University of Richmond
UR Scholarship Repository

Law Student Publications

School of Law

Fall 2012

Technologies-That-Must-Not-Be-Named: Understanding and Implementing Advanced Search Technologies in E-Discovery

Jacob Tingen
University of Richmond

Follow this and additional works at: http://scholarship.richmond.edu/law-student-publications

Part of the Internet Law Commons, and the Legal Ethics and Professional Responsibility Commons

TECHNOLOGIES-THAT-MUST-NOT-BE-NAMED: UNDERSTANDING AND IMPLEMENTING ADVANCED SEARCH TECHNOLOGIES IN E-DISCOVERY

By Jacob Tingen*

I. INTRODUCTION

[1]   The Federal Rules of Civil Procedure were created to promote the "just, speedy, and inexpensive determination of every action and proceeding."[1]   Unfortunately, in the world of e-discovery, case determinations are often anything but speedy and inexpensive.[2]   The manual review process is notoriously one of the most expensive parts of litigation.[3]   Beyond expense, the time and effort required to carry out large-scale manual review places an immense burden on parties, nearly destroying the possibility of assessing the merits of early settlement before expensive review has already been carried out.[4]   Due to the difficulty inherent in the manual review process and the potential for human error,

---

* Jacob Tingen is a licensed Virginia attorney and a graduate of the University of Richmond School of Law. In the summer of 2011 he interned with Vault26, an e-discovery startup, where he consulted on current e-discovery practices. Living on the cutting edge of technology, Jacob maintains a home on the web at http://jacobtingen.com. He would like to thank Professor James Gibson for his guidance and help in preparing this article.

[1] FED. R. CIV. P. 1.

[2] David Degnan, *Accounting for the Costs of Electronic Discovery*, 12 MINN. J.L. SCI. & TECH. 151, 152 (2011).

[3] *See* George L. Paul & Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, ¶ 4 (2007), http://law.richmond.edu/jolt/v13i3/article 10.pdf (noting that manual review is too time-consuming and expensive).

[4] *See, e.g., id.* at ¶ 20 (providing an example showing the time it takes for manual review of one billion e-mail records).

courts have become tired of seeing what they view as incompetence among attorneys.[5] All acknowledge that technology is the main culprit—e-mail alone produces *100 billion new messages daily*.[6] At the same time, technology may in fact provide the solution to the e-discovery problem.[7]

[2]   In response to the e-discovery challenge, courts and commentators have begun to refer to "new" and "emerging" search technologies.[8] Some tout them as the holy grail of e-discovery, while others dismiss the new technologies as unfit for the task or unable to compete with the raw capability of hundreds of attorneys reviewing documents for hours on end.[9] Even now, doubts exist as to whether new technologies really can help resolve the difficulties experienced by attorneys tasked with increasingly demanding discovery requests.[10]

[3]   Even for those who are aware of the existence of advanced search and review tactics beyond keyword search, many questions remain for attorneys and judges alike. First, what are the new and emerging technologies? While courts and commentators mention the existence of

---

[5] *See* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, 17 RICH. J.L. & TECH. 9, ¶ 13 (2011), http://jolt.richmond.edu/v17i3/article9.pdf.

[6] *See* Paul & Baron, *supra* note 3, at ¶ 12.

[7] *See, e.g.*, H. Christopher Boehning & Daniel J. Toal, *Assessing Alternative Search Methodologies*, N.Y. L.J. TECH. TODAY, Apr. 22, 2008, at 5.

[8] *See* discussion *infra* Part II.

[9] *See* Boehning & Toal, *supra* note 7, at 6 (comparing classic Boolean keyword searching with new technological approaches to e-discovery).

[10] *See* Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, 17 RICH. J.L. & TECH. 11, ¶ 1 (2011), http://jolt.richmond.edu/v17i3/article11.pdf (stating that there has been little scientific evidence proving whether advanced search and review tactics are more effective than keyword search and manual review).

the technologies, there is not much guidance with regard to what the new technologies are and what they accomplish.[11] Second, are the new technologies superior to the manual review process? Understandably, attorneys are hesitant to use an unfamiliar e-discovery product that may not work better than the e-discovery process to which they are already accustomed.[12] Third, if attorneys do use a new search and review process, what standards of accuracy or defensibility is a court likely to impose? When managing the discovery process, attorneys want to be sure that the method of production satisfy the expectations of the court.[13]

[4]   This article answers those questions. It demonstrates that attorneys have a legal duty to understand and use advanced conceptual search and review technologies as part of an e-discovery review process when dealing with large amounts of information. It then briefly explains how those technologies actually work, why they are superior in both accuracy and efficiency to traditional manual review, and how one can defend use of these new technologies in court.

[5]   Part II of this article discusses the need for lawyers to reconsider which search methodologies to use in the e-discovery review process. It reveals that lawyers currently have a duty to understand technology to competently represent their clients and argues that this duty should extend

---

[11] *See* Baron, *supra* note 5, at ¶ 34 (noting that only two cases even mention the existence of conceptual search technologies). Since the publication of Jason R. Baron's article in 2011, two additional cases have spoken in more detail regarding the use of advanced search technologies. *See* Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *1, *12 (S.D.N.Y. Feb. 24, 2012) (approving the use of predictive coding in e-discovery for the first time); Case Management Order: Protocol Relating To Production of Electronically Stored Information at 1-26 Actos (Pioglitazone) Products Liability Litigation, No. 6:11-md-2299 (W.D. La. July 27, 2012) [hereinafter *Actos* Order] (emphasizing the importance of collaboration when using an advanced e-discovery process for all pending and future related litigation involving Actos products).

[12] *See* Grossman & Cormack, *supra* note 10, at ¶ 1.

[13] *See* Baron, *supra* note 5, at ¶ 37.

to a cursory understanding of e-discovery search tactics. It discusses the reluctance of the legal community to adopt new search technologies for a variety of reasons, including economic concerns and lack of experience with technology. It briefly explains recent judicial decisions advocating the use of advanced search and review technologies.

[6]   Part III provides a background of advanced search technologies, some explanation of what they are, and information on their use in the e-discovery context. It analyzes recent research finding that advanced search and review methodologies are more effective than a keyword process followed by extensive manual review. Furthermore, it discusses steps to ensure that counsel's implementation of advanced search technologies will be defensible in court.

[7]   Finally, Part IV addresses some concluding issues. It identifies when advanced search technologies should be used as opposed to other search and review methods. It discusses the issue of attorney-client privilege and argues that courts should be lenient when evaluating whether privilege has been waived by inadvertently produced documents after an advanced search of millions of documents. It argues that when practitioners properly implement advanced search technologies, they meet their legal duty and help further the goals of the Federal Rules of Civil Procedure by making e-discovery more efficient, more accurate, and less expensive.

II. Adoption of New Search Technologies

[8]   "Lawyers need to rethink how they perform 'searches.'"[14] Familiar with keyword and Boolean search operations from widespread experience with popular legal research services, attorneys tend to apply the same skill set when they approach e-discovery.[15] Unfortunately,

---

[14] Paul & Baron, *supra* note 3, at ¶ 36.

[15] *Id.* at ¶ 37.

simple keyword searches followed by extensive manual review have proven inadequate when it comes to finding the responsive documents necessary to litigate a case on its merits.[16] Overcoming the shortcomings of keyword search and the high expense of complete manual review has become an important goal in e-discovery practice.[17] Courts and commentators have pointed to emerging search and review technologies as the answer to the manual review problem.[18] In effect, attorneys must have a basic understanding of e-discovery and the available search technologies to competently represent their clients.[19]

A. A Legal Duty To Use Advanced Search Technologies?

[9]   Requiring attorneys to have a foundational understanding of technology is not without precedent.[20] In the seminal *Zubulake* cases, Judge Scheindlin went so far as to delineate a new legal duty, requiring attorneys to understand their client's technology infrastructure.[21] *Zubulake*, while instructive mainly from the context of determining when the duty to preserve is triggered, also provides helpful background in

---

[16] *See id.* at ¶ 40.

[17] The Sedona Conference, *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 194 (2007) [hereinafter *Best Practices*].

[18] *See* Paul & Baron, *supra* note 3, at 36-37.

[19] *See* Monica Bay, *Georgetown E-Discovery Conference Opens With Case Law Update*, LAW TECH. NEWS (Nov. 18, 2011), http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202532791193 (quoting U.S. District Court Judge James Francis: "I don't see how you can provide competent representation if you don't have some basic understanding of e-discovery.").

[20] *See, e.g.*, Zubulake v. UBS Warburg LLC, 229 F.R.D. 422, 432 (S.D.N.Y. 2004).

[21] *Id.*

examining whether attorneys have a duty to cultivate an understanding of technology.[22]

[10]   The plaintiff in *Zubulake* leveled charges of gender discrimination against her former employer in August of 2001.[23] While the factual and procedural background makes for an interesting read for anyone interested in e-discovery, the primary thrust of the e-discovery problems in the case arose from the plaintiff's request for certain e-mails which the defendant repeatedly failed to produce.[24] The 2006 Amendments to the Federal Rules of Civil Procedure were unavailable to the parties involved, and as a result, Judge Scheindlin's commentary throughout the entire series of *Zubulake* cases in some way set the stage for the new rules and continues to prove influential in modern e-discovery practice and discussion.[25] In particular, Judge Scheindlin held that "counsel must become fully familiar with her client's document retention policies, as well as the client's data retention architecture."[26] That's legalese for saying lawyers must learn to speak tech.[27]

[11]   In the future, lawyers must become competent when dealing with and talking about technology.[28]   Judge Scheindlin clarified this expectation in 2004 when she said that during the discovery process,

---

[22] *Id.* at 441.

[23] *Id*. at 425.

[24] *Id*. at 425-29.

[25] *See* RALPH C. LOSEY, INTRODUCTION TO E-DISCOVERY: NEW CASES, IDEAS, AND TECHNIQUES 441-42 (2009) [hereinafter LOSEY, INTRODUCTION TO E-DISCOVERY].

[26] *Zubulake*, 229 F.R.D. at 432.

[27] *See* RALPH C. LOSEY, E-DISCOVERY CURRENT TRENDS AND CASES 56 (2008) [hereinafter LOSEY, CURRENT TRENDS].

[28] *See Zubulake*, 229 F.R.D. at 440.

attorneys must speak with their client's information technology personnel to learn about their client's system-wide information storage procedures and policies.[29] In short, attorneys have a legal duty to understand technology.[30] This article argues that this duty should also extend to understanding and implementing "emerging" search technologies.

[12]   In the years since *Zubulake*, the field of e-discovery has experienced further advances in research and sophistication, including the 2006 Amendments to the Federal Rules of Civil Procedure,[31] guidelines and standards developed by the Sedona Conference,[32] a rising level of education among the bench,[33] and the development of new technologies to assist in searches.[34]

[13] The *Sedona Conference Cooperation Proclamation* ("*Proclamation*") suggests a more collaborative approach to e-discovery in litigation.[35] Endorsed by judges across many jurisdictions, the *Proclamation* promotes education in e-discovery technology to ensure the

---

[29] *Id.* at 432.

[30] *Id.*

[31] *See generally* FED. R. CIV. P. (2006); *see also* LOSEY, CURRENT TRENDS, *supra* note 27, at 241-63 (explaining the 2006 amendments to the Federal Rules of Civil Procedure).

[32] *About The Sedona Conference*, THE SEDONA CONFERENCE, http://www.thesedona conference.org (last visited Oct. 27, 2012) [hereinafter *About The Sedona Conference*].

[33] *See* THE SEDONA CONFERENCE, THE SEDONA CONFERENCE COOPERATION PROCLAMATION 4-11 (2008) [hereinafter COOPERATION PROCLAMATION], *available at* https://thesedonaconference.org/publication/The%20Sedona%20Conference%C2%AE% 20Cooperation%20Proclamation (listing judicial endorsements of the Cooperation Proclamation).

[34] *See* Paul & Baron, *supra* note 3, at ¶ 66.

[35] *See* COOPERATION PROCLAMATION, *supra* note 33, at 1-3.

"just, speedy, and inexpensive determination of every action."[36]   In particular, the *Proclamation* identifies the need for cooperation and understanding between not only plaintiff and defense counsel, but also among technology professionals.[37] Furthermore, it advocates educating attorneys about the tools available through law school programs and classes to help new lawyers understand the technical, legal, and cooperative aspects of e-discovery, as well as programs to help businesses understand how to manage their electronic records.[38] The need for training with regard to e-discovery strategies and technologies is widely expressed, and endorsed, by the judiciary in many states.[39]

[14]   Indeed, some believe that the "legal profession is at a crossroads: the choice is between continuing to conduct discovery as it has 'always been practiced' . . . or, alternatively, embracing new ways of thinking in today's digital world."[40] Clients can no longer bear the mounting costs of e-discovery, and overburdened judges are beginning to recommend newer search and review methodologies to attorneys.[41] Extensive manual review of every document in litigation is already impossible in many cases and manual review guided by keyword search alone has proven ineffective in

---

[36] *Id.* at 3 (quoting FED. R. CIV. P. 1).

[37] *See id.*

[38] *See id.*

[39] *Id.* at 4-11 (providing a detailed list of judicial endorsements).

[40] THE SEDONA CONFERENCE, COMMENTARY ON ACHIEVING QUALITY IN THE E-DISCOVERY PROCESS 1 (2009), *available at* https://thesedonaconference.org/publication/ The%20Sedona%20Conference%C2%AE%20Commentary%20on%20Achieving%20Qu ality%20in%20the%20E-Discovery%20Process [hereinafter ACHIEVING QUALITY].

[41] *See id.*

others.[42] The attorney of the future must embrace new technologies or face being drowned in an overwhelming sea of data.[43]

B. Resistance To New Search Technologies

[15]   Despite the need for adoption of better technologies, some attorneys assert that keyword and Boolean searches are the industry standard and that newer technologies are cost-prohibitive and less accurate.[44] This assertion is incorrect.[45] In the face of a growing amount of evidence showing that new search technologies can make the e-discovery process easier and more efficient,[46] the legal community tends to push back against newer search technologies for a variety of reasons.

[16]   First, the manual review process is notorious for being the most expensive piece of an e-discovery request.[47] With upwards of fifty-percent of e-discovery costs attributed to the manual review process, an attorney's potential earnings can be tough to ignore.[48] The conflict between the legal industry's self interest and the just, speedy, and "*inexpensive* determination" of a case creates serious ethical concerns.[49] Typical manual review costs can range anywhere from two hundred and

---

[42] *See* Paul & Baron, *supra* note 3, at ¶¶ 4, 39-40.

[43] *See id.* at ¶ 36.

[44] *Cf.* Boehning & Toal, *supra* note 7.

[45] *See* Grossman & Cormack, *supra* note 10, at ¶ 52; *see also* discussion *infra* Part III.

[46] *See* Grossman, & Cormack, *supra* note 10, at ¶ 52; *see also discussion infra* Part III.

[47] *See* Degnan, *supra* note 2, at 161.

[48] *See id.*

[49] FED. R. CIV. P. 1.

fifty dollars to five hundred dollars per hour to scan through mountains of documents, a process which can take months.[50] In many situations, law firms charge a premium to boost profits. For example, one contract attorney recently learned that her firm billed its client two hundred and fifty dollars per hour during a manual review while only paying her thirty-five dollars per hour.[51] Firms clearly have an economic incentive to continue using a manual review process that has a potential for huge profits. Acknowledging that new search technologies are more effective than manual review may mean giving up revenues the legal industry is accustomed to receiving.[52]

[17]   Other attorneys may not like the idea of learning a new set of technologies. In general, lawyers are not known for being tech savvy.[53] Some commentators have mentioned their dismay with the legal profession's inability to keep up with the technology industry.[54] Perhaps in e-discovery, this failure to keep up with newer technologies results from over-familiarity with keyword search.[55] Many attorneys are of the opinion that keyword search is the industry standard and that it effectively finds the majority of relevant documents in a given data set.[56] Recognizing that a better method exists may amount to a significant investment of time,

---

[50] *See* Degnan, *supra* note 2, at 160.

[51] *See, e.g.*, Kashmir Hill & David Lat, *Top Lawyers*, WASHINGTONIAN (Sept. 6, 2012, 1:27 PM), http://www.washingtonian.com/print/articles/6/171/14536.html.

[52] *See* Justin Scheck, *Tech Firms Pitch Tools For Sifting Legal Records*, WALL ST. J., (Sept. 6, 2012, 1:35 PM), http://online.wsj.com/article/SB121936262421062033.html.

[53] *See* LOSEY, INTRODUCTION TO E-DISCOVERY, *supra* note 25, at 72.

[54] *See, e.g.*, *id.*

[55] *See* Paul & Baron, *supra* note 3, at ¶ 37.

[56] *See id.* at ¶¶ 37, 40.

classes, and hardware in order to understand and implement new technologies.[57]

[18]   Even though more e-discovery resources are available today than ever before, some attorney behavior demonstrates a lack of understanding in how to meet a client's e-discovery needs.[58] In many cases, counsel's "apparent lack of savvy" is to blame for overbroad, expensive, or poorly implemented discovery.[59] For example, in a 2010 case, it seemed that both the court and counsel involved were unaware of the possibility of using alternate search methodologies to assist in a more accurate or expedited review.[60]

[19]   In fact, with merely two exceptions, there were no judicial opinions prior to 2012 that even mentioned the use of alternative search methods to expedite document review, much less explain what those search methods might be or provide guidance on how to implement them.[61] Only very recently has counsel received explicit judicial approval of the use of advanced search methodologies in e-discovery, as evidenced by Judge Peck's groundbreaking opinion in *Moore v. Publicis Groupe*.[62]

---

[57] *See* COOPERATION PROCLAMATION, *supra* note 33, at 3.

[58] *See* Baron, *supra* note 5, at ¶ 13.

[59] *Id.*

[60] *See id.* at ¶ 14 (citing Helmert v. Butterball, LLC, No. 4:08CV00342 JLH, 2010 WL 2179180, at *1-5 (E.D. Ark. May 27, 2010)).

[61] *See id.* at ¶ 34.

[62] Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *1, *12 (S.D.N.Y. Feb. 24, 2012). The parties in *Moore* have hotly contested the judicial order in the case, and even though predictive coding met with Judge Peck's approval, it is now uncertain whether the parties will even use an advanced search and review methodology. Since this article's writing, another case has emerged where the court approved the use of predictive coding. *See Actos* Order, *supra* note 11. In *Actos*, the order emphasizes the

C. Judicial Approval of Advanced Search Technologies

[20]   The first two opinions to broach the subject of the potential of advanced search technologies address the issue only anecdotally.[63] In *Disability Rights Council of Greater Washington v. Washington Metropolitan Transit Authority*, an advocacy group brought disability discrimination claims against the transit authority.[64] During discovery, the plaintiff requested information that could only be found on backup tapes because the original e-mails in question had been destroyed.[65] The court ordered restoration and search of the backup tapes.[66] In its order, the court requested that the parties consider how the information on the backup tapes would be searched and directed the parties to recent scholarship arguing that conceptual search technologies could provide more efficient, comprehensive, and accurate results than a keyword search process.[67]

[21]   The only other case to recommend alternative or advanced search methodologies was Judge Grimm's decision in *Victor Stanley, Inc. v. Creative Pipe, Inc.*[68] Included in Judge Grimm's criticism of the plaintiff's discovery efforts, he repeatedly discussed the lack of qualification of the members in the plaintiff's party to build a targeted

---

collaboration between the parties that made the use of predictive coding in the e-discovery process possible. *Id.*

[63] *See* Baron, *supra* note 5, at ¶¶ 34-35 (discussing *Disability Rights Council of Greater Washington v. Washington Metropolitan Transit Authority*, 242 F.R.D. 139 (D.C. Cir. 2007) and *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D. 251 (D. Md. 2008)).

[64] *Disability Rights Council*, 242 F.R.D. at 141.

[65] *Id.* at 145-46.

[66] *Id.* at 148.

[67] *Id.*

[68] *See generally Victor Stanley, Inc.*, 250 F.R.D. 251.

search.[69] This language highlights the expectation that attorneys be competent or seek competent help in conducting e-discovery. Furthermore, Judge Grimm cites the potential shortcomings of keyword search and mentions other options that counsel could use in the e-discovery process.[70] In a footnote, the opinion explains some of the potential search alternatives currently available.[71]

[22]   In contrast, Judge Peck's decision in *Moore* is the first judicial opinion to approve a document review process that leverages advanced search and review technologies.[72] The basic facts of the case along with a summary of Judge Peck's discussion of the application of advanced search and review technologies are outlined below. Because his opinion provides guidance as to how counsel should proceed when using a technology-assisted review process, it is addressed further throughout this article.[73]

[23]   In *Moore*, plaintiffs claimed that the defendant violated numerous gender discrimination laws.[74] As part of their discovery effort, plaintiffs sought numerous e-mails and other electronically stored information to prove the gender bias.[75] During the parties' discussion as to how the requested information should be reviewed, plaintiffs raised objections to the defendant's proposed use of technology-assisted document review in

---

[69] *See id.* at 256.

[70] *Id*. at 259-60.

[71] *Id.* at 259 n.9.

[72] *See* Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *1 (S.D.N.Y. Feb. 24, 2012).

[73] *See id.* at *8-12.

[74] *Id.* at *1.

[75] *See generally id.* at *4-5.

the form of predictive coding.[76] The court took an active role in the discovery dispute, pointing out that advanced search and review technologies often lead to better results than traditional keyword search and document review, and encouraged the parties to continue to work out an acceptable discovery plan.[77] During various discovery conferences, the parties and the court discussed how to proceed with discovery issues, such as the number of custodians and other sources of electronically stored information ("ESI"), the number of phases in which to review documents, the predictive coding or technology-assisted review process, and the level of transparency in the review process.[78] At various points in the opinion, the court emphasized that advanced search and review technologies typically produce more accurate results than keyword search and manual review.[79] Finally, the court ordered the parties to go forward with their agreed upon technology-assisted review process, becoming the first court to judicially approve the use of advanced search and review technologies.[80]

[24]   Judge Peck's order has recently come under intense scrutiny by both the plaintiff's attorneys in *Moore* and the legal community at large.[81] Even though U.S. District Judge Andrew Carter initially confirmed the order, Judge Peck granted a motion to stay discovery after the plaintiff's continued calls for his recusal and for revision of the e-discovery

---

[76] *Id.* at *3-6.

[77] *See Moore*, 2012 WL 607412, at *3.

[78] *See id.* at *3-6.

[79] *See, e.g., id.* at *10-11.

[80] *See id.* at *12.

[81] *See, e.g.,* Alison Frankel, *That federal court e-discovery breakthrough? Not so fast…,* THOMSON REUTERS (May 15, 2012), http://newsandinsight.thomsonreuters.com/ Legal/News/ViewNews.aspx?id=47523.

protocol.[82] Despite the predictable pushback by the plaintiffs in this case, attorneys should recognize that widespread use of advanced search technologies in e-discovery will one day be the standard;[83] it simply makes more sense to use a specialized machine to find a needle in a haystack as opposed to manually searching through each individual piece of hay.[84]

[25]   Much of the commentary already examined, as well as the opinions coming from the bench, provides the clear message that, "[l]awyers [*still*] need to rethink how they perform 'searches.'"[85] Even with this clear instruction to use new technology, practitioners have important questions about how to use them. It is essential to find answers about what the new search technologies are, how they work, and how to defend their use in court. Part III of this article provides those answers.

III. EXAMINATION OF SEARCH TECHNOLOGIES IN E-DISCOVERY

[26]   Part II considered the current climate of search technologies in e-discovery and an attorney's duty to understand those technologies. A legal duty to understand technology is not without precedent. Court opinions and commentary lead to the conclusion that some e-discovery

---

[82] *Id.* (noting that Judge Peck "issued an order staying MSL's discovery of electronically stored information until there's a ruling on whether the case can be certified as a collective action").

[83] *See* Andrew Peck, *Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?*, LAW TECHNOLOGY NEWS (Oct. 1, 2011) (stating that more attorneys are using advanced search technology as the technology and methods improve).

[84] *See Mythbusters: Exploding House*, Episode 23 (Discovery television broadcast Nov. 16, 2004) (showing that a needle can literally be found in a haystack, but only by using a specialized machine or process).

[85] Paul & Baron, *supra* note 3, at ¶ 36.

technologies, like keyword search, may be insufficient and therefore attorneys should educate themselves about alternate search technologies.

[27]   Given the scarcity of information regarding advanced search technology and how it operates, Part III begins by providing a lay-lawyer description of conceptual search technologies and how they are employed in e-discovery. It analyzes recent research proving that advanced search technologies lead to a more complete, accurate, and cost-efficient e-discovery process. Furthermore, it provides practical guidance to ensure that an attorney's use of conceptual search technologies is defensible in court.

A. An E-Discovery Search Vocabulary

[28]   The purpose of this article is not to provide an in-depth technical examination of search methodologies or to advocate the use of a particular e-discovery vendor or product. Its purpose is merely to present in ordinary language current search methodologies that are now available and that may help counsel and clients throughout the American justice system to better coordinate e-discovery efforts. This paper accomplishes this task by discussing advanced search technologies in lay-lawyer terms that any member of the bar practicing in the twenty-first century should be capable of understanding. The rationale behind a lay-lawyer explanation of advanced search technologies is to use the vocabulary framework that has been developed through commentary in the Sedona Conference and other articles[86] that technical consultants will also understand[87] and upon

---

[86] *E.g.*, *About The Sedona Conference*, *supra* note 32; *see, e.g.*, *The Sedona Conference Glossary: Commonly Used Terms for E-Discovery and Digital Information Management (3d ed.)*, THE SEDONA CONFERENCE (Oct. 2010), https://thesedonaconference.org/download-pub/471.

[87] *See* Jonathan Jaffe, Comment to *Hash*, E-DISCOVERY TEAM (Dec. 2, 2009, 9:03 AM), http://e-discoveryteam.com/computer-hash-5f0266c4c326b9a1ef9e39cb78c352dc/ (describing a language inconsistency between the legal and technology worlds manifested in the actual blog post's discussion regarding hashing algorithms). In order for attorneys

which further commentators can build and provide new insight as technology advances.[88] To this aim, the following technologies are defined and some of their uses are outlined to a limited extent to help readers understand and apply, or at least defend themselves when discussing, modern e-discovery search methodologies.

[29]   To begin, it is also important to recognize that this is more than a theoretical discussion. Courts and commentators have at times referred to the following technologies using vague terms such as "emerging" search methods.[89] However, since it is clear that the technologies exist, they have officially "emerged."[90] By clearly identifying these search methods, it should help practitioners overcome any fear of dealing with Technologies-That-Must-Not-Be-Named.[91] As a group, the technologies should perhaps be acknowledged as "advanced" search technologies or often as "concept" or "conceptual" searches, though never referred to as "new" or "emerging."[92] No one should suggest that the technology is unavailable, untried, or not yet suited to the e-discovery task.

[30]   Furthermore, given the rapidly evolving state of technology, this should not be considered a comprehensive list requiring no further

---

and technology consultants to work together in a multidisciplinary field like e-discovery, they must both learn to speak the same language.

[88] Zubulake v. UBS Warburg LLC, 229 F.R.D. 422, 440 (2004) ("The subject of the discovery of electronically stored information is rapidly evolving.").

[89] *See, e.g.*, Paul & Baron, *supra* note 3, at ¶ 43.

[90] *Best Practices*, *supra* note 17, at 204.

[91] *Cf.* THE HARRY POTTER LEXICON, http://www.hp-lexicon.org/wizards/voldemort.html (last visited Nov. 22, 2011) (discussing the added fear inherent in not naming Voldemort, or, He-Who-Must-Not-Be-Named).

[92] *See* Paul & Baron, *supra* note 3, at ¶ 43.

learning on the part of the practitioner or judge.[93] Rather, the explanations that follow should be considered a starting point, allowing lawyers to quickly gain a basic understanding of some of the overarching search technologies and concepts currently in use in e-discovery practice.

1. Keyword Search

[31]   Most practicing attorneys are already familiar with keyword searches due to their experience with popular legal research services like Westlaw and Lexis, as well as society's general experience with web search engines like Google and Yahoo!.[94] Given the legal industry's general familiarity with keyword and Boolean search technology, more time will be given to explaining the more advanced conceptual search technologies. Of course, keyword searches will continue to play a part in e-discovery. The simplicity in its use makes it possible to immediately sift through a data set and gain some general ideas about the use of certain keywords.[95]

[32]   However, the main problem with keyword search is the very simplicity that has given it widespread use.[96] In its most basic form, keyword search can only find documents with the exact keyword searched for.[97] This means that potentially relevant documents that do not contain any of the keywords searched for will not be found, notwithstanding the

---

[93] *Best Practices*, *supra* note 17, at 217.

[94] *See* Paul & Baron, *supra* note 3, at ¶ 37.

[95] *See id.*

[96] *See id.* at ¶¶ 4, 6.

[97] *Id.* at ¶ 37 n. 92 (citing J.C. Smith, *Machine Intelligence and Legal Reasoning*, 73 CHI.-KENT L. REV. 277, 334-35 (1998)).

expertise in choosing the keywords.[98] At the same time, the use of a keyword in a document does not guarantee relevance.[99]

[33]   Two variants on keyword search attempt to overcome this problem: Boolean operators and fuzzy search technologies.

a. Boolean Operators

[34]   Boolean operators may help resolve the shortcomings of keyword search to some degree, thereby allowing a user to request documents with multiple keywords, find specific phrases, or even find keywords within a specified proximity to each other.[100] One advanced Boolean tactic allows the use of wildcard operators, a practice known as truncation or stemming, to find keywords that use the same word root.[101] For example, a search for "read*" would find documents containing the words "reads," "reader," and "reading."[102] In this way, Boolean operators extend a keyword search, making it more likely to find relevant documents by combining keywords.[103] However, as an extension of keyword, it suffers from the same weakness: it is still guesswork.[104]

---

[98] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 253 F.R.D. 251, 257 (D. Md. 2008).

[99] *Id.*

[100] *Best Practices*, *supra* note 17, at 217.

[101] *See id.* at 218.

[102] *See id.*

[103] *See id.*

[104] *See* Paul & Baron, *supra* note 3, at ¶¶ 37-40.

b. Fuzzy Search

[35]   As another attempt at overcoming keyword's simplicity, fuzzy search assists parties in finding *misspelled* keywords.[105] As mentioned, keyword search is strict in the sense that it can only find documents with the exact keywords used, misspellings notwithstanding.[106] Fuzzy search overcomes the occasional typo by giving more weight to words whose middle letters match since English usage tends to have more word variants or misspellings at the beginning and end of words.[107] This would, for example, recall alternate spellings like "theatre" and "theater," as well as find words that are simply mistyped.[108] Despite fuzzy search's utility, it only presents part of the picture in overcoming limitations associated with keyword search.[109]

[36]   Even with the help of Boolean and fuzzy technology, keyword search is still guesswork.[110] Parties may make educated guesses about which keywords may have been used in a universe of documents, but the problem remains that keyword can only find documents with the keyword searched.[111] Unfamiliarity with a case and industry specific language or slang used by the key parties may lead to the inability to form an adequate

---

[105] *See Best Practices*, *supra* note 17, at 219.

[106] Susan W. Brenner & Barbara A. Frederiksen, *Computer Searches and Seizures: Some Unresolved Issues*, 8 MICH. TELECOMM. & TECH. L. REV. 39, 60-61 (2002).

[107] *Best Practices*, *supra* note 94, at 219.

[108] *Id.*

[109] *Id.* at 202.

[110] Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *10 (S.D.N.Y. Feb. 24, 2012).

[111] *Id.*

search, which in turn leads to disappointing search results.[112] In effect, keyword search is an attempt at divination; it is a gamble that hopes to find a majority of relevant documents based on informed guesswork about an industry in a particular set of documents.[113] This is why a study has shown that keyword searching reveals only one in five relevant documents.[114] The *Moore* decision previously discussed lamented this limitation of keyword search technology and cited this weakness in justifying the use of advanced search and review technology.[115] Given the lack of crystal balls in e-discovery, attorneys must instead turn to advanced conceptual searches.

2. Conceptual Search Technologies

[37]   Conceptual search technologies overcome the weakness of keyword search by recalling more than just documents containing the exact words in the search query.   Instead, conceptual searches find documents based on their relevance or similarity to the *ideas* expressed in the search query.[116]   These advanced technologies can take words, phrases, or even a "training set" of documents as an input query, as the parties did in the *Moore* decision,[117] and then recall material that is *conceptually* related to the search query. A very basic overview of how these technologies find conceptually related material is provided below.

---

[112] *See* Paul & Baron, *supra* note 3, at ¶ 38.

[113] *Cf. id.* at ¶ 39 (stating that searching via keyword is "fraught with technological difficulties").

[114] *See id.* at ¶ 40 (citing a study where a keyword search only revealed 20% of the relevant documents in the litigation).

[115] *Moore*, 2012 WL 607412, at *10-12.

[116] *See Best Practices*, *supra* note 17, at 202.

[117] *Moore*, 2012 WL 607412, at *5.

Again, the following is not an exhaustive discussion of available concept search and review technologies, but is provided to give practicing attorneys a general idea of the kinds of technologies that are available and a quick view of how they work.

a. Ontology and Taxonomy

[38]   Perhaps the simplest way to think of taxonomy and ontology search technologies is to consider them from the perspective of a thesaurus.[118] Again, the problem with keyword search is that if the exact word searched for is not present in the document, the document is not included in the realm of potentially relevant documents.[119] Taxonomy and ontology tools overcome this problem by automatically searching for synonyms of keywords.[120] However, taxonomy is more than just finding synonyms; it is finding relationships, which is the science of classification.[121] For example, a search for "shoes" using taxonomy technology might find, boots, slippers, loafers, heels, and many other variations. Ontologies tend to be more generic, leading away from mere shoe types to other topics that are related to shoes.[122] For example, a search for "shoes" using ontology technology might find podiatrists or shoe manufacturers.

[39]   A taxonomy is generally represented in graphic form as a tree with a root word and branches to other related words.[123] As provided by

---

[118] *See Best Practices*, *supra* note 17, at 221.

[119] *See* Paul & Baron, *supra* note 3, at ¶ 37.

[120] *Best Practices*, *supra* note 17, at 221.

[121] *Id.*

[122] See *id.* at 222.

[123] *See, e.g., id.* at 221.

example in this article's appendix, another conceptual way to consider taxonomy and word relationships is to imagine a web of interrelating words.[124]

[40]   Viewing and analyzing words in the context of the relationships between them can also be helpful in the e-discovery context.[125] Not only is it important to find relevant material that uses words synonymous to the main keywords selected, but determining the documents' relationship to each other helps attorneys to determine where it will be most useful to concentrate one's e-discovery efforts.[126]

b. Document Clustering

[41]   Clustering tools use statistical methods to automatically group documents with similar content.[127] Similarity of content can be defined a number of ways, but a typical way is to automatically group documents by the number of words that overlap from one document to another.[128] The more words that a document has in common with another document, the greater the likelihood that the documents are related.[129]

---

[124] *Id.* at 221-22. An online tool—www.visualthesaurus.com—may be helpful in visualizing what a taxonomy looks like. A screenshot of a taxonomy example taken from an online visual thesaurus is provided in the Appendix. THINKMAP VISUAL THESAURUS, http://www.visualthesaurus.com (last visited Sept. 5, 2012).

[125] *See Best Practices*, *supra* note 17, at 222.

[126] *See id.*

[127] *Id.* at 219.

[128] *Id.*

[129] *Id.*

[42]   There are a number of parameters that users can control when using a document-clustering tool.[130] For example, a fixed number of possible clusters can be set and topics for the clusters can even be identified.[131] One effective way of guiding the clustering process is to choose certain documents, analyze them manually, and arrange them as "seed" documents.[132] Subsequently, when the clustering engine is run, it will base its document clusters off those seed documents and parameters placed by the user.[133]

[43]   In e-discovery, document clustering can provide a quick snapshot of the data and how all of the documents are related.[134] Many e-discovery vendors boast early case assessment technologies ("ECA").[135] It is likely that some of these ECA tools include some form of document clustering capability to group similar documents.[136] Document clustering could provide insight into a case by identifying additional key players, creating estimates of the potential number of documents that may eventually need to be produced, and laying the groundwork for deciding which keywords might be important in further identifying relevant documents. [137] Combined with powerful visualization tools that showcase data on graphs

---

[130] *Best Practices*, *supra* note 17, at 219.

[131] *Id.*

[132] *Id.*

[133] *See id.*

[134] *See id.*

[135] *See, e.g.*, *Early Case Assessment*, CLEARWELL SYSTEMS, http://www.clearwellsystems.com/e-discovery-customers/early-case-assessment.php (last visited Sept. 7, 2012).

[136] *Cf. Best Practices*, *supra* note 17, at 219.

[137] *See id.* at 203.

and charts that are more easily read than a host of documents, counsel can establish the merits of a given claim and make educated decisions about a case from the very beginning instead of waiting until the end of a long and expensive manual review process.[138]

c. Bayesian Classifiers

[44]   In contrast to *statistics-based* clustering tools that look at the number of common words between documents, Bayesian technologies are based on *probability* algorithms that determine the likelihood that a document is relevant by placing a value on words, their relationships to each other, and their proximity and frequency in comparison with other documents. [139]   In clustering tools, all overlapping words between documents may hold the same value.[140] While that method may be useful to provide a quick comparison, Bayesian systems go the extra mile by setting up a formula that weighs and ranks words and their relationships.[141] One can customize how a Bayesian system ranks words and documents per implementation. However, Bayesian systems typically weigh factors, such as the frequency of certain words in the document, the location of keywords in the document, and the proximity of certain words to other important words.[142] Bayesian systems are also informed by feedback on the relevance of documents and therefore learn during the review process.[143] Before a Bayesian system is even implemented, a set of

---

[138] *See id.* at 222.

[139] *Id.* at 218.

[140] *See id.* at 219.

[141] *Best Practices*, *supra* note 17, at 218.

[142] *Id.*

[143] *See id.* at 218-19.

documents are typically reviewed in order to "train" the system to identify which kinds of documents are relevant or irrelevant.[144]

[45]   A complete explanation of Bayesian technology is well outside the scope of this paper. However, Bayesian technology's application to e-discovery can be informed from other disciplines. To provide two examples, Bayesian technology has been employed in e-mail spam filtering[145] and facial recognition software.[146]

[46]   Bayesian technology has been used to filter spam e-mails since the late 1990s.[147] A Bayesian spam filter has one job, which is to determine whether a message is junk.[148] The filter works by comparing new e-mail messages with current messages that have already been organized into junk and non-junk folders.[149] For example, when a new message is received, the spam filter will automatically compare the words in the recent message against the messages in the junk folder.[150] It will compare the frequency of certain words like "Nigerian Prince" and "wire transfer."[151] The filter might compare the proximity of those keywords to

---

[144] *See id.* at 218.

[145] *See generally* Mehran Sahami et al., *A Bayesian Approach to Filtering Junk E-Mail*, *in* LEARNING FOR TEXT CATEGORIZATION 55 (1998) *available at* http://robotics.stanford.edu /users/sahami/papers-dir/spam.pdf.

[146] *See generally* Baback Moghaddam et al., *Bayesian Face Recognition*, 33 PATTERN RECOGNITION 1771 (2000).

[147] *See, e.g.*, Sahami, *supra* note 145, at 56.

[148] *See id.*

[149] *See, e.g., id.*

[150] *See, e.g., id*.

[151] *See id.*

each other.[152] It might also consider where the junk-implicating words are located, whether in the subject line, or the body of the e-mail.[153] Further parameters can also be programmed, such as whether the user has previously received a message from the sender of the e-mail.[154] Ultimately, any e-mail containing the words "Nigerian Prince," "wire transfer," and "bank routing number" should end up in the junk folder.[155]

[47]   The utility of Apple iPhoto's "Faces" capability also demonstrates how a Bayesian search and review process might work.[156] In iPhoto, a user can categorize photos by face.[157] To streamline the process, iPhoto allows a user to identify the faces in a photo.[158] After a face has been identified, iPhoto searches through all the photos in the application for a face with matching characteristics.[159] When it begins, iPhoto may draw a large number of false positives. However, after a number of iterations, iPhoto "learns" which photos match or do not match the first "training set"

---

[152] *See* Sahami, *supra* note 145.

[153] *Id.*

[154] *See id.*

[155] *See id.*

[156] *See* Wilson Rothman, *What to Know About iPhoto '09 Face Detection and Recognition*, GIZMODO (Jan. 29, 2009, 8:00 AM), http://gizmodo.com/5141741/what-to-know-about-iphoto-09-face-detection-and-recognition. The author is unaware whether iPhoto uses Bayesian classifiers as part of its iPhoto facial recognition software; however, Bayesian technology has been employed in facial recognition software and iPhoto provides a popular example of technology that is at least similar to how a Bayesian search and review system might work.

[157] *Id.*

[158] *Id.*

[159] *Id.*

of faces identified.[160] Eventually, the program does so with a high level of accuracy.[161]

[48]   In a similar fashion, Bayesian technology in the e-discovery context allows a user to begin by identifying certain documents as relevant, irrelevant, privileged, or not privileged.[162] Instead of having these decisions made by an army of low-level legal associates or contract attorneys, sensitive relevance determinations can be made by senior attorneys familiar with the case in order to produce a high-quality "training set" of documents.[163] This "training set" of documents is then used to search through the universe of documents and ask the user if the next set of documents is relevant to the litigation and whether or not it is privileged.[164] Over time, the computer can learn which documents are relevant with a high level of accuracy.[165] This technology allows attorneys to review wide swaths of documents in short periods of time for both relevance and privilege.[166]

---

[160] *Id.*

[161] Rothman, *supra* note 156.

[162] *See Best Practices*, *supra* note 17, at 218.

[163] This appears to be the approach in the now influential *Moore* decision. Arguably, by having senior attorneys carefully review a smaller "seed" or "training" set of documents, the overall document review process is honed and more attuned to the issues being litigated, leading to a more complete, accurate, and efficient review. *See* Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *5 (S.D.N.Y. Feb. 24, 2012).

[164] *See id.* at *5-6.

[165] *See id.*

[166] *See, e.g., id.*

3. **Maintaining Quality and the Role of Sampling**

[49]   Regardless of the search process used, courts may expect attorneys to use safeguards to help ensure the quality of the document review.[167] Sampling is a quality control method urged by courts that consists of manually sampling files identified as relevant or privileged to test whether or not the review process was accurate.[168]   However, that explanation might be overly simplistic since courts have also stated that expert assistance may be required to develop an effective sampling protocol.[169] Sampling can serve as a check on advanced search methodologies, thereby helping technology-skeptic attorneys to rest easy by ensuring that machine-assisted search and review results are as accurate and complete as possible.[170]   Wise practitioners can leverage sampling techniques to improve overall search and review, informing their process through sampling relevant documents that were not identified, and modifying search processes to increase accuracy.[171]   This sampling process, used at various phases of the search and review, can then be used to explain the efficacy of the search tools used.[172]

---

[167] ACHIEVING QUALITY, *supra* note 40, at 11.

[168] *Id.* at 9.

[169] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260 n.10 (D. Md. 2008).

[170] *See* ACHIEVING QUALITY, *supra* note 40, at 11.

[171] *See id.*

[172] *See id.* at 11-12.

B. Do Conceptual Search and Review Technologies Work Better than Keyword Search and Manual Review?

[50]   In Part III Section A, we discussed the importance of a common language to discuss e-discovery technologies and reviewed current conceptual search technologies using lay-lawyer terms. The examples provided were intended to give practitioners insight into the complexity of conceptual search and a framework for understanding its use. Part III Section B goes one step further by answering a question every astute reader should be asking: do the technologies work better than the status quo?

[51]   Understanding the technologies and putting them into practice is not enough.[173] The true measure of potentially helpful conceptual search technologies is whether they actually do a better job than the traditional keyword search followed by a manual review of every single responsive document.[174] Until recently, a comparative test and analysis of the two methods had been lacking.[175] However, this was changed in 2009 when the Text Retrieval Conference ("TREC") conducted a study comparing traditional search and review methods to advanced search technologies.[176]

1. Factors for TREC Analysis

[52]   The analysis focused on three key indicators to determine whether the groups using conceptual search technologies actually performed better during the e-discovery process.[177] The first important factor was recall,

---

[173] *See* Boehning & Toal, *supra* note 7, at 2.

[174] *See* Grossman & Cormack, *supra* note 10, at ¶ 6.

[175] *See id.* at ¶ 27.

[176] *See id.* at ¶¶ 44-46.

[177] *See id.* at ¶ 34.

which is the percentage of relevant documents a group finds out of the total number of relevant documents in a data set.[178] Thus, if there are 1,000 relevant documents in a universe of 10,000, and an e-discovery process finds 200 of the relevant documents, then its recall is 20% because it only found 200 of the possible 1,000 documents.[179] As will be discussed later, 20% recall is about par for the course with keyword search alone.[180]

[53]   The second factor, precision, measures how well the process retrieved *only* the relevant documents.[181] Using the same example of 10,000 documents with 1,000 relevant documents, assume an e-discovery process identified 400 documents, but only 200 of those documents were

---

[178] *See id.* at ¶ 7.

[179] The numbers used here are merely provided as an example to help explain the concepts of recall and precision. When determining the total number of potentially relevant documents in the TREC study, a series of mathematical formulas was applied to the data resulting from the various reviews. Four calculations were applied to each group's review to determine: (1) the proportion of relevant documents within the group of documents reviewed; (2) the number of relevant documents within the group of documents reviewed; (3) the estimate of variance within the produced documents; and (4) the estimate of variance within all the documents reviewed by the given groups. Because of time and resource constraints, the groups were only able to review portions of the full document collection available for the study. Further estimates of the total number of relevant documents were determined as well as the variance calculation on the full document collection. Using information from the review itself, and applying the formulas mentioned above, the TREC study was able to determine a probability estimate of the total number of relevant documents in each sample tested in addition to the full document collection. More information regarding estimating the denominator of potentially relevant documents, including an in-depth analysis and the specific formulas used, can be found online in the TREC 2008 report. *See* DOUGLAS W. OARD ET AL., OVERVIEW OF THE TREC 2008 LEGAL TRACK: ESTIMATION OF METRICS—INTERACTIVE TASK, at 21-23, 40-44 (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf.

[180] *See Best Practices*, *supra* note 17, at 206.

[181] *See* Grossman & Cormack, *supra* note 10, at ¶ 7.

relevant while the remaining 200 documents were irrelevant to the litigation. The resulting precision calculation would be 50%.[182]

[54]   The third factor utilized in the study to determine the quality of an e-discovery process is entitled F_1, which is derivative of the first and second factors of recall and precision.[183] Using the same example with a 20% recall rate and a 50% precision rate, the resulting F_1, or harmonic mean,[184] would fall somewhere in between the two numbers at about 28.57%.[185] This third factor is the most important as it measures a balance between the two important factors involved in determining the quality of the document search and review process.[186] The higher the F_1, the more complete and more accurate the review process is.[187]

2. Advanced Technology Used

[55]   With regards to the actual search and review technologies employed in the TREC study, half of the groups used a manual review process and the other half used custom search technologies developed by the parties themselves.[188] Of the advanced technology groups, one group described their technology as "deterministic," beginning the review by tailoring a highly detailed definition of relevance.[189] Then, documents

---

[182] *See id.*

[183] *See id.* at ¶ 9.

[184] *See id.* at ¶ 9 n.30.

[185] *See id.*

[186] *See* Grossman & Cormack, *supra* note 10, at ¶ 9.

[187] *See id*.

[188] *See id.* at ¶ 37 tbl.7.

[189] *Id.* at ¶ 39.

could easily be compared against the relevance parameters to determine if it was responsive.[190] The intent was to bring a high level of precision to the review process, rejecting the practice of using broad keyword searches and later narrowing down the data set.[191]

[56]   Another advanced technology group used a computer assisted learning approach that estimated the probability that a document was relevant.[192] The system used had previously been developed to assist in spam filtering.[193] As the technology "learned" from new documents, so did the reviewers, adjusting the search and judging system to improve the review throughout the process.[194]

[57]   In the real world, e-discovery vendors may not describe their systems as "conceptual search." However, the vocabulary framework of Part III Section A should help attorneys identify and understand how a given technology might work. Even though a "training set" of documents may not have been used, detailed relevance parameters and computer learning systems helped parties in the TREC study identify and group responsive documents.[195] These strategies seem similar to a Bayesian classification system.[196]

---

[190] *See id.* at ¶¶ 38-39.

[191] *See* Grossman & Cormack, *supra* note 10, at ¶¶ 38-39.

[192] *See id.* at ¶¶ 40-41.

[193] *See id.*

[194] *See id.*

[195] *See Id.* at ¶¶ 38-41.

[196] *See* discussion *supra* Part III.A.

3. Study Findings

[58]   Each group involved in the 2009 TREC study was assigned a topic and requested to sift through the data provided to build a "case" as if for litigation.[197] The results of the varying review processes, manual review versus technology assisted review, were compared and analyzed. [198] Across the wide majority of the topics tested, the groups using advanced search technologies performed at a statistically significant rate higher than the groups who used traditional review methods.[199] The average recall and precision for the traditional review groups was 59.3% and 31.7% respectively, while the recall and precision for the concept search and review was 76.7% and 84.7% respectively.[200] On average, the F_1 for the traditional review groups was 36%.[201] Among those who used advanced search technologies, the F_1 was 80%.[202]

[59]   Clearly, the advantages of conceptual search technologies can be understood on a superficial level. After discussing the available search technologies and their possible uses in the e-discovery context, a number of strategies can be imagined to automatically organize documents in a data set, see relationships among the information, search more accurately and widely to find the relevant documents, and then use automated learning tools to speed up the review process, all more accurately than

---

[197] *See* Grossman & Cormack, *supra* note 10, at ¶¶ 30-32.

[198] *Id.* at ¶ 45.

[199] *See id.* at ¶ 45 tbl.7.

[200] *Id.*

[201] *Id.*

[202] *See* Grossman & Cormack, *supra* note 10, at ¶ 45 tbl.7.

with manual review.[203] This is not the future of search technology; this is now.[204]

C. Defending E-Discovery Process Through Conceptual Search Technologies

[60] Even though a practitioner may now be aware of conceptual search technologies after reading Part III Section A, and understand that conceptual search and review produces better results after reviewing Part III Section B, opposing counsel and courts may still need some convincing. Part III Section C discusses how to defend conceptual search in court. Regardless of the e-discovery search methodology employed, whether keyword or conceptual, the parties must be able to defend the methodology used before a judge.[205] While using a defensible process throughout the entirety of a discovery request is beyond the scope of this paper, this discussion would be incomplete without reviewing aspects of defensibility applicable to advanced search methods.

1. Accuracy

[61] First, a defensible search methodology is not a perfect search methodology.[206] In the Sedona Conference Best Practices Commentary on the Use of Search, the authors discuss a 1985 study by David Blair and

---

[203] *See* discussion *supra* Part II.A.

[204] *See* Grossman & Cormack, *supra* note 10, at ¶ 45 (proving that a technology assisted review process using advanced search and review methods is more effective than manual review).

[205] *See Best Practices*, *supra* note 17, at 214.

[206] *See* Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *11 (S.D.N.Y. Feb. 24, 2012) ("While this Court recognizes that computer-assisted review is not perfect, the Federal Rules of Civil Procedure do not require perfection."); *see also Best Practices*, *supra* note 94, at 206.

M.E. Maron.[207] The case dealt with an unfortunate train accident in the San Francisco area that resulted in an e-discovery workload of 40,000 documents and some 350,000 pages.[208] After a thorough review of the documents, presumably based on some form of keyword search to identify potentially relevant documents, attorneys in the case estimated they had found seventy-five percent of all relevant documents.[209] However, a detailed analysis of the documents involved revealed that attorneys on the case had only identified about twenty percent of the relevant documents.[210] The article attributes this lack of accuracy to the ambiguity inherent in word usage, giving more weight to the idea that the assistance of search experts may become necessary, as Judge Grimm implied in *Victor Stanley*.[211]

[62]   Regardless, the key point is that even though keyword and Boolean search has been the "state-of-the-art" in terms of e-discovery search for many years, keyword search has never led to perfection in the e-discovery process.[212] It follows, then, that any search performed using conceptual searches must merely meet the low threshold of a keyword search process.[213] As previously discussed in Part III Section B, technology assisted search and review has proven to be *more* accurate.

---

[207] *See Best Practices*, *supra* note 17, at 206.

[208] *See id.*

[209] *See* Grossman & Cormack, *supra* note 10, at ¶ 45.

[210] *Best Practices*, *supra* note 17, at 206.

[211] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260 (D. Md. 2008) ("[The] proper selection and implementation [of keywords] obviously involves technical, if not scientific knowledge.").

[212] *See id.*

[213] *Cf.* Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412, at *10-11 (S.D.N.Y. Feb. 24, 2012) (pointing out the low accuracy threshold of keyword searches).

2. Efficiency

[63]   With regards to efficiency, the more quickly the universe of documents can be culled and reviewed, the better.[214] However, efficiency and accuracy do not exist on a sliding scale where accuracy can be sacrificed. Quickly reviewing a universe of 350,000 documents without finding a single responsive file would not be defensible.[215] Certain standards of accuracy must be met while using efficient and cost-effective means at the same time.[216] Some practitioners are experiencing success with conceptual search, thereby providing a much quicker review period with appropriate levels of accuracy.[217] One report stated that a body of 20,933 documents reviewed first using traditional review methods took 180 hours to review.[218] Afterward, the same documents were loaded into a system that learned as a separate review progressed and grouped documents according to topic.[219] This second review took 18.5 hours, nearly one-tenth of the manual review time,[220] a speedy determination indeed.[221]

---

[214] *See* ACHIEVING QUALITY, *supra* note 40, at 5.

[215] *See id.* at 1-3.

[216] *See id.*

[217] *See generally* Bennett B. Borden et al., *Why Document Review is Broken*, THE WILLIAMS MULLEN EDGE (May 2011), http://www.umiacs.umd.edu/~oard/desi4/ papers/borden.pdf.

[218] *See id.* at 2 (explaining the amount of time taken to complete a review of 20,933 documents using the traditional method).

[219] *See id.*

[220] *See id.* (explaining the amount of time taken to complete a review of 20,933 documents was ten times faster using linear review than the traditional method).

[221] *See* FED. R. CIV. P. 1 ("They shall be construed and administered to secure the just, speedy, and inexpensive determination of every action").

3. Transparency

[64]   This aspect of defending the use of advanced search technology, transparency, goes not to the efficacy of the technology itself, but the cooperation of the parties with regard to its use.[222] The *Moore* Court went so far as to state that the defendant's willingness to be transparent in their implementation of advanced technologies made it possible for the court to approve the use of the technology-assisted review process.[223] In *Moore*, the defendant agreed to provide plaintiffs with a complete copy of all "seed" documents they had reviewed, except for privileged documents, which they then used to "train" the computer in the review process.[224] By providing the 2,399 documents in the "seed set," plaintiffs and the court would be able to plainly evaluate and provide guidance in setting the parameters of the advanced review. [225]   Arguably, this level of transparency also gives unprecedented power to plaintiffs who can effectively provide input to the decisions in the defendant's review process.[226] While this level of transparency may not be required in every situation using advanced technologies, it may help opposing counsel and the court to feel more at ease with technologies that are admittedly difficult to understand.[227]

---

[222] *See* Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412 at *11 (S.D.N.Y. Feb. 24, 2012) ("Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI.").

[223] *See id.*

[224] *See id.* at *5.

[225] *See id.* at *3.

[226] *See, e.g.*, *id.* at *5.

[227] *See Moore*, 2012 WL 607412, at *11.

4. Other Factors of Defensibility

[65]   The Sedona Commentary also mentions defensibility guidelines, such as cost effectiveness and a showing of fairness and good faith.[228] While no one factor seems to predominate, "the just, speedy, and inexpensive determination of every action and proceeding" appears to be the underlying factor of defensible e-discovery.[229] Counsel should be prepared to articulate how the search methods employed helped meet the ends of the Federal Rules of Civil Procedure.[230] Additionally, evidence regarding the efficacy of a search methodology must be introduced through experts.[231] Attorneys and experts should be prepared to explain that a well-implemented conceptual search speeds up the review process and leads to more accurate results.[232] Saved costs are the logical byproduct and should be included as part of any defense concerning the efficacy of advanced search technologies.[233] In the end, attorneys defend their use of advanced conceptual search by demonstrating that it is more just, speedier, and less expensive than keyword search followed by manual review.

---

[228] *See Best Practices*, *supra* note 17, at 195.

[229] FED. R. CIV. P. 1.

[230] *See supra* Part II.A-B.

[231] *See* United States v. O'Keefe, 537 F. Supp. 2d. 14, 24 (D.D.C. 2008) ("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 [regarding the introduction of evidence via experts].").

[232] *See* discussion *supra* Part II.B.

[233] *See* Borden, *supra* note 217, at 4.

IV. REQUIRED USE OF ADVANCED SEARCH TECHNOLOGIES

[66]   In Part III, we defined some of the advanced search technologies currently employed in e-discovery, determined that they are indeed more accurate than keyword search alone followed by manual review, and also considered how to defend and explain their use for courts and opposing counsel. Part IV provides guidance on when the use of advanced search technologies should be required. It also discusses how courts should deal with inadvertent production of privileged documents after using an advanced e-discovery process. In addition, Part IV provides guidance to practitioners regarding what technology-related legal duties have formed around the e-discovery process and how those duties help fulfill the purposes of the Federal Rules of Civil Procedure.

A. Requiring Conceptual Search

[67]   Cases that should require the use of advanced search technologies involve millions of documents.[234] These cases involve situations where a keyword search followed by manual review would be truly unfeasible and overly expensive.[235] Knowing that one of the underlying purposes behind the duty to use conceptual search technologies is to save money, and recognizing that the biggest expense in e-discovery is manual review, an understanding of advanced search technologies should help courts and counsel draw the conclusion that an advanced process saves both time and money.[236] Although some pushback from counsel is to be expected, court enforcement of a duty to use advanced search technologies, accompanied with further research and learning about conceptual search technology, should help allay any concerns about the efficacy of conceptual search.[237]

---

[234] *See Best Practices*, *supra* note 17, at 194.

[235] *See id.*

[236] *See* ACHIEVING QUALITY, *supra* note 40, at 309.

[237] *See* Boehning & Toal, *supra* note 7, at 2; *see also* discussion *infra* Part III.B.

[68]   Should the use of concept search technologies be required in all cases? No. Given the discussion of the technologies above, there is clearly some level of preparation and analysis required before conceptual search and review is initiated.[238] In some cases, it may be more effective to formulate a basic keyword strategy, especially when dealing with smaller data sets where manual review is feasible and less expensive than employing the services of an e-discovery vendor.[239] Using advanced technology in those situations would be akin to cutting the Thanksgiving turkey with a chainsaw—simply overkill.

B. Dealing with Privileged Documents

[69]   When courts evaluate discovery productions that result from a conceptual search and review process, they should keep in mind the impossibility of manually reviewing millions of documents.[240] Given that *review* of documents does not necessarily equal *viewing* documents, courts should respect clawback agreements between parties and be hesitant to find waiver of privilege from documents that were inadvertently produced. Despite an attorney's best efforts, it is possible and even likely that after a review of millions of documents, some privileged material will be produced to opposing counsel.[241] Courts and counsel, in the interest of the speedy and just determination of a case, should expect a certain level of

---

[238] *See Best Practices*, *supra* note 17, at 194 ("[A]ny automated search method or technology will be enhanced by a well-thought out process with substantial human input on the front end.").

[239] *See id.* at 209-10. If substantial human input is required to initiate an advanced search methodology, then smaller data sets that take less time to review manually than it would to create the right search environment should not use advanced search and review methods.

[240] *See id.* at 194.

[241] *See Achieving Quality*, *supra* note 40, at 320-21.

inaccuracy involved with sifting through large quantities of documents, regardless of the search and review methodologies employed.[242]

[70]   In fact, the 2006 Amendments to the Federal Rules of Civil Procedure anticipated a margin of error when dealing with a large universe of documents, by purposefully crafting the ability to institute clawback agreements between parties.[243] Clawback agreements should be discussed as part of any electronic discovery plan and when disputes over privileged documents occur, judges should use an "appropriate mathematical yardstick" when determining whether to waive privilege.[244]

[71]   Compare, for example, the application of certain factors in the *Mt. Hawley* case with the *Victor Stanley* case.[245] In *Victor Stanley*, the court found that privilege had been waived on 165 documents out of a universe of 9,000 documents that had been inadvertently produced.[246] In contrast, the *Mt. Hawley* Court found that privilege had been waived on 377 documents out of a universe of five million documents.[247] The court's

---

[242] *See id.* at 321.

[243] *See* FED. R. CIV. P. 26(b)(5)(B); FED. R. EVID. 502 (b)(1).

[244] *See* Baron, *supra* note 5, at 40.

[245] *Compare* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 257 (D. Md. 2008) (waiving privilege on 165 documents out of a universe of 9000 documents), *with* Mt. Hawley Ins. Co. v. Felman Prod., Inc., 271 F.R.D. 125, 136, 139 (S.D. W. Va. 2010) (waiving privilege on 377 documents out of a universe of millions of documents because 377 was *double the number* of privileged documents produced in the Victor Stanley case).

[246] *See Victory Stanley*, 250 F.R.D. at 257.

[247] *See Mt. Hawley*, 271 F.R.D. at 138-39; Baron, *supra* note 5, at 40 (citing Ralph Losey, *The Good, the Bad, and the Ugly: "Mt. Hawley Ins. Co. v. Felman Production, Inc."*, E-DISCOVERY TEAM (June 10, 2010, 7:11 AM), http://e-discoveryteam.com/2010/06/10/the-good-the-bad-and-the-ugly-"mt-hawley-ins-co-v-felman-production-inc-"/).

conclusion that privilege had been waived is not necessarily the problem. However, the reasoning behind the *Mt. Hawley* holding with regard to those 377 documents is.[248] In stating that privilege had been waived on the inadvertently produced documents, the court relied in part on the *Victor Stanley* holding, concluding that 377 documents was more than double the number of documents at issue in the *Victor Stanley* case.[249] However, the *number* of documents that were inadvertently produced provides a poor comparison.[250] Using instead the number of documents in terms of a proportion or a percentage of the possible privileged documents that could have been inadvertently produced, the parties in the *Mt. Hawley* case did much better.[251] For this reason, courts should approach inadvertent disclosure problems with a relative mindset instead of thinking in terms of bright-line non-proportional rules.[252]

[72]   Ironically, the *Mt. Hawley* decision highlights much of the information that supports a conclusion that data should be evaluated on a relative basis. In its analysis, the court examines a five-factor test for determining whether privilege has been waived, the parties' own clawback agreement, the Federal Rules of Evidence and Procedure that authorize clawback agreements, and finally, the Advisory Committee Notes discussing how to evaluate clawback agreements.[253] Specifically, the Advisory Committee states that:

---

[248] *See Mt. Hawley*, 271 F.R.D. at 136, 139.

[249] *See id.*

[250] *See* Ralph Losey, *The Good, the Bad, and the Ugly: "Mt. Hawley Ins. Co. v. Felman Production, Inc.",* E-DISCOVERY TEAM® (June 10, 2010; 7:11 AM), http://e-discoveryteam.com/2010/06/10/the-good-the-bad-and-the-ugly-%E2%80%9Cmt-hawley-ins-co-v-felman-production-inc-%E2%80%9D/.

[251] *See* Baron, *supra* note 5, at ¶ 40.

[252] *See id.*

[253] *See Mt. Hawley*, 271 F.R.D. at 133-34.

> [o]ther considerations bearing on the reasonableness of a
> producing party's efforts include the *number of documents
> to be reviewed* and the *time constraints for production*.
> Depending on the circumstances, *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.[254]

Essentially, the Advisory Committee was aware of the potential need to
review huge amounts of data and that perfection in the review process
would be impossible.[255] Courts and practitioners alike should be prepared
for a margin of error in the discovery process and should be flexible
enough to work out and enforce clawback agreements that preserve
privilege while speeding along the review process.[256] Some practitioners
have begun to claim that if the advanced review is carried out properly,
privilege should not be a worry because the same systems that help
quickly and efficiently identify relevance can also make privilege
determinations with a high level of accuracy.[257] It may be that in the
future, privilege stops being a concern for parties who use advanced
search technology. Until then, courts should always consider a review
augmented by advanced conceptual searches to be found to have taken
"reasonable steps" to preserve privilege under the meaning of the Federal
Rules.[258]

---

[254] FED. R. EVID. 502 advisory committee's note (emphasis added).

[255] *Cf. Best Practices*, *supra* note 17, at 194.

[256] *See* Baron, *supra* note 5, at ¶ 40.

[257] *See* Borden, *supra* note 217, at 3.

[258] *See, e.g.*, Moore v. Publicis Groupe, No. 11 Civ. 1279(ALC)(AJP), 2012 WL 607412,
at *11 (S.D.N.Y. Feb. 24, 2012) ("[T]he Federal Rules of Civil Procedure do not require
perfection.").

C. Legal Duties

[73]   The purpose of the Federal Rules of Civil Procedure is to secure "the just, speedy, and inexpensive determination of every action and proceeding."[259] In the past, speed and expense were sacrificed in the name of justice, giving time to long-term manual review projects to ensure that the most accurate and complete set of information was discovered and produced.[260] Today, a more accurate, complete, and just process exists through conceptual search tools.[261] The fact that the same tools also give way to speedier and less expensive determination is a bonus.[262]

[74]   Judges who have recommended advanced search technologies in the past may require parties to use them in the near future, especially in litigation with large data sets.[263] It would not be the first time a judge has required counsel to learn about and become familiar with technology.[264] Legal duties in terms of the e-discovery process will continue to emerge and become more defined.[265] Just as Federal Rule of Civil Procedure 26(f) requires parties to discuss an in-depth discovery plan—including discovery subjects, production format, and privilege issues—future evolutions of the rule may require discussion of, and plans to use, advanced search technologies in order to secure the just, speedy, and inexpensive determination of the case.[266]

---

[259] FED. R. CIV. P. 1.

[260] *Cf. Best Practices*, *supra* note 17, at 192.

[261] *See* discussion *supra* Part III.B.

[262] *See* ACHIEVING QUALITY, *supra* note 40, at 1.

[263] *Cf. Best Practices*, *supra* note 17, at 194.

[264] *See* Zubulake v. UBS Warburg LLC, 229 F.R.D. 422, 432 (S.D.N.Y. 2004).

[265] *See* Baron, *supra* note 5, at ¶ 37.

[266] *See* FED. R. CIV. P. 26(f).

[75]   Requiring the attorneys in a case to use advanced search technologies may raise competency concerns.[267]   Furthermore, any mandate to use conceptual search should have at its root the purpose of helping to resolve cases on their merits rather than e-discovery issues.[268]

1. Competency

[76]   The legal community should embrace new conceptual search technologies.[269] Where expertise is lacking, attorneys should not hesitate before seeking help with managing a large database of electronic information.[270]   A defensible e-discovery strategy for large data sets should employ the review of documents through a variety of search tactics, including document clustering and keyword search assisted by Bayesian and ontology search mechanisms.[271]   Since attorneys do not typically have access to those tools on their desktop computers, *in some cases*, attorneys should be required to seek help either by a firm or a vendor who specializes in e-discovery.[272]   In the past, courts have sanctioned parties for botching e-discovery requests and requirements and as a result, the legal community should consider themselves "on notice" with regard to their competency qualifications, or lack thereof, in the e-discovery context.[273]

---

[267] *See* Baron, *supra* note 5, at ¶ 13.

[268] *Cf.* Paul & Baron, *supra* note 3, at ¶ 28.

[269] *See* ACHIEVING QUALITY, *supra* note 40, at 17.

[270] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260 n.10 (D. Md. 2008).

[271] *See Best Practices*, *supra* note 17, at 194.

[272] *See Victor Stanley, Inc.*, 250 F.R.D. at 260 n.10.

[273] *Cf.* Zubulake v. UBS Warburg LLC, 229 F.R.D. 422, 440 (S.D.N.Y. 2004).

2. Resolve Cases on their Merits

[77]   One of the best-named tools in opposing counsel's arsenal is the Weapon of Mass Discovery.[274] Counsel can sometimes try to make overbroad discovery requests, hoping for settlement from larger defendants because it would be more cost-effective for the defendant to settle than to try the case on its merits.[275] This is due in part to the impossibility of manually reviewing millions of documents.[276] The capacity of advanced search technologies to conceptually organize a universe of documents should help larger defendants avoid this threat by analyzing the merits of a claim from day one.[277] Some work and preparation for litigation will be required on the part of the defendant to effectively use this strategy, but in the long term, a strategy that includes preparation and use of conceptual search will help cases resolve on their merits instead of the difficulty of the e-discovery process.[278] This purpose should be at the core of any mandate to implement advanced search technology.

V. CONCLUSION

[78]   Courts have been developing a legal duty to understand and implement advanced search technologies in the e-discovery process.[279]

---

[274] *See* LOSEY, CURRENT TRENDS, *supra* note 27, at 41.

[275] *See id.* at 29.

[276] *See Best Practices*, *supra* note 17, at 194.

[277] *See* discussion *supra* Part II.B.

[278] *See* COOPERATION PROCLAMATION, *supra* note 33, at 331.

[279] *See* Bennett B. Borden et al., *Four Years Later: How The 2006 Amendments To The Federal Rules Have Reshaped The E-Discovery Landscape And Are Revitalizing The Civil Justice System*, 17 RICH. J.L. & TECH. 10, ¶ 36 (2011).

This duty is informed by scholarship demonstrating the efficacy of advanced search technologies and their advantage over the status quo of a keyword search method followed by an extensive manual review. To meet the needs of clients, practitioners must strive to gain some technical knowledge regarding available search and review methods. Given that manual review is the most expensive piece of the e-discovery process and that using conceptual search inevitably erases much of the manual review process along with its accompanying high cost, attorneys should implement conceptual search technologies as often as possible. The understanding that conceptual searches are more effective and efficient should help attorneys defend an advanced search process in court. Finally, as the review process is shortened considerably and the burden of review is lifted from the shoulders of counsel and courts, cases can again be resolved on their merits instead of diving down the rabbit hole of e-discovery disputes.

APPENDIX: A SAMPLE TAXONOMY

representation

affirm
sustain
support

put down

substantiate

record

corroborate

confirm

document ◀ ◀

enter

text file

papers
written document

communication