74. *See supra* notes 51-52 and accompanying text.
75. United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008).

technical character and specialized features required of electronic search protocols, including keyword searches.[76]

> As noted above, defendants protest the search terms the government uses. 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.[77] Indeed, a special project team of the Working Group on Electronic Discovery of the Sedona Conference is studying that subject and their work indicates how difficult this question is.[78] 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.[79]

Judge Facciola's analysis in *O'Keefe* is focused on the nature of keyword searching and concludes that keyword searching is an expert function because it relies upon the application of specialized knowledge and concepts.[80] Judge Facciola incorporated into the *O'Keefe* decision the research and knowledge of the combined areas of computer technology, linguistics, and statistics, and determined that challenges to the use of electronic search protocols must be based on expert testimony.[81]

After the *Ganier* and *O'Keefe* cases, the issue remained whether creating keyword search protocols was, itself, an expert function demanding special competencies on the part of the attorneys or law

---

76. *Id.*

77. *See* Paul & Baron, *supra* note 3.

78. *See Best Practices*, *supra* note 4.

79. *Id.* at 23-24.

80. *Id.* at 24.

> Accordingly, if the 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.

Id.

81. *Id.* ("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.").

firms creating such searches. In May 2008, Magistrate Judge Grimm decided that issue in *Victor Stanley v. Creative Pipe et al*—a case addressing inadvertent production of privileged data to a party opponent in response to a request for production of documents.[82]

In the *Victor Stanley* matter, defendants Creative Pipe Inc. and Mark and Stephanie Pappas produced data to plaintiff Victor Stanley, Inc. in response to plaintiff's request for production of documents.[83] Prior to producing the data, defendants' counsel conducted an electronic search of the client data for privileged documents.[84] Unfortunately, the electronic search did not identify all privileged material, and 165 privileged documents were disclosed to plaintiff.[85] Defendants requested the return of the 165 documents, but plaintiff insisted that the privilege had been waived by disclosure.

The issue before Magistrate Judge Grimm was whether defendants had waived the attorney-client privilege by reason of their inadvertent production of the privileged documents.[86] Basically, if the defendants had acted in a reasonable manner to prevent the inadvertent disclosure, there would be no waiver of privilege. Whether defendants acted reasonably, in turn, required the court to analyze the manner in which defendants created their keyword search strategy.[87]

To create the keyword search, counsel for Creative Pipe met and conferred with their client and with co-defendant Mark Pappas.[88] Together they devised a keyword search strategy to locate privileged document consisting of seventy keywords that they believed ought to identify all privileged data.[89] Counsel ran those keywords over all client documents and any document that contained one or more of the keywords was withheld from production on the ground of privilege.[90] It is significant to note that the privilege search undertaken in the *Victor Stanley* case appears to be identical to the manner in which any law firm might use electronic keyword searches to identify privileged documents contained within a client data set of documents relevant to litigation.

Although counsel was in control of the client's data set, counsel took no action other than to "guess" the keywords that ought to be used

82. Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251 (D. Md. 2008).
83. *Id.* at 255.
84. *Id.*
85. *Id.*
86. *Id.* at 257.
87. Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 259-61 (D. Md. 2008).
88. *Id.* at 254.
89. *Id.*
90. *Id.*

to identify privileged documents. Counsel took no actions to determine the actual language used by the client's employees to create the data in the client data set. No attempt was made to identify any taxonomies or ontologies; no sampling was done to identify the ambiguities or variations used by the creators of the client data. It is also interesting to note that this technique (meeting to confer and "guess" keywords) is the manner in which most litigation counsel agrees with a party opponent regarding search terms to be used to identify potentially relevant data as part of discovery. Indeed, in *Victor Stanley*, the litigants had previously met and agreed upon search terms designed to locate potentially relevant data that would then need to be reviewed for privilege.[91] The fact that the seventy keywords used by counsel in *Victor Stanley* did not completely identify all privileged documents indicates that there was sufficient ambiguity in the client data set to defeat the keyword search. The issue for the court, however, was whether the keyword strategy used by counsel and client was defensible, notwithstanding its failure to capture all privileged documents.[92]

In order to decide whether the seventy keywords comprised a defensible search of the client data set for privileged documents, the court looked to the defendants to provide the court with information about the people and the process that was used to create (or guess) the seventy keywords.[93] Magistrate Judge Grimm demanded that defendants produce evidence in the nature or testimony that demonstrated the protocols chosen by defendants were appropriate for the task.[94] It is significant—indeed essential—for attorneys to note that the process of keyword searching for privileged information was not shielded by work product or attorney-client privilege. Nor was the process of constructing

---

91. *Id.*

92. *Id.* at 257.

93. *Victor Stanley, Inc.*, 250 F.R.D. at 256.

> While it is known that M. Pappas (a party) and Mohr and Schid (attorneys) selected the keywords, nothing is known from the affidavits provided to the court regarding their qualifications for designing a search and information retrieval strategy that could be expected to produce an effective and reliable privilege review.

Id.

94. *Id.* at 262.

> Use of search and information retrieval methodology, for the purpose of identifying and withholding privileged or work product protected information from production, requires the utmost care in selecting methodology that is appropriate for the task because the consequence of failing to do so, as in this case, may be the disclosure of privileged/protected information to an adverse party, resulting in a determination by the court that the privilege/protection has been waived.

Id.

keywords assumed to be a function subsumed by the ordinary practice of law—such as one might characterize the manual review of client data by attorneys in the past. Rather, the court characterized the use of electronic search tools as a *process* that must be defended and explained as any other process or methodology:

> Selection of the appropriate search and information retrieval technique requires careful advance planning by persons qualified to design effective search methodology. The implementation of the methodology selected should be tested for quality assurance; and the party selecting the methodology must be prepared to explain the rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented.[95]

Applying this *process-oriented analysis* to the facts in the *Victor Stanley* case, Magistrate Judge Grimm held that counsel for defendants had:

> [F]ailed to provide the court with information regarding: the keywords used; the rationale for their selection; the qualifications of M. Pappas and his attorneys to design an effective and reliable search and information retrieval method; whether the search was a simple keyword search, or a more sophisticated one, such as one employing Boolean proximity operators; or whether they analyzed the results of the search to assess its reliability, appropriateness for the task, and the quality of its implementation.[96]

The *Victor Stanley* court grounded its process-oriented analysis upon the science of information retrieval. The court demanded defendants provide testimony in support of the search protocols used—not mere legal argument.[97] Citing Judge Faciolla in *United States v. O'Keefe*, Judge Grimm noted, in the *Creative Pipe* case, that keyword searches may, indeed, be the proper method for searching in a matter; but "there are well-known limitations and risks associated with them, and proper selection and implementation obviously involves technical, if not scientific knowledge."[98]

> Judge Facciola made the entirely self-evident observation that challenges to the sufficiency of keyword search methodology

95. *Id.*

96. *Id.* at 259-60.

97. *Id.* at 260 ("While keyword searches have long been recognized as appropriate and helpful for ESI search and retrieval, there are well-known limitations and risks associated with them, and proper selection and implementation obviously involves technical, if not scientific knowledge.").

98. *Id.* at 260.

> unavoidably involve scientific, technical and scientific [sic] 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.[99]

VI. SEARCHING ELECTRONICALLY STORED INFORMATION IS AN EXPERT PROCESS

The *Victor Stanley* court is not alone in characterizing electronic searching as a process that must be defended as an expert process. The Sedona Conference's "Practice Point 7" related to Search and Information Retrieval Methods also describes electronic searching as a process or methodology based on the science of information retrieval.

> Counsel should be prepared to explain what keywords, search protocols, and alternative search methods were used to generate a set of documents, including ones made subject to subsequent manual searches for responsiveness and privilege. This explanation may best come from a technical "IT" expert, a statistician, or an expert in search and retrieval technology. Counsel must be prepared to answer questions, and indeed, to prove the reasonableness and good faith of their methods.[100]

Characterizing electronic searching as an expert process subjects the search to analysis and challenges, requires the search process be defended, and triggers significant implications for attorney issues related to malpractice.[101]

---

99. *Victor Stanley, Inc.*, 250 F.R.D. at 261 n.10.

100. *Best Practices*, *supra* note 4, at 212.

101. *See* Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 136 (S.D.N.Y. 2009).

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

Id.

VII. CHALLENGING ELECTRONIC SEARCH PROCESSES AS REASONABLE

In any particular case, challenges to electronic search protocols will probably be raised by a party opponent to measure the degree to which a responding party reasonably searched client data, especially if there has been little or no discussion amongst or between counsel and little or no transparency of the search tools and protocols.[102] Traditionally, the producing party to a discovery request has enjoyed a presumption that it knows best the location and method by which to identify and produce documents.[103] However, this presumption may not always apply to the use of electronic search tools.[104] While courts and litigants may be willing to accept representations and statements on the record that counsel has performed a manual review of client data in a competent manner, recent case law discussed herein suggests no similar deference is afforded electronic searching.[105] Perhaps this is because courts and litigants intuitively understand that searching electronically stored information is an expert process much more difficult to properly design and execute than the manual review of paper documents. Additionally, the efficacy of manual review was never directly challenged but was hidden behind professional representations and assumptions of competency. When using electronic search tools to perform electronic searching functions, however, the efficacy of the search can be addressed directly and measured.[106]

---

102. *See* Smith v. Life Investors Ins. Co. of Am., 2009 WL 2045197, at *7 (W.D. Pa. 2009).

> Defendant must do more than summarily list the number of pages it has produced and the time and effort it has invested. Rather, Defendant has a burden to demonstrate that its search for documents was reasonable. A thorough explanation of the search terms and procedures used would be a large step in that direction.

Id.

103. *See* FED. R. CIV. P. 26. The language used 26(a)(1) and 26(a)(2) states: "[A] party must disclose." *Id.* There is a presumption based on this language that a disclosing party is aware of what information they need to disclose and therefore is best suited to determine the location and method of producing the discovery items. However, 26(c) permits a court to intervene in the discovery process if necessary to facilitate disclosure of necessary information. *Id.*

104. *Best Practices*, *supra* note 4, at 204.

> Absent agreement, a party has the presumption, under Sedona Principal 6, that it is in the best position to choose an appropriate method of searching and culling data. However, a unilateral choice of a search methodology may be challenged due to lack of a scientific showing that the results are accurate, complete, and reliable.

Id.

105. *See* United States v. Gainer, 468 F.3d 920 (6th Cir. 2006); United States v. O'Keefe, 252 F.D.R. 26 (D.D.C. 2008); Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.D.R. 251 (D. Md. 2008).

106. *Best Practices*, *supra* note 4, at 205.

> One can often adjust a system to retrieve more documents, thereby increasing recall, but at the expense of retrieving more irrelevant documents, and thus decreasing precision.

Challenges to the effectiveness of electronic search methodologies can reasonably be expected to focus upon the expert nature of electronic searching and upon the reasonableness of the search and configuration decisions made while using the tools and protocols.[107] Indeed, a "challenging party may argue that the process used by the responding party is essentially an expert technology which has not been validated by subjecting it to peer review, and unbiased empirical testing or analysis."[108]

Automated software solutions that enter the marketplace may also be challenged as a process—and some fear that these challenges could be difficult to overcome.

> The probability of such a challenge is greater if the technology is patented or proprietary to a developer or vendor (i.e. in a so-called "Black Box"). In such circumstances, e-discovery and litigation support vendors that use these technologies may be several degrees of separation from the original developers. A requesting party may demand the responding party to "prove up" the use of such search technology. This could set the stage for a difficult and expensive battle of experts.[109]

Perhaps future electronic searching software will need to follow the path of computer forensic software.[110] Forensic software is used by experts to identify, preserve, and extract relevant electronic data (content and artifacts).[111] Expert opinions based upon the results of using the

---

> One can cast either a narrow net and retrieve fewer relevant documents along with fewer irrelevant documents, or cast a broader net and retrieve more relevant documents, but at the expense of retrieving more irrelevant documents.

Id.

107. *See Best Practices*, *supra* note 4. *See also supra* note 100 and accompanying text.

108. *Best Practices*, *supra* note 4, at 204.

109. *Id*.

110. See, e.g., United States v. Tank, 200 F.3d 627 (9th Cir. 2000) (validating the use of Encase software to create image of Defendant's computer and the authenticity of computer evidence in the general context of Fed. R. Evid. 901(a)); Smith v. Slifer Smith & Frampton/Vail Assocs. Real Estate, LLC, 2009 WL 482603 (D.Colo. 2009) (upholding bad faith in the destruction of documents on the part of the respondent in producing computer records related to discovery).

111. *See* Hanger Prosthetics & Orthotics, Inc. v. Capstone Orthopedic, Inc., 2008 WL 2441067, at *3 (E.D. Cal. 2008).

> Alcock's proposed testimony also includes identifying "files and fragments of files previously deleted from the Laptop," which he indicates involved expert reasoning since his work was "a time-consuming process due to the number of computer drives and files involved, and the complexity of retrieving files and artifacts damaged due to the attempts of sterilizing the drive to conceal or deprive the use of data once present on the laptop." . . . Accordingly, Alcock's proposed testimony constitutes expert testimony.

Id.

software are usually validated by independent means, thereby authenticating the results of the software.[112]

In a similar manner, future electronic searching software tools may need to be validated by comparing the ability of the software to locate relevant documents in a universe of test data in which the identity of all relevant documents is known. The challenges of developing such electronic searching software, however, are formidable. They include the tremendous flexibility of the language and the creativity of humans which continues to frustrate attorneys who attempt to "rationally" guess effective keywords. For example, attorneys appear to be only 20 percent effective "at thinking up all of the different ways that document authors could refer to words, ideas, or issues in their case."[113]

> The limitations on search and retrieval methodology exposed in the Blair and Maron study was not the ability of the computer to find documents that met the attorneys' search criteria, but rather the inability of the attorneys and paralegals to anticipate all the possible ways that people could refer to the issues in the case. The richness of human language causes a severe challenge in identifying informational records.[114]

If search and information retrieval methods are measured against the accuracy of attorneys "guessing" the language used by key players, then it may be relatively easy to demonstrate sufficient accuracy and precision to satisfy a Daubert/Frye challenge.

> The Daubert challenge raised by [Judge] Facciola, then, may be met not by judging the scientific validity of a search engine in an absolute way, but by judging how valid it is to suit the purposes of e-discovery production, an undertaking which involves many factors, such as the costs in time, money and energy to the producing party and their marginal benefit to the requesting party and the litigation, that have no bearing on the scientific validity of the search engines.[115]

112. Leonard Deutchman, *When E-Discovery Is Put to the Test*, L. TECH. NEWS 1, May 14, 2008 (discussing issues of authentication related to the proprietary nature of search engines).

113. *Best Practices*, *supra* note 4, at 206 (citing David Blair and M.E. Maron BART study, at 1985).

114. *Id*.

115. Leonard Deutchman, *supra* note 112.

VIII. DEFENDING ELECTRONIC SEARCH PROCESSES AS REASONABLE AND FEDERAL EVIDENCE RULE 502(B)

Challenges to electronic search methodologies will require litigants to defend their electronic search processes and will force counsel to consider evidentiary issues at the beginning of the process, when electronic searching protocols are being created or negotiated.[116] Based upon the language analysis in *Ganier*,[117] and the expert process analysis in *O'Keefe* and *Victor Stanley*, prudent attorneys will treat electronic searching as an expert function comprised of skills in the area of computer technology, linguistics, and statistics.[118] Prudent attorneys will base their electronic search strategies upon the advice of an expert or other authoritative source that is willing and able to defend those search strategies when challenged.

> It cannot credibly be denied that resolving contested issues of whether a particular search and information retrieval method was appropriate—in the context of a motion to compel or motion for protective order—involves scientific, technical or specialized information. If so, then the trial judge must decide a method's appropriateness with the benefit of information from some reliable source—whether an affidavit from a qualified expert, a learned treatise, or, if appropriate, from information judicially noticed.[119]

Requiring litigants to defend their search protocols with expert testimony is similar to requiring expert testimony to explain and defend random sampling protocols.[120] The requirement is a direct and necessary result of the court's recognition of the technical aspects of electronically searching data. While some attorneys may view the requirement as a burden, Magistrate Judge Grimm suggests that this requirement ought to benefit the discovery process by reducing costs through cooperation between or amongst litigants.[121]

---

116. *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.D.R. 251, 262 (D. Md. 2008). *See also supra* note 42 and accompanying text.

117. *See* United States v. Ganier III, 468 F.3d 920 (6th Cir. 2006).

118. *See* United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008); *Victor Stanley, Inc.*, 250 F.R.D. at 251.

119. *See Victor Stanley, Inc.*, 250 F.R.D. at 261 n.10. *See also supra* note 6 and accompanying text.

120. In re Vioxx Prods. Liab. Litig., 2006 WL 1726675, at *2 n.5 (5th Cir. 2006) ("By random sampling, we mean adhering to a statistically sound protocol for sampling documents. . . . The parties must provide expert assistance to the district court in constructing any protocol.").

121. *Victor Stanley, Inc.*, 250 F.R.D. at 261 n.10. *See also supra* note 6 and accompanying text.

If attorneys are to defend the process of electronically searching client data sets, they will need to better understand the features of that process. Compiling an exhaustive analysis of the many features of electronic searching is beyond the scope of this article, but a few fundamental features ought to be noted by all attorneys. For example, a key feature of electronic searching is its iterative nature.[122] Rarely, if ever, will an initial keyword search yield satisfactory results. Keyword searches are notoriously over or under-inclusive.[123] Part of the problem is the ambiguity of language; another is the failure of attorneys to recognize that the client data set, itself, can be analyzed for information to accurately create keyword searches.

For example, client data can be indexed. Indexing the entire client data set identifies every word in every document, accurately states the number of times the word appears, and keeps track of the documents in which the word resides and the key player that created the document. Rather than guess the keywords that client personnel may have used to create relevant or privileged data, attorneys could use indexing tools to know the universe of words actually used in the client data set and the frequency of their use. While this method is used commonly to locate "code" words or phrases used by cliques or clans in networks,[124] it ought to be included in every electronic search strategy to help lessen the "guesswork" from keyword searches. This technique might also be considered by courts that mandate litigants agree to search terms as part of discovery conferences. Rather than forcing litigants to "guess" at the language used by their respective clients to designate relevant, work product, or privileged matters, courts perhaps could agree upon an iterative, index-enhanced, protocol that would substantially improve keyword searching.[125]

Successfully defending electronic search methodologies will be especially important in light of changes to the federal rules of evidence

---

122. Paul &Baron, *supra* note 3, at 50.

123. *Best Practices*, *supra* note 4, at 201.

> [A]lthough 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 wind 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.

Id.

124. Wouter de Nooy, Andrej Mrvar &, Vladimir Batagelj, *Exploratory Network Analysis with Pajek* 73 (2005).

125. Paul & Baron, *supra* note 3, at 50.

that prohibit the use of inadvertently produced client data only if, *inter alia*, reasonable precautions were taken to avoid the disclosure. New Evidence Rule 502(b) was designed to respond to:

> widespread complaints that litigation costs for review and protection of material that is privileged or work product have become prohibitive due to the concern that any disclosure of protected information in the course of discovery (however innocent or minimal) will operate as a subject matter waiver of all protected information.[126]

As amended, Rule 502(b) provides that the inadvertent disclosure of privileged information in a federal proceeding, or to a federal officer or agency, does not waive the attorney-client privilege if:

(1). The disclosure is inadvertent;
(2). The holder of the privilege or protection took reasonable steps to prevent disclosure, and
(3). The holder promptly took reasonable steps to rectify the error, including (if applicable) following Fed. R. Civ.P. 26(b)(5)(B)[127]

Whether the holder of the privilege took "reasonable" steps to prevent disclosure will be the focus of analysis on a case by case basis.[128] So long as the steps taken can be proven to be reasonable, Rule 502(b)(2) ought to be satisfied. As noted by the Advisory Committee to Rule 502, the rule does not explicitly codify the reasonable test, because the rule is really a set of non-determinative guidelines that vary from case to case.[129]

Rule 502(b) clearly invites attorneys to anticipate the technical, linguistic, and statistical challenges related to the use of electronic search tools, and create an electronic search process that can be defended in any particular case.[130] Evidence Rule 502(b) is an attempt to provide attorneys some relief from the overwhelming task of manually reviewing all client documents for privilege by expressly protecting client privilege while using reasonable electronic search protocols.

Evidence Rule 502(b) appears to incorporate the "expert process analysis" set out in *O'Keefe* and *Victor Stanley* with a particular emphasis upon computer technology to derive electronic search

126. FED. R. EVID. 502(b), Advisory Committee's Note.
127. FED. R. EVID. 502(b).
128. *Id.*
129. *Id.*
130. Other electronic discovery processes may also fail to be reasonable. *See* Amersham Biosciences Corp v. Perkinelmer, Inc., 2007 WL 329290 (D.N.J. 2007) (corrupt files included in production set that were readable by receiving party not reasonably protected from disclosure).

solutions that will be reasonable. It states: "A party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken 'reasonable steps' to prevent inadvertent disclosure."[131] Greater use of sampling[132] and the implementation of an efficient system of records management may also be relevant to the issue whether reasonable precautions have been taken to avoid disclosure of privileged data in any particular case.

IX. CONCLUSION

As the volume of client data increases in litigation, economic pressure to reduce or eliminate manual review of client data for privilege, work product, and relevance will increase. Attorneys will be forced to use electronic searching tools and protocols to identify privileged, work product, or relevant data. These electronic tools, however, are fundamentally different from manual review. Electronic search and information retrieval tools represent an expert process that can be properly used and defended only if attorneys recognize that these tools must be used and configured in accordance with properly designed search protocols, results measured in accordance with accepted metrics such as recall and precision, and implemented in a technically valid manner. Challenges to the use of electronic search and information retrieval protocols will focus upon their technical features and will force attorneys to recognize that electronic searching is an expert process.

By focusing upon the expert process of electronic searching, and by judging the "reasonableness" of that process, courts are properly moving away from focusing discovery on measurements of the completeness of production. This shift in focus represents a significant "relief" from the economics of manual review. By creating an electronic search and information retrieval process that is defensible for the particular case in which it is to be used, attorneys will be able to incorporate technology into discovery and "dial in" the amount of precision and recall necessary. The end result will be an a reasonable process, of sufficient scope, precision, and recall to satisfy discovery without undue burden and expense.

131. FED. R. EVID. 502, Advisory Committee's Note.
132. Paul & Baron, *supra* note 3, at 47.