eDISCOVERY INSTITUTE

MANDATING REASONABLENESS IN A REASONABLE INQUIRY

PATRICK OOT,[†] ANNE KERSHAW,[‡] & HERBERT L. ROITBLAT[*]

PREFACE: DATA WARS

> "Don't try to frighten us with your sorcerer's ways, Lord Vader. Your sad devotion to that ancient religion has not helped you conjure up the stolen data tapes, or given you clairvoyance enough to find the rebels' hidden fortress."
>
> – Admiral Motti to Darth Vadar, *Star Wars*

Just as today's litigators struggle with searching for relevant documents to respond to an opponent's discovery request, fictional adversar-

[†] Patrick Oot is an experienced corporate attorney and co-founder of The Electronic Discovery Institute. He has extensive experience in discovery practices involved in commercial litigation, regulatory filings, employment, and antitrust matters. He has testified before the United States Judicial Conference's Advisory Committee on the Federal Rules of Evidence. Mr. Oot actively participates in The Sedona Conference's working groups. He is a member of the advisory board for American Lawyer Media's LegalTech, The Georgetown University Law Center's Advanced E-Discovery Institute, and The Council on Litigation Management. Mr. Oot lectures regularly at educational events and legal conferences internationally, has appeared on National Public Radio's Morning Edition, and was interviewed for the August 2008 edition of The Economist. Mr. Oot has appeared as an expert panelist with United States Supreme Court Justice Stephen Breyer at Georgetown University Law Center's Summit on Electronic Discovery. He received both his B.A. and J.D. from Syracuse University and his LL.M. from Georgetown University Law Center.

The authors would like to thank Judge James Francis for his gracious guidance in drafting this article. The Electronic Discovery Institute ("EDI") would like to thank Judges Facciola, Grimm, Hedges, Peck, Rosenthal, and Waxse for their assistance and contribution to EDI through interviews, decisions, and public presentations. EDI would also like to thank John Frantz, Jennifer Tomaino, Laura Kibbe, Brian Golbère, Jonathan Nystrom, Rich Tobey, Peter Conner, Caleb King, Lisa Schofield, Jean Gavin, and Larry Center for their commitment to helping EDI conduct and distribute the study discussed throughout this Article. Finally, the authors gratefully acknowledge the assistance and input of Jason R. Baron, Director of Litigation at the U.S. National Archives and Records Administration.

[‡] Anne Kershaw is a nationally recognized expert on electronic discovery and litigation data management issues, and is the founder of A. Kershaw, P.C. Attorneys & Consultants, a litigation management consulting firm (more information at www.AKershaw.com). Ms. Kershaw is also the President and co-founder of the Electronic Discovery Institute. In addition, Ms. Kershaw is a Faculty Member of Columbia University's Executive Master of Science in Technology Management Program, teaches at the Georgetown University e-Discovery Academy and is also an Advisory Board Member of the Georgetown University Law Center's Advanced E-Discovery Institute.

[*] Dr. Herb Roitblat is a recognized expert in cognitive science, information management, data mining, statistics, and eDiscovery processes. He is co-founder of The Electronic Discovery Institute and a co-founder and Principal at OrcaTec LLC. Dr. Roitblat is the author of numerous papers on dolphin biosonar and neural network models of the dolphin sensory system. More recently he has been writing about data mining and how technology can ease the burden of eDiscovery. Herb was an award-winning Professor of Psychology at the University of Hawaii from 1985 until 2002. He taught courses in cognitive science and research methods. In 2002, he received the Clifford T. Morgan award for distinguished contributions to behavioral neuroscience and comparative psychology. Herb is a member of the Sedona Working Group on Electronic Document Retention and Production.

ies that existed "[a] long time ago in a galaxy far, far away" had similar trouble locating missing data tapes on a planet-sized battleship. Searching for lost items is a difficult task—especially if your search team doesn't know what to look for.

A physical search using a traditional brute-force technique does not necessarily result in the desired outcome of locating everything requested for the least amount of effort and cost, all while avoiding the wrong look-alike objects. Knowing where to look and how to deploy the proper tools can help garner success; whereas hiring a squadron of "newbies," who lack subject matter expertise, to search every nook and cranny for a treasure that might not even exist, might be a fool's errand. After all, Darth Vader could not find the stolen data tapes because Princess Leah loaded them onto R2D2.

Similarly, American lawyers lacking jedi-like clairvoyance must develop subject matter expertise in knowing what to search for, and creating strategies and techniques to search for and produce discovery material in litigation using a cost-effective and reasonable process. Failure may result in a wrath of sanctions for underperforming, or heavy expenses for "over discovery."[1] Thus, the most successful lawyers in the electronic discovery age will be those who overcome the difficulty of reasonably responding to their opponent's requests in the shortest amount of time and for the least amount of effort and expense.

Yet litigants are at an impasse. The advent of electronically stored information ("ESI") has rendered old-fashioned, spoon-fed document review operations—which places three pairs of very expensive eyeballs on every document—impractical and arguably ineffective. Inappropriate search methodology not only increases litigation costs for the parties, but such tasks also waste the precious time of the courts with avoidable motions and unnecessary orders.

The litigation community must reconsider traditional search and retrieval techniques, or we will face either a nation without justice or a profession full of document reviewers. Traditional approaches to discovery now lead counsel away from the path toward a just, speedy, and inexpensive determination of the law, and away from the mandate that discovery responses be both reasonable and proportional to the controversy they surround.[2]

[1] "Over discovery" is the practice of collecting and producing data that is largely irrelevant, but for the fact that it may reside in the vicinity of relevant information.

[2] *See* FED. R. CIV. P. 26 Advisory Committee's notes to 1983 amendment ("These practices impose costs on an already overburdened system and impede the fundamental goal of the 'just, speedy, and inexpensive determination of every action.'" (quoting FED. R. CIV. P. 1)); *see also* Herbert v. Lando, 441 U.S. 153, 179 (1979) (Powell, J., concurring) "[T]he widespread abuse of discovery . . . has become a prime cause of delay and expense in civil litigation.").

The authors of this Article—after responding to hundreds of discovery requests by deploying massive traditional brute-force document reviews—sought peer-reviewed alternatives that are as efficient, defensible, and at least as good as traditional document review. This Article, and *The Electronic Discovery Institute Study* ("EDI study") that supports it, seek to provide insight for legal practitioners who face the significant challenge of navigating a response to an opponent's discovery requests.[3] We conclude that alternate approaches to manual human document review are both valid and reasonable, perhaps even more reasonable than traditional methodologies.

On a broader level, the EDI study and this Article underscore the tremendous value and need for the measurement and understanding of the effects of technology on litigation. As technology continues to imbed itself in the social fabric of our lives, lawyers representing clients must take on the task of learning how technology works and how it affects us, both before and after litigation is commenced. Commentators and attorneys in the so-called "e-discovery" arena often blame the problems on the technology. But we submit that it's not the technology; the problem is the attorneys that fail to learn and embrace it. Accordingly, the authors respectfully request that judges start to rule in favor of reasonableness, and that the Advisory Committee on the Federal Rules of Civil Procedure consider future supplemental commentary and committee notes that (1) encourage attorneys to learn and study technology; (2) help them better understand their options for meeting discovery obligations in litigation; and (3) assist courts with properly evaluating a disclosing party's process for meeting those obligations.[4] Quite simply, litigants require tools and guidance to help them develop discovery processes to meet the standards of Rule 1 and provide a clear path to alternatives to traditional methodologies.[5]

[3] *See generally* Herbert L. Roitblat, Anne Kershaw & Patrick Oot, *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y FOR INFO. SCI. & TECH. 70 (2010).

[4] *See* Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 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 rationale for the method chosen to the court, demonstrate that it is appropriate for the task, and show that it was properly implemented.").

[5] The Sedona Conference Working Group Series, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery,* 8 SEDONA CONF. J. 189, 198 (2007), http://www.thesedonaconference.org/dltForm?did=Best_Practices_Retrieval_Methods___revised_cover_and_preface.pdf [hereinafter Sedona Conference WGS] ("However, with increasingly complex computer networks, and the exponential increase in the volume of information existing in the digital realm, the venerated process of 'eyes only' review has become neither workable nor economically feasible.").

I. GENERAL COMMENTARY OF THE DISCOVERY PROCESS

A. Discovery Defined

For the non-lawyer, discovery is best defined as an investigation of the facts surrounding a lawsuit or case. The Federal Rules of Civil Procedure grant each party in a lawsuit the broad investigative right to retrieve information (both technical and theoretical) from the party opponent.[6] Typical methods of discovery include document requests, written interrogatories, or depositions of an opponent's representatives. During this investigation of the facts, a responding party must identify the potential locations of relevant information, collect that information, examine it for both responsiveness and privilege, and then finally produce it to the requesting party.

Similarly, The Sedona Conference® defines discovery as:

> [T]he process of identifying, locating, securing and producing information and materials for the purpose of obtaining evidence for utilization in the legal process. The term is also used to describe the process of reviewing all materials that may be potentially relevant to the issues at hand and/or that may need to be disclosed to other parties, and of evaluating evidence to prove or disprove facts, theories or allegations. There are several ways to conduct discovery, the most common of which are interrogatories, requests for production of documents and depositions.[7]

B. Discovery Prior to the Federal Rules[8]

Prior to the final draft of the Federal Rules of Civil Procedure, discovery was significantly limited in both England and the United States.[9] Broad discovery rules were not a part of litigation in the United States until Charles Clark, first reporter for the Advisory Committee on the Federal Rules of Civil Procedure, sought the assistance of scholar Edson Sunderland to draft the federal discovery provisions. Relying on George Ragland, Jr.'s extensive research on pre-trial discovery, Sunderland's additions "included every type of discovery that was known in the

[6] FED. R. CIV. P. 26 (b)(1). The Federal Rules state that "[p]arties may obtain discovery regarding any nonprivileged matter that is relevant to any party's claim or defense." *Id.* To be discoverable, the information sought "need not be admissible at the trial if the discovery appears reasonably calculated to lead to the discovery of admissible evidence." *Id.*

[7] THE SEDONA CONFERENCE WORKING GROUP SERIES, THE SEDONA CONFERENCE GLOSSARY: E-DISCOVERY AND DIGITAL INFORMATION MANAGEMENT 15 (Conor R. Crowley & Sherry B. Harris eds., 2d ed. 2007), http://www.thesedonaconference.org/dltForm?did=TSCGlossary_12_07.pdf.

[8] *See generally* INST. FOR THE ADVANCEMENT OF THE AM. LEGAL SYS., HISTORICAL BACKGROUND TO THE FEDERAL RULES OF CIVIL PROCEDURE (2009), http://www.du.edu/legalinstitute/pubs/History%20FINAL.pdf [hereinafter IAALS].

[9] *Id.* at 2 (citing Stephen N. Subrin, *Fishing Expeditions Allowed: The Historical Background of the 1938 Federal Discovery Rules*, 39 B.C. L. REV. 691, 694 (1998)).

United States and probably England up to that time."[10] As a result, the Advisory Committee included every major discovery device proposed (but for a mandatory disclosure provision) in the final draft of the Federal Rules, but excluded a number of constraining devices.[11] The Supreme Court approved the rules in December of 1937, which took effect on September 16, 1938 as a result of Congressional inaction.[12]

C. Discovery Today

Courts and treatises state that the discovery rules, together with pretrial procedures, remove the risk of surprises at trial and provide for more a fair contest "by requiring disclosure of *all* relevant information."[13] Others have argued that full discovery will, among other things, "(1) help focus controversies on the substantive issues; (2) make trials and settlements more rational; and (3) reduce pleading disputes."[14] But as practicing litigators know all too well, the nature of the adversarial system has pushed discovery well beyond its original purpose "to inform the adversary of what theories [a] party proposes to 'develop' at trial, and on what basis a jury will be asked to award damages."[15] By requiring disclosure of all possible relevant information in our electronic world, the discovery rules allow the ultimate resolution of disputed issues to be based not only on the full and accurate understanding of true facts, but also on which party has more money to spend and whether litigants are before a judge who does not understand the full cost implications of certain discovery decisions.

D. Rule 26(g): A Reasonable Inquiry

The Federal Rules of Civil Procedure mandate certain obligations of counsel when responding to discovery requests. Like all obligations, attorneys must adhere to a "standard of care." Rule 26(g) promulgates this standard of care by requiring the responding party's attorney to certify "that to the best of the person's knowledge, information, and belief formed after a *reasonable inquiry*: with respect to a disclosure, [the response] is complete and correct as of the time it is made."[16] Furthermore:

> The duty to make a "reasonable inquiry" is satisfied if the investigation undertaken by the attorney and the conclusions drawn therefrom are reasonable under the circumstances. It is an objective standard

[10] *Id.* at 6 (internal quotation marks omitted) (quoting Subrin, *supra* note 9, at 718).

[11] *Id.*

[12] *Id.*

[13] *In re* PE Corp. Sec. Litig., 221 F.R.D. 20, 23 (D. Conn. 2003) (citing United States v. Procter & Gamble Co., 356 U.S. 677, 682 (1958)); 6 JAMES WM. MOORE ET AL., MOORE'S FEDERAL PRACTICE § 26.02 (3d. ed. 1997 & Supp. 2009) (citing New Haven Temple SDA Church v. Consol. Edison Corp., No. 94 Civ. 7128 (AGS) (BAL), 1995 WL 358788 (S.D.N.Y. June 13, 1995)).

[14] IAALS, *supra* note 8, at 6 (citing Subrin, *supra* note 9, at 709).

[15] *New Haven Temple SDA Church*, 1995 WL 358788, at *5.

[16] FED. R. CIV. P. 26(g)(1) (emphasis added).

> similar to the one imposed by Rule 11. . . . Ultimately, what is reasonable is a matter for the court to decide on the totality of the circumstances.[17]

Thus, Rule 26(g) retains the principle that attorneys are obliged to be reasonable in their discovery objectives and processes, and to refrain from conduct that frustrates the objectives of Rule 1.

1. The Standard of Care for Today's Litigator: Preserve, Collect, Review, and Produce

The concept of "reasonableness" that the drafters intertwined throughout the Federal Rules of Civil Procedure stems from tort law.[18] The now-familiar objective reasonable person standard originally came from *Vaughan v. Menlove*.[19] As many lawyers will recall from first-year torts, in *Vaughn* the defendant negligently stacked hay on his property in a way that caused it to spontaneously combust and subsequently destroy his neighbor's homes. The court ruled that the defendant's lack of intelligence did not override his duty of care to his neighbors. The defendant was required to maintain the care of reasonably prudent person. Thus, ignorance was not an excuse for Mr. Menlove.

Similarly, in analyzing a litigant's actions in response to a discovery request, "reasonableness" requires courts to place themselves in the shoes of a reasonably educated litigator when determining if the responding party acted appropriately in conducting a *reasonable inquiry*.[20] Yet

[17] *Id.* 26 advisory committee's notes to 1937 adoption; *see also* Mancia v. Mayflower Textile Servs. Co., 253 F.R.D. 354, 357 (D. Md. 2008) ("The duty to make a 'reasonable inquiry' is satisfied if the investigation undertaken by the attorney and the conclusions drawn therefrom are reasonable under the circumstances."); Kinee v. Abraham Lincoln Fed. Sav. & Loan Ass'n, 365 F. Supp. 975, 982–84 (E.D. Pa. 1973). In *Kinee*, the court imposed sanctions for violation of Rule 11 because the plaintiffs decided to sue every lending institution in the phone book, rather than conducting an adequate investigation:

> The plaintiffs' attorneys set out a dragnet. Having put a large number of parties to the inconvenience, expense and possible anxiety of being sued, they then were able conveniently to separate the wheat from the chaff without great effort. . . . If the plaintiffs had attempted reasonable investigation, and if some of the lending institutions who were not proper parties had not cooperated in that investigation, then perhaps they would have been justified in undertaking the course of action which they undertook. But under the circumstances of this case, the course of action which they chose was grossly improper.

Kinee, 365 F. Supp. at 982–83.

[18] *See* Vaughan v. Menlove, 132 E.R. 490, 494 (C.P. 1837). In *Vaughan*, in an action concerning the defendant's liability for starting a fire, the court held:

> Instead . . . of saying that the liability for negligence should be co-extensive with the judgment of each individual, which would be as variable as the length of the foot of each individual, we ought rather to adhere to the rule which requires in all cases a regard to caution such as a man of ordinary prudence would observe.

Id.

[19] *Id.*

[20] GEORGENE M. VAIRO, AM. BAR ASS'N, RULE 11 SANCTIONS app. 02[b][1] (2004) (excerpts from the REPORT OF THE JUDICIAL CONFERENCE COMMITTEE ON RULES OF PRACTICE AND PROCEDURE TO THE CHIEF JUSTICE AND JUDICIAL COUNCIL OF THE UNITED STATES (1983)) ("The [1983] amendments of Rule 26 are aimed at protecting against excessive discovery and evasion of reasonable discovery demands. As amended Rule 26(b) would require the court, when certain condi-

judges usually know less about the facts of the case, the true amount in controversy, and the level of search and retrieval appropriate for each case than the litigators themselves.[21] The reasonableness assumption also requires that the court possess the requisite technological expertise to assess a discovery process objectively, or seek educational assistance from a special master.[22] It is the job of the attorneys to educate the court on these matters.

What are the traits of a reasonably educated litigator in our electronic world of discovery? Recent surveys show that not only do lawyers fail to learn the technical knowledge they need to properly assess what is reasonable in terms of preservation and document review, they also fail to follow the fundamental principle of the electronic discovery world: it is vital to discuss these discovery issues with your adversary as soon as possible after the matter commences.[23]

In a recent Federal Judicial Center ("FJC") survey, only one in three respondents reported that their 26(f) conference to plan discovery included a discussion of ESI.[24] More than half of all respondents reported that the conference did not include discussion of ESI.[25] Equally frightening was the finding that only one in five court-ordered discovery plans included provisions relating to ESI.[26] 46.5% of plaintiffs' lawyers and 55.5% of defense lawyers reported issues related to the retention (preservation) of ESI; more than 30% of the plaintiffs' lawyers reported issues

tions exist, to limit the frequency and extent of use of discovery methods. Rule 26(g) would impose upon each party or attorney the duty, before proceeding with respect to any discovery matter, to make a reasonable inquiry and to certify that certain standards have been met. A violation of this duty would result in the imposition of sanctions.").

[21] Bell Atl. Corp. v. Twombly, 550 U.S. 544, 560 n.6 (2007) ("The judicial officer always knows less than the parties, and the parties themselves may not know very well where they are going or what they expect to find. A magistrate supervising discovery does not—cannot—know the expected productivity of a given request, because the nature of the requester's claim and the contents of the files (or head) of the adverse party are unknown. . . . The portions of the Rules of Civil Procedure calling on judges to trim back excessive demands, therefore, have been, and are doomed to be, hollow. We cannot prevent what we cannot detect; we cannot detect what we cannot define; we cannot define 'abusive' discovery except in theory, because in practice we lack essential information." (internal quotation marks omitted) (quoting Frank H. Easterbrook, *Discovery as Abuse*, 69 B.U. L. REV. 635, 638–39 (1989)).

[22] Telephone interview with Hon. James C. Francis, U.S. Magistrate Judge, S.D.N.Y. (Nov. 23, 2009) ("The reasonable standard of discovery isn't one of an average random person. A court should assume that the responding party had training on the subject matter [i.e., how to conduct an investigation to respond to a discovery request]. Consider the duty of care of an airline pilot: the pilot's conduct will be judged against the standard of a well-trained pilot, not an average person on the street.").

[23] *See* Anne Kershaw, *Talking Tech: Automated Document Review Proves Its Reliability*, DIGITAL DISCOVERY & E-EVIDENCE: BEST PRACS. & EVOLVING L., Nov. 2005, at 10, 10–12, *available at* https://www.lexisnexis.com/applieddiscovery/NewsEvents/PDFs/200511_DDEE_LegalLandscape.pdf.

[24] *See* EMERY G. LEE III & THOMAS E. WILLGING, FED. JUDICIAL CTR., CASE-BASED CIVIL RULES SURVEY: PRELIMINARY REPORT TO THE JUDICIAL CONFERENCE ADVISORY COMMITTEE ON CIVIL RULES 15 (2009), http://www.fjc.gov/public/pdf.nsf/lookup/dissurv1.pdf/$file/dissurv1.pdf.

[25] *Id.*

[26] *Id.* at 16.

with preservation, collection, review and productions; 41.9% of the defense lawyers reported issues with restricting the scope of discovery of ESI; 37.5% had issues with respect to the preservation of ESI; and 36% had issues pertaining to the collection, review and production of ESI.[27]

These numbers are staggering for a profession that imposes on all lawyers an ethical obligation to be qualified to undertake the representation of their clients. Unfortunately, because the lawyers and judges do not understand the technology, they blame the technology and volume of ESI for these issues, failing to recognize how so many of these issues could have been avoided with a little self-education.

Problems and conflict arise when litigators lack the requisite knowledge and education (just like Mr. Menlove) to build and deploy defensible inquiry strategies.[28] An ignorant counsel's failure to seek guidance from appropriate outside resources only heightens the problem.[29] As a result of poor education and inadequate guidance, litigators lack the ability to define the reasonable inquiry standard.[30] Litigators that misjudge the reasonable inquiry fall into two categories: attorneys that failed to meet their obligations due to the lack of effort in conducting a reasonable inquiry (the "under-inquiry") to the detriment of the requesting party; and attorneys that cast a risk-averse overbroad net of inquiry (the "over-inquiry") to the financial disadvantage of his client.

a. The Under-Inquiry

The reasonable inquiry of a certification of discovery responses requires counsel to conduct an independent investigation of the facts and produce what is specifically requested.[31] Courts tend to rule in favor of sanctions against a certifying attorney for lack of a reasonable inquiry

[27] *Id.* at 16–17.

[28] In considering if a responding attorney acted reasonably, a court might consider the methods an educated counsel chose to collect and produce the relevant information. The court might consider how counsel weighed the value of the material sought against the burden of providing it, and how he communicated that analysis to his opponent. For example, a litigant might restrict the number of searched custodians based upon the type of case, amount in controversy, and number of key players surrounding the case. Although courts have generally accepted an employee-centric "key player" preservation and production model, the definition of "key players" has caused some debate amongst clients and counsel. *See* Concord Boat Corp. v. Brunswick Corp., No. LR-C-95-781, 1997 WL 33352759, at *1 (E.D. Ark. Aug. 29, 1997) (finding that an employee-centric preservation model under attorney supervision was reasonable and did not indicate a failure to meet obligations imposed by law). *See generally* The Pension Comm. of the Univ. of Montreal Pension Plan, et al. v. Banc of Am. Sec. LLC, et al., No. 05 Civ. 9016 (SAS), 2010 WL 184312 (S.D.N.Y. Jan. 15, 2010) (discussing an employee-centric preservation model).

[29] Assessment of appropriate outside resources should include considerations of experience, case history, education, and whether the resource stands to gain financially from proffering certain advice.

[30] *See generally In re* Fannie Mae Sec. Litig., 552 F.3d 814 (D.C. Cir. 2009) (reviewing the district court order holding party in contempt for failing to comply with a discovery deadline).

[31] *See* FED. R. CIV. P. 26(g).

when that attorney failed to discover the obvious.[32] For example, in *R & R Sails, Inc. v. Insurance Co. of Pennsylvania*, a defendant insurance company certified that it conducted a reasonable inquiry, but failed to produce an electronic claim log after the plaintiff repeatedly requested the log.[33] After the defendant's employee certified that the claim log did not exist, the same employee realized that the log was accessible on his own computer.[34] United States Magistrate Judge Porter held (1) the defendant lacked evidence of the inquiry made by counsel; (2) the log could have been easily found; and (3) the certifying employee was accessing the log immediately prior to the certification and, therefore, no reasonable inquiry was completed, nor was the lack thereof substantially justified.[35] Thus, a responding party must make a reasonable inquiry and document the steps it took to respond to the discovery request if the party expects to defend his search in court.

Similarly, in the trademark dispute *Gucci America, Inc. v. Costco Wholesale*, the court ruled for sanctions against defendant Costco for failing to timely disclose cost figures related to the infringing jewelry sold by defendant.[36] Magistrate Judge Ronald Ellis held that Costco "failed in its obligation under [Rule] 26(g) to make a 'reasonable inquiry' to ensure its disclosure was complete and accurate."[37] In his rationale, Judge Ellis indicated that defendant Costco was notified of the request early in the litigation, did not produce the cost information after Gucci reiterated its request, and did not produce the information after the court ordered it to do so.[38] Furthermore, the court found that Costco

[32] *See, e.g.*, R & R Sails Inc. v. Ins. Co. of Penn., 251 F.R.D. 520, 525 (S.D. Cal. 2008) ("[T]o give meaning to the certifications provided on discovery responses, Rule 26(g) requires attorneys or parties to sign their responses 'after a reasonable inquiry.'").

[33] *Id.*

[34] *Id.* ("Evidence of such an inquiry prior to January 2007 [certification] may provide this Court with justification for the incorrect certifications provided to Plaintiff. Instead, this Court is presented with evidence that Lombardo was maintaining a claim log on his own computer using the AEGIS system while failing to recognize that this log was the same 'record/log' being requested by Plaintiff. Lombardo entered notes of a communication with counsel into the AEGIS system on November 16, 2007, immediately prior to counsel's representation to this Court that such a system was not possessed by Defendant and close in time to his signing a declaration that no such notes are maintained. The Court cannot find that a reasonable inquiry was made into whether Defendant possessed discovery responsive to Plaintiff's requests, and therefore the Court does not find Defendant's incorrect certifications to be substantially justified." (citations omitted)).

[35] *Id.*

[36] No. 02 Civ. 3190 (DAB) (RLE), 2003 WL 21018832, at *2 (S.D.N.Y. May 6, 2003) ("A simple inquiry to Costco's Accounts Payable Operations department revealed that the cost information was easily retrievable. Costco has not shown that producing the document earlier would have required any extraordinary diligence. Early in this litigation, Costco was on notice that Gucci sought information on costs for commercial dealings in Gucci items. At a conference held on October 7, 2002, the Court instructed Costco to produce information for jewelry items bearing the terms 'Gucci link' or 'Gucci style.' On November 21, 2002, the Court again reminded Costco of its obligation to timely disclose information pertaining to the items. After Costco produced records lacking cost figures, Gucci questioned its jewelry buyer who lacked any knowledge about records. In February 2003, Gucci reiterated its requests for the information. Yet Costco waited until the end of discovery, after the issue was brought before the Court, to conduct a thorough search.").

[37] *Id.*

[38] *Id.*

lacked any evidence indicating that an earlier document production would have required any extraordinary diligence and waited until the end of discovery to conduct a thorough search.[39] Consequently, Costco was ordered to pay costs and attorneys fees as a sanction for its poor discovery conduct.[40]

Other courts look to a set of factors when determining if the responding party conducted a reasonable inquiry. For example, in the insurance dispute *St. Paul Reinsurance Co. v. Commercial Financial Corp.*, the plaintiff filed consistent boilerplate objections to defendants repeated discovery requests.[41] In turn, the defendant filed a motion for expedited relief pursuant to Rule 57. Chief Judge Mark Bennett opened his order with a memorable quote:

> Anatole France, a late 19th and early 20th century French writer, urbane critic and Nobel Prize winner penned: "It is human nature to think wisely and to act in an absurd fashion." Little could France foresee that he would decades later capture the essence of plaintiffs' counsel's "Rambo" style discovery tactics in this litigation.[42]

Judge Bennett ruled that "[c]ounsel need not conduct an exhaustive investigation, but only one that is reasonable under the circumstances," and he provided four relevant circumstances to consider when challenging a reasonable inquiry.[43] These included: "(1) the number and complexity of the issues; (2) the location, nature, number and availability of potentially relevant witnesses or documents; (3) the extent of past working relationships between the attorney and the client, particularly in related or similar litigation; and (4) the time available to conduct an investigation."[44] The court held that counsel's behavior "plummet[ed] far below any objective standard of reasonableness. Indeed, every single objection is not only obstructionist and frivolous, but, as demonstrated above, is contrary to the Federal Rules of Evidence and well-established federal law."[45]

[39] *Id.*

[40] *Id.* at *1.

[41] 198 F.R.D. 508, 511 (N.D. Iowa 2000).

[42] *Id.* at 510. Anatole France (1844-1924), a pseudonym for Jacques Anatole François Thibault, was one of the major figures of French literature in the late nineteenth and early twentieth centuries. He was awarded the Nobel Prize for Literature in 1921. *See* Petri Liukkonen, *Anatole France* (2008), http://www.kirjasto.sci.fi/afrance.htm. Other variations of the quoted aphorism include: "It is human nature to think wisely and act foolishly" and "It is in human nature to think wisely and to act in an absurd fashion."

[43] *St. Paul Reinsurance Co.*, 198 F.R.D. at 516 n.3.

[44] *Id.* (citing Dixon v. Certainteed Corp., 164 F.R.D. 685, 691 (D. Kan. 1996)). "Under Rule 26(g), a 'signature certifies that the lawyer has made a reasonable effort to assure that the client has provided all the information and documents available to him that are responsive to the discovery demand. What is reasonable is a matter for the Court to decide on the totality of the circumstances.'" *Id.* (citation omitted) (quoting FED. R. CIV. P. 26(g)). "'Under Rule 26(g)(2) . . . [the subject of the inquiry] is the thoroughness, accuracy and honesty (as far as counsel can reasonably tell) of the responses and the process through which they have been assembled.'" *Id.* (alterations in original) (quoting Poole *ex rel.* Elliott v. Textron, Inc., 192 F.R.D. 494, 503 (D. Md. 2000)).

[45] *Id.* at 517.

"Because Rule 26(g) 'mandates that sanctions be imposed on attorneys who fail to meet the standards established in the first portion of 26(g),'"[46] the judge issued sanctions.[47]

b. The Over-Inquiry

Many litigators may surmise that the best way to avoid sanctions is to do everything possible to satisfy discovery demands. After all, the thinking goes, if a responding party does everything possible, how could anyone argue that you did not make a reasonable inquiry? Unfortunately, when attorneys couple this thinking with the challenges of electronic data, the results can be disastrous for clients.

Courts tend to hold litigants to higher standard of inquiry if they expressly agree to it, even without a reasonable inquiry into the facts before agreement. For example, in *In re Fannie Mae Securities Litigation*, the attorneys for the Office of Federal Housing Enterprise Oversight (OFHEO), a non-party responding to subpoenas, entered into a stipulated order to search, *inter alia*, disaster-recovery tapes without first assessing the time and costs associated with doing so.[48] The OFHEO spent over $6 million—more than 9% of its annual budget—attempting to comply, but ultimately failed in its efforts. The OFHEO was subject to a finding of contempt and the imposition of sanctions. The findings were affirmed on appeal.

In the recently settled case, *Advanced Micro Devices, Inc. v. Intel Corp.*, Intel engaged in a massive preservation effort, issuing a litigation hold initially to 4,000 employees and preserving thousands of backup tapes. [49] When mistakes were discovered, Intel fell on its sword and proffered an elaborate and expensive "discovery remediation" plan. This naturally led to discovery regarding compliance with the discovery remediation plan and surely placed the costs for discovery well outside the proportionality requirements set forth in the Rules. Regrettably, trying to boil the ocean to comply with discovery does little more than fuel the notion that discovery can and should be perfect when in fact it never can be. Indeed, the process of over-inquiry is more likely to lead to sanc-

[46] *Id.* at 516 (quoting *Poole*, 192 F.R.D. at 503).

[47] *Id.* at 517. Another interesting case looks to an attorney's conduct outside of the case in dispute. *See* Thibeault v. Square D Co., 960 F.2d 239, 246 (1st Cir. 1992). ("In considering sanctions for lapses in the course of pretrial discovery, a district court should consider all the circumstances surrounding the alleged violation. The totality of the circumstances can include events which did not occur in the case proper but occurred in other cases and are, by their nature, relevant to the pending controversy. Once the district court has recognized a pattern of misbehavior on an attorney's part, the court would be blinking reality in not taking counsel's proven propensities into account. . . . [A] trial court may properly give some consideration to a lawyer's behavior in previous cases when determining whether to accept the attorney's explanation of why he failed to comply with Rule 26(e) in a current case." (citations omitted)).

[48] 552 F.3d 814, 816 (D.C. Cir. 2009).

[49] *See* 258 F.R.D. 280, 282–83 (D. Del. 2008).

tions and excess costs than a reasoned, contained, and sustainable scope and process for discovery.

The "over-under" tug-of-war inquiry proves only one thing: A responsible attorney must balance the costs of electronic discovery with the duty to thoroughly review all of the relevant documents. This requires new tools and education in electronic discovery. Defaulting to traditional methods is no longer an option.

2. Attorney Document Review

Typical case law addressing Rule 26(g)'s reasonable inquiry requirement focuses on the investigation, preservation, and collection of discovery material. In the examples above, either the responding attorney did too little, or alternatively promised "the moon and the stars" only to face unwanted consequences. Yet, beyond an investigation, preservation plan, and collection, stands another significant "reasonable inquiry" methodology under scrutiny: attorney document review.

Just as attorneys require education and diligence in balancing reasonableness for an early investigation of the case, counsel must possess the very same skills in searching, analyzing, and categorizing discovery material once sources have been identified and collected. Prior to the proliferation of computers and e-mail, when discovery was limited to file cabinets that interviewing attorneys could search during the client interviews, human review seemed logical. Now, when the average employee hard drive has the capacity to store 160 gigabites (the equivalent of 12,000,000 text pages), the "reasonable inquiry" standard does not end at the desk drawer of the employee; attorneys must now search supertankers full of documents.[50] And that is just the beginning: although the cases above discuss the physical search efforts, what about the time spent after documents have been collected?[51]

Review and analysis methodology for the reasonable inquiry has not kept pace with every changing technology, and the rapid data growth that drives it. Unfortunately, many attorneys, judges, and other practitioners still maintain the mindset that traditional brute-force page-by-page attorney document review *is* a best practice when responding to massive dis-

[50] *See, e.g.*, Ralph Losey, How Much Data Do You Have?, http://e-discoveryteam.com (last visited Mar. 7, 2010).

[51] *See* Sedona Conference WGS, *supra* note 5, at 198 ("Historically, outside counsel played a key role in the discovery process, and the process worked simply. Litigants, assisted by their counsel, identified and collected information that was relevant to pending or foreseeable litigation. Counsel reviewed the information and produced any information that was relevant and not otherwise protected from disclosure by the attorney–client privilege, the attorney work product or by trade secret protections. This worked fine in the days where most of the potentially relevant information had been created in or was stored in printed, physical form, and in reasonable volumes so that it required only 'eyes' to review and interpret it.").

covery requests.[52] Regardless of its effectiveness, the human review process simply cannot keep pace with the speed at which society is accumulating data.[53] In 2008 alone, the Interactive Data Corporation says the world created 487 billion gigabytes of information, up 73% from 2007. That was 3% more than it forecasted at the beginning of the year. This trend shows no sign of slowing: going forward, the IDC forecasts "the digital universe will *double every 18 months*."[54]

E. Old Dog, Old Tricks

If the legal world generally understands that the volume of data is increasing exponentially, and that firms and their clients cannot sustain the overwhelming burden of reviewing, why do some litigators continue to rely primarily upon manual review of information to conduct a reasonable inquiry?[55] Attorneys have many different defensible search options available to deploy, some more technologically advanced than others.[56] The problem is not technology; it is attorneys' lack of education and the judicial system's inattentiveness to ensure that attorneys have the proper education and training necessary for a proportional and efficient discovery process. Lack of attorney education aggravates the problem because uneducated litigators are unable to make informed judgments as to where to draw the line on discovery, thereby creating unrealistic expectations from the courts—particularly as to costs and burdens. For example, failing to understand how different methods of search methodology work, some judges will unnecessarily mandate traditional and expensive "brute force" attorney review.

Why do attorneys overestimate their judgment, yet lack a realistic view of their level of precision? Uneducated lawyers tend to rely on old

[52] *See id.* at 198–99 ("Accordingly, the conventional discovery review process is poorly adapted to much of today's litigation. . . . It is not possible to discuss this issue without noting that there appears to be a myth that manual review by humans of large amounts of information is as accurate and complete as possible—perhaps even perfect—and constitutes the gold standard by which all searches should be measured. Even assuming that the profession had the time and resources to continue to conduct manual review of massive sets of electronic data sets (which it does not), the relative efficacy of that approach versus utilizing newly developed automated methods of review remains very much open to debate.")

[53] *Id.* at 198.

[54] William M. Bulkeley, *The Exploding Digital Universe*, WALL ST. J. BLOGS, May 18, 2009, http://blogs.wsj.com/digits/2009/05/18/the-exploding-digital-universe (emphasis added); *see also* EMC, Digital Universe, http://www.emc.com/digital_universe (last visited Mar. 7, 2010) (providing a real-time "Worldwide Information Growth Ticker" that measures the bytes of information created since Jan. 1, 2010).

[55] *See* KROLL ONTRACK, THIRD ANNUAL ESI TRENDS REPORT 9 (2009), http://www.krollontrack.com/library/esitrends3_krollontrack2009.pdf (discussing challenges in responding to ESI in Finding 12); *see also* JOHN GANTZ & DAVID REINSEL, IDC, AS THE ECONOMY CONTRACTS, THE DIGITAL UNIVERSE EXPANDS 1 (2009), http://idcdocserv.com/EMC_MMWP_Digital_Universe (discussing how increased data volume is the most significant factor bolstering the increased spending on electronic discovery).

[56] For a listing of search and retrieval techniques see Sedona Conference WGS, *supra* note 5, at 217–18.

tricks and the oxymoron of "gold-standard attorney document analysis," which does not necessarily amount to a high level of precision when attempting to review documents for relevancy. "For example, in the *Blair and Maron* study, attorneys over-estimated their ability to create and develop queries to assess the relevancy of 40,000 documents relevant to a transit accident."[57] Additionally, "[l]awyers estimated that their refined search methodology would find 75% of relevant documents, when in fact the research showed only 20% or so had been found."[58] Clearly, a more educated approach would eliminate such unjustified conclusions.

Counsel's overestimate of human ability could be based on a variety of factors. First, attorneys' false notion of accuracy could emanate from the core training attorneys receive early in their careers—including law school—where using keyword search methods and basing decisions on history and precedent is encouraged.[59] Such training results in an aversion to a change in methodology, and attorneys view that change, however reasonable, as risky. Accordingly, attorneys continue to rely on keyword search methods bolstered by attorney review. A baseless lack of education and knowledge by the bar only seems to accelerate the problem.[60]

Furthermore, there is a noticeable lack of positive feedback when attorney *do* conduct efficient discovery. Judges don't award "gold stars" in published opinions and orders to practitioners that conduct a reasonable search and review of information.[61] Unfortunately, litigators only read the horror stories of when things go wrong, or how counsel failed to perform a reasonable inquiry in the discovery phase by using inappropriate or overbroad keyword search terms.[62] Although a properly researched and executed keyword search strategy can meet the reasonable inquiry

[57] Patrick L. Oot, *The Protective Order Toolkit: Protecting Privilege with Federal Rule of Evidence 502*, 10 SEDONA CONF. J. 237, 239 (2009) (citing David C. Blair & M. E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMM. ACM 289 (1985)).

[58] *Id.* (internal quotation marks omitted) (quoting Sedona Conference WGS, *supra* note 5, at 206).

[59] Attorneys learn their inquiry strategy from Lexis, Westlaw, and Google. Keyword searching works well in these structured databases as they tend to exist in a far more categorized state than litigation data. For example, Lexis and Westlaw have a team of editors to spell-check and categorize cases by key concepts and headnotes. In addition, these databases only categorize a very limited number of document types.

[60] *See, e.g.*, William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 134 (S.D.N.Y. 2009) (issuing a "wake-up call" to the Bar of the district).

[61] *See* Roitblat, Kershaw & Oot, *supra* note 3, at 71 ("Web searches are generally fairly specific, for example, 'What are the best sites to visit in Paris?' In contrast, the information need in eDiscovery is generally much broader and more vague. Discovery requests include statements like 'All documents constituting or reflecting discussions about unfair or discriminatory allocations of [Brand X] products or the fear of such unfair or discriminatory allocations.' These requests will not typically be satisfied by one or a few documents"). Standards of reasonableness lack certainty, as a result attorneys show reluctance to change from a "generally accepted standard" to a new untested method, even if that method is more reasonable under the circumstances.

[62] *See* cases cited *infra* notes 84, 86, 91, and 93; *see also William A. Gross Constr. Assocs., Inc.*, 256 F.R.D. 134.

standard under Rule 26, many alternative forms of search criteria exist beyond the traditional human review[63]—yet the opinions tend to be limited to keyword search terms and attorney review.

Simply put, the legal system has a crisis of education. Both attorneys and judges need to better understand technology as it applies to the reasonable inquiry. Education initiatives mandated by state bars, law schools, and advisory committee notes could help alleviate the problem.

II. Toward a Reasonable Inquiry: Research in Finding a Better Way to Conduct Electronic Discovery

Fortuitously, legal commentators have already started educational and research programs to target the problem of discovery response. However, it is the obligation of the bar, law schools, educators, and courts to ensure that these research programs reach the intended audience.

A. TREC

The United States government has taken an interest in text retrieval generally in a venture co-sponsored by the National Institute of Standards and Technology ("NIST") and U.S. Department of Defense.[64] "TREC (Text Retrieval Conference) is a multi-track project sponsored by the National Institute for Standards and Technology and others to conduct comparative research on text retrieval technologies."[65] Since 2006, "TREC has included a legal track whose goal is to assess the ability of information retrieval technology to 'meet the needs of the legal community for tools to help with retrieval of business records.'"[66] In support of this goal, they seek to develop and apply collections and tasks that approximate the data, methods, and issues that real attorneys might use

[63] *See generally* DOUGLAS W. OARD ET AL., OVERVIEW OF THE TREC 2008 LEGAL TRACK, http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf (last visited Mar. 7, 2010); Feng C. Zhao et al., Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback (DESI III Global E-Discovery/E-Disclosure Workshop at ICAIL 2009), http://www.law.pitt.edu/DESI3_Workshop/Papers/DESI_III.Zhao_Oard_Baron.pdf. A properly executed keyword search method includes developing keyword search terms from interviews with data users to identify terms-of-art, acronyms, or other non-traditional search terms, testing those search terms by sampling the terms that were hits, as well as the misses, and communicating the search strategy to your opponent. Some litigants suggest checking the extracted text index for likely misspellings and a second meet and confer could provide better search results.

[64] *See* Text REtrieval Conference, Nat'l Inst. of Standards and Tech., Overview, http://trec.nist.gov/overview.html (last visited Mar. 7, 2010) ("The Text REtrieval Conference (TREC), co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense, was started in 1992 as part of the TIPSTER Text program. Its purpose was to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies.").

[65] Roitblat, Kershaw & Oot, *supra* note 3 at 71.

[66] *Id.* at 71–72 (quoting OARD ET AL., *supra* note 63, at 1).

during civil litigation, and to apply objective criteria by which to judge the success of various search methodologies.[67]

B. Electronic Discovery Institute Study

Similarly, *The Electronic Discovery Institute* has conducted a recent study (the "EDI study") to determine if service providers offering auto-categorization technology can equal or surpass the performance of a human based attorney document review system. The study hypothesized that if the service providers could equal the response of a real-life attorney review team in a significantly disputed regulatory filing, the methodology used by the service provider would meet the reasonable inquiry test of Rule 26(f). Little public information then existed regarding the comparison of traditional litigation document review methodologies with alternative technology approaches.[68]

The EDI study based its inquiry on a completed Verizon matter. In 2005, as part of an acquisition of a competitor, Verizon responded to a governmental request for additional information for material relevant to the proposed acquisition; also known as a "Second Request" under the Hart-Scott-Rodino Antitrust Improvements Act of 1976.[69] To respond to the Second Request, Verizon completed a voluminous document review project: the company collected documents from 83 employees in 10 states, consisting of 1.3 terabytes of electronic files in the form of 2,319,346 documents. The collection included close to 1.5 million email messages, 300,000 loose files, and 600,000 scanned documents. After eliminating duplicates, 1,600,047 items were submitted for attorney review. It took the attorneys four months, working sixteen hours per day seven days per week, for a total cost of $13,598,872.61 or about $8.50 per document. This sum included the fees of outside counsel specifically assigned to document review tasks (but not attorneys working on other aspects of the case), and hourly fees billed by contract attorneys hired specifically for the review.

After spending many long hours managing the document review for the Second Request response, Verizon attorneys John Frantz and Patrick Oot agreed to seek a better alternative to the process of traditional human document review. Frantz and Oot teamed up with Anne Kershaw and Herb Roitblat, and shortly thereafter, Oot, Kershaw, and Roitblat (the authors of this Article) formed *The Electronic Discovery Institute*, a non-

[67] For general information about TREC, see TREC Legal Track, http://trec-legal.umiacs.umd.edu/ (last visited Mar. 7, 2010).

[68] *See* Kershaw, *supra* note 23, at 12. Commentator Anne Kershaw also completed a non-public study that compared a traditional human review process with an automated electronic review process. *Id.* Ms. Kershaw's study revealed that using an electronic process to assist the document review reduced the chances of missing relevant documents by as much as 90%. *Id.*

[69] *See* Roitblat, Kershaw & Oot, *supra* note 3, at 73. Verizon is New York based telecommunication carrier-corporation listed on the New York Stock Exchange (VZ). For more information about Verizon, see http://www.verizon.com.

profit research institution organized to complete a peer-reviewed study comparing human document review with technology assisted document review.

Unlike many data compilations available to researchers that have been culled (perhaps using keyword search terms or some other form of analysis) prior to human review, the data set of the EDI Study included attorney relevancy decisions made by the original review team on every user created computer readable document collected from Verizon employees.[70] The EDI study sought to compare these final decisions of Verizon's original attorney review with the results of any provider willing to undertake the significant task of responding to an already completed Second Request.[71] Two providers responded. Over 75 people and organizations donated their time, services, and work product to the study.[72]

1. Methodology

The EDI study compared two different methodologies of computer assisted document categorization with the original attorney review. Similar to the human review, the study required the computer-assisted systems to submit document relevancy decisions on the entire corpus of computer readable user documents. The computer assisted systems did not have any knowledge of the original attorney review team's categorization decisions.

A second human review was also completed on a sample set of 5,000 documents to compare the decision made by a second set of attorneys to the decisions made by the original review team. Again, the computer assisted systems did not have any knowledge of the original attorney review team's categorization decisions.

The EDI study sought to compare the level of agreement between the original attorney review system and a second attorney review system. The EDI study also sought to compare the levels of agreement of the original attorney review with the two computer assisted systems individually. The hypothesis of the study stated that the computer systems

[70] Many studies rely on pre-culled or reviewed datasets to analyze the reasonableness of search retrieval methodology because these sets are all that is publically available. Both the Enron litigation data set and the tobacco litigation data set are publicly available to those desiring to study search and retrieval techniques. However, the Verizon data set was a real-life compilation of company documents collected from employees containing confidential and privileged information. The data is not available to the public and never left the custody or control of Verizon, its legal service providers, and law firms. Search and retrieval researchers are seeking raw data sets. Please contact info@electronicdiscoveryinstitute.org for more information.

[71] In 2006, EDI gave an open invitation to the litigation technology community to participate in the study, three service providers responded, of which, two finally agreed to participate under the methodology created by the founders. In the years after the invitation closed, many other service providers requested the ability to participate. Although the founders limited the first study to the two original providers, there has been internal discussion on whether to expand participation. EDI will revisit this discussion in November, 2010.

[72] The authors suggest reading the study in conjunction with this law review article.

will agree with the original attorney review at least as frequently as the second attorney review. The teams were not instructed to achieve the best results possible, but to equal or surpass traditional manual review.[73] Specifically, the study set out to demonstrate that auto-categorization technology will agree with the court-accepted standard of attorney review at least as frequently as a second attorney review of the same material.[74]

For the first of two auto-categorization systems, the participant deployed two new attorney teams that re-reviewed a sample set of documents. The attorneys received the same training materials and documentation that the original second request attorney review team received. A senior litigator also reviewed and categorized documents on which the two teams disagreed. The senior litigator arbitrated the conflicting document decisions without the knowledge of either teams' document decisions. The participating service provider then deployed proprietary automation algorithms that categorized the remaining documents dependently upon the original sample set. The service provider then validated the results by sampling another subset, and then repeated the sampling until it reached what it felt was a reasonable confidence level. No guidance was offered to the service provider from the client beyond the original sampling and senior litigator document decision arbitration. The entire process was completed in less than four weeks.

The second auto-categorization system did not request attorney review teams. The second system relied upon the documentation that the original second request attorney review team received. In addition, the service provider received answers to a set of approximately thirty interrogatories that it prepared and were answered by a senior litigator. The second service provider deployed proprietary automation algorithms, linguistics-based queries, legal professionals, computer scientists, computational linguists, mathematicians, and statisticians. The workflow included: computer assisted categorization, testing by sample review, assessment of different sample responses, adjustments and multiple supplemental iterations. No guidance was offered to the service provider from the client beyond the senior litigator interrogatory answers. The entire process was completed in less than four weeks.

[73] Both service providers indicated that auto-categorization systems could actually perform even better than their submitted responses in this study given unlimited time, expense, and resources, but the EDI challenge was to perform as well as humans, an accepted standard.

[74] *See* Roitblat, Kershaw & Oot, *supra* note 3, at 73 ("In the ideal case, we would like to know how accurate each classification is. Ultimately, measurement of accuracy implies that we have some reliable ground truth or gold standard against which to compare the classifier, but such a standard is generally lacking for measures of information retrieval in general and for legal discovery in particular. In place of a perfect standard, it is common to use an exhaustive set of judgments done by an expert set of reviewers as the standard (e.g., as is the practice in the TREC studies). Under these circumstances, agreement with the standard is used as the best available measure of accuracy, but its acceptance should be tempered with the knowledge that this standard is not perfect.").

The EDI study "set out to answer the question of whether there was a benefit to engaging a traditional human review or whether computer systems could be relied on to produce comparable results."[75] It concluded that "the performance of the two computer systems was at least as accurate (measured against the original review) as that of a human re-review."[76]

In addition to the core conclusion of the EDI study, the authors of the study made several additional observations:

- Many lawyers and judges need education regarding "reasonable inquiry" discovery response techniques.

- Litigants should consider cooperation with an opponent early to establish a search protocol.

- All categorization systems require some level of educated interaction. Better results result occur when knowledge is transferred early and continuously throughout the process.

- The use of auto-categorization systems can potentially *reduce document request response times* from over four months to as little as thirty days for even the largest datasets. Assumingly, requesting parties desire their documents faster, as speedy response will allow a receiving party to conduct a more thorough and complete investigation.

- Government agencies should consider acceptance guidelines for responding to document requests using auto-categorization technology.

- Human review is of unknown accuracy and consistency.

- Measurement against an accepted standard is essential to evaluating reasonableness.

- A litigant should sample at least 400 results of both responsive and non-responsive data.[77]

- Using auto-categorization *will save money and time*.

[75] *Id.* at 79.

[76] *Id.*

[77] E-mail from Maura R. Grossman, Counsel, Wachtell, Lipton, Rosen & Katz, to Patrick Oot, Director, Electronic Discovery Institute (Nov. 21, 2009, 13:38 EST) (on file with authors) ("If you have a document collection containing 100,000 or more documents (which is not atypical these days in the litigation or investigatory context), and you take a random sample of 384 documents—meaning that every document in the collection has an equal chance of being selected for inclusion in the sample—you can have 95% confidence (that is, if you repeated the exercise 100 times, you could expect to get a similar result 95 out of 100 times), that your sample has an error rate of no more than plus or minus 5%. If, instead, you were to use a random sample of 596 documents, you would have the same 95% confidence interval, but an error rate of no more than plus or minus 4%. Therefore, it seemed to me that, for the average matter with a large amount of ESI, and one which did not warrant hiring a statistician for a more careful analysis, a sample size of 400 to 600 documents should give you a reasonable view into your data collection, assuming the sample is truly randomly drawn.").

- Based upon the service provider cost submissions, had Verizon used auto-categorization in its Second Request response, EDI concludes that there would have been a minimal measurable cost savings of $5 million using 2006 pricing.

- As data volumes increase, auto-categorization may be the only practical solution to massive data sets common in today's corporations.

C. The Sedona Conference®

Both TREC and the Electronic Discovery Institute Study are educational initiatives proffered by the legal community to support considerations for alternate search and retrieval methodology. Also, The Sedona Conference® has developed an extensive commentary on selecting an appropriate search and retrieval method, titled *The Sedona Conference® Best Practices Commentary On The Use Of Search And Information Retrieval Methods In E-Discovery* ("Sedona Commentary").[78] It is interesting to note how the practice points set forth in The Sedona Commentary dovetail with the points made by the ongoing research.

For example, Practice Point 1 of the Sedona Commentary states: "In many settings involving electronically stored information, reliance solely on a manual search process for the purpose of finding responsive documents may be infeasible or unwarranted. In such cases, the use of automated search methods should be viewed as reasonable, valuable, and even necessary."[79] The EDI study clearly validates Practice Point 1. For example, the complete review of the dataset took many millions of dollars, many months, and the efforts of hundreds of attorneys to complete.[80] Had Verizon used either of the auto-categorization technologies used in the EDI study, it might have saved millions of dollars and several months of attorney labor.

Practice Points 2 and 3 of the Sedona Commentary hold respectively that "[s]uccess in using any automated search method or technology will be enhanced by a well-thought out process with substantial human input on the front end" and "[t]he choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed."[81] The EDI study again supports these points, as both systems relied upon a well-developed process with human interaction early on. For example, one system used real human review for input; the other relied on subject matter experts and interrogatories to develop a categorization system. Moreover, both service providers in the EDI study deployed resources appropriate for the legal context; a Second

[78] Sedona Conference WGS, *supra* note 5, at 189.

[79] *Id.* at 194.

[80] Roitblat, Kershaw & Oot, *supra* note 3, at 73.

[81] Sedona Conference WGS, *supra* note 5, at 194.

Request analyzes massive amounts of data. Therefore, auto categorization is more appropriate than in a request where the document volume is low or limited. Both service providers understood the options available for designing a well thought out process and selecting the appropriate search and retrieval method.

Sedona Commentary Practice Point 4 further underscores the need for attorney education and due diligence in stating that "[p]arties should perform due diligence in choosing a particular information retrieval product or service from a vendor."[82] Again, both service providers met the practice point standard. Through EDI's inquiries, both service providers were able to explain their techniques, justify their results based upon sampling, and maintain a reference list of prior cases and contacts EDI could call upon to test results. In addition, both service providers offered expert witnesses if auto-categorization techniques were challenged by an opponent or others.

Even so, perhaps the most important Sedona Commentary point is Practice Point 8, which states that "[p]arties and the courts should be alert to new and evolving search and information retrieval methods."[83] Redundant to the guidance of this article, Practice Point 8 mandates the need for attorney education and training.

D. The End of Keyword Search Methods

Search and retrieval techniques for discovery responses have rapidly evolved since the turn of the century. Just a few years ago, courts ordered keyword culling techniques in discovery without the use of any sort of testing methodology.[84] Although many attorneys still rely on keyword search culling techniques, recent decisions have placed significant scrutiny on keyword search term culling, especially where a responding party failed to sample, check, and verify the results.

Courts tend to rule against parties that fail to sample, test, and verify keyword search results. For example, in the drug liability case *In re Se-*

[82] *Id.* at 194.

[83] *Id.* at 195. Sedona Commentary Practice Points 5, 6 and 7 all fall under the category of being a "good lawyer," as follows: "Practice Point 5. The use of search and information retrieval tools does not guarantee that all responsive documents will be identified in large data collections, due to characteristics of human language. Moreover, differing search methods may produce differing results, subject to a measure of statistical variation inherent in the science of information retrieval." *Id.* at 194. "Practice Point 6. Parties should make a good faith attempt to collaborate on the use of particular search and information retrieval methods, tools and protocols (including as to keywords, concepts, and other types of search parameters)." *Id.* at 195. "Practice Point 7. Parties should expect that their choice of search methodology will need to be explained, either formally or informally, in subsequent legal contexts (including in depositions, evidentiary proceedings, and trials)." *Id.*

[84] *See* Medtronic Sofamor Danek, Inc. v. Michelson, 229 F.R.D. 550, 559 (W.D. Tenn. 2003) ("Using the vendor of its choice, [the plaintiff] shall search the 300gb of electronic data using the Boolean search terms."); *see also* Tessera, Inc. v. Micron Tech., Inc., No. C06-80024MISC-JW(PVT), 2006 WL 733498, at *8 (N.D. Cal. Mar. 22, 2006) ("The following search terms shall be run through electronic document databases for production to plaintiff.").

roquel Products Liability Litigation, defendant AstraZeneca selected keyword search terms to cull data prior to producing it to the requesting plaintiff.[85] The court concluded that the defendant significantly failed in meeting its discovery obligations. Magistrate Judge Baker ruled that:

> [W]hile key word searching is a recognized method to winnow relevant documents from large repositories, use of this technique must be a cooperative and informed process. Rather than working with Plaintiffs from the outset to reach agreement on appropriate and comprehensive search terms and methods, [counsel] undertook the task in secret. Common sense dictates that sampling and other quality assurance techniques must be employed to meet requirements of completeness.[86]

As a result, Judge Baker ordered sanctions against the defendant for its failure to produce readable and assessable documents.

Similarly, courts have identified the difficulty in selecting keyword search terms. In *United States v. O'Keefe*, the government charged defendants with receiving gifts for expediting visas while working at the Department of State in Canada.[87] The government searched for and produced documents using a self-selected Boolean search query of keyword search terms.[88] In *O'Keefe*, Judge Facciola held that "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."[89] The court emphasized the difficulty in selecting keyword search terms in its rationale for a Rule 702 analysis:

> 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 angels fear to tread.[90]

Coincidentally, Judge Facciola revisited his *O'Keefe* ruling shortly thereafter in the employment case *Equity Analytics, LLC v. Lundin*.[91]

[85] 244 F.R.D. 650, 651 (M.D. Fla. 2007).

[86] *Id.* at 662. The defendant failed to provide information "as to how it organized its search for relevant material, [or] what steps it took to assure reasonable completeness and quality control." *Id.* at 660 n.6.

[87] 537 F. Supp. 2d 14, 16 (D.D.C. 2008).

[88] *Id.* at 18. The query deployed by the government to locate relevant documents was "early or expedite* or appointment or early & interview or expedite* & interview." *Id.*

[89] *Id.* at 24.

[90] *Id.*

[91] 248 F.R.D. 331, 333 (D.D.C. 2008).

Again, Judge Facciola emphasized the difficulty in selecting search methodology, and the court's need of evidence to determine the validity of the search technique.[92]

Other courts have ruled against a particular search technique as unreasonable while identifying a multi-factor test to determine if a search methodology meets a reasonableness standard. In *Victor Stanley, Inc. v. Creative Pipe, Inc.*,[93] a litigant ineffectively settled on keyword search terms to cull for privilege.[94] Finding the search methodology unreasonable, Judge Grimm put forth a multi-factor analysis litigants should deploy when selecting search techniques:

> 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]

Courts are also requiring litigators to cooperate with one another by invoking a multi-step framework when selecting search methodology. For example, in the construction dispute *William A. Gross Construction Associates, Inc. v. American Manufacturers Mutual Insurance Co.*,[96] the responding party deployed overbroad and imprecise keyword search terms to respond to a discovery request.[97] Judge Peck warned:

> This opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or "keywords" to be used to produce emails or other electronically stored information ("ESI"). . . .
>
> . . . .
>
> This case is just the latest example of lawyers designing keyword searches in the dark, by the seat of the pants, without adequate (in-

[92] *Id.* ("[D]etermining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a lay person (and a lay lawyer) and requires expert testimony that meets the requirements of Rule 702 of the Federal Rules of Evidence.").

[93] 250 F.R.D. 251 (D. Md. 2008).

[94] *Id.* at 256–57.

[95] *Id.* at 262 ("[T]he Defendants have failed to demonstrate that the keyword search they performed on the text-searchable ESI was reasonable. Defendants neither identified the keywords selected nor the qualifications of the persons who selected them to design a proper search; they failed to demonstrate that there was quality-assurance testing; and when their production was challenged by the Plaintiff, they failed to carry their burden of explaining what they had done and why it was sufficient.").

[96] 256 F.R.D. 134 (S.D.N.Y. 2009).

[97] *Id.* at 134–35.

> deed, here, apparently without any) discussion with those who wrote the emails.[98]

Furthermore, the court ordered a multi-step framework that the litigators must use when selecting a keyword search strategy.[99] Judge Peck ordered that the litigators "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.'"[100] Again, the opinion illustrates how courts are developing factors to determine the reasonableness of a litigants search methodology.

E. Avoiding Costly Evidence Motions and Proceedings by Collaboration and Informal Conferences

Some litigants have argued that the *O'Keefe* and *Equity Analytics* rulings implicate the reliability of expert witness testimony under Federal Rule of Evidence 702 and the *Daubert* factors during pretrial discovery.[101] We disagree.

In *Daubert*, two minor children born with serious birth defects alleged that their defects were caused by their mothers' ingestion of Bendectin, a prescription anti-nausea drug marketed by Merrell Dow. After extensive discovery, the parties submitted conflicting reports experts on Bendectin causation. The issue in *Daubert* focused on how courts should analyze the validity of the scientific data in expert reports as an attempt to avoid junk science. On appeal, The U.S. Supreme Court ruled Federal Rule of Evidence 702 applied.[102] Thus, when courts must analyze conclusions of expert reports that draw difficult and often theoretical causa-

[98] *Id.*

[99] *See id.* at 136.

[100] *Id.*

[101] *See id.* at 592. In *Daubert*, the Supreme Court said that "[f]aced with a proffer of expert scientific testimony . . . the trial judge must determine at the outset . . . whether the expert is proposing to testify to (1) scientific knowledge that (2) will assist the trier of fact to understand or determine a fact in issue." *Id.* Accordingly, the district judge is generally required to "ensure that any and all scientific testimony or evidence admitted is not only relevant, but reliable." *Id.* at 589. In determining reliability, the Supreme Court further noted four non-exhaustive factors the district court may use in determining the reliability of scientific expert testimony: (1) whether a theory has been tested; (2) whether it has been subject to peer review; (3) whether a technique has a potential rate of error, or standard operating procedures; and (4) whether a theory is generally accepted within the scientific community. *See id.* at 592–94.

[102] The task of ensuring that an expert's testimony rests on both a reliable foundation and is relevant to the task at hand is assigned to the trial judge. Pursuant to Rule 104(a), the judge must make a preliminary assessment of whether the testimony's underlying reasoning or methodology is scientifically valid and can be properly applied to the facts at issue. Considerations include whether the theory or technique in question can be (and has been) tested, whether it has been subjected to peer review and publication, its known or potential error rate, and the existence and maintenance of standards controlling its operation, and whether it has attracted widespread acceptance within a relevant scientific community. The inquiry is a flexible one, and its focus must be solely on principles and methodology, not on the conclusions that they generate.

tion arguments, and the only way to measure probability of causation is through that expert opinion, a *Daubert* analysis is appropriate.

However, in testing the reasonableness of a producing party's inquiry in discovery, when the subject of the inquiry—the dataset—is available for repeated testing and sampling without harm to the data, then perhaps less formal methods of judicial analysis and questioning are appropriate. More simply, the causative factors that achieve a discovery outcome are not hidden.[103] In testing discovery response techniques, a litigant can test the search and retrieval criteria for validity easily, transparently, and repeatedly. The technical processes and results for discovery can be observed, measured, and reported to the court directly without expert testimony. All that the court and litigants really need in order to assess the reasonableness of a discovery process is the information as to what the parties did and why they believe it worked.

Moreover, neither *O'Keefe* nor *Equity Analytics* specifically held that FRE 702 conclusively applies to discovery proceedings.[104] Rather, the courts ruled that litigators consider FRE 702 to underscore the point that lawyers needed to look beyond their ordinary knowledge when dealing with matters involving technical concepts, such as the accuracy of keyword searches.[105]

We do not suggest limiting the court system's ability to discover truth.[106] We simply anticipate that judges will deploy more reasonable

[103] Unlike discovery, one cannot unring a bell in cases of toxic exposure to test and retest historical exposure levels. Since we cannot see the actual working of Bendectin in the human body, we have no choice but to rely on epidemiological studies and other evidence from which experts draw conclusions.

[104] Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 261 n.10 (D. Md. 2008) ("The *O'Keefe* and *Equity Analytics* opinions have raised the eyebrows of some commentators who have expressed the concern that they 'engraft [FED. R. EVID.] 702 (and [FED. R. EVID.] 104(a) into discovery . . . [which, it is feared] would multiply the costs of discovery', [sic] and, it is argued, this is a 'path [that] is rife with unintended consequences.' A careful reading of *O'Keefe* and *Equity Analytics*, however, should allay these concerns. In neither case did the court expressly hold that FED. R. EVID. 702 and 104(a) were 'engrafted' into the rules of discovery in civil proceedings (indeed, neither opinion even mentions Rule 104(a)). Instead, Judge Facciola made the entirely self-evident observation that challenges to the sufficiency of keyword search methodology unavoidably involve scientific, technical and scientific subjects, and *ipse dixit* pronouncements from lawyers. Observations 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. Certainly those concerned about the *O'Keefe* and *Equity Analytics* opinions would not argue that trial judges are not required to make fact determinations during discovery practice. Indeed, such fact determinations inundate them." (alterations in original) (citation omitted)).

[105] *See id.*

[106] Furthermore, courts have tremendous leeway in their quest for validated proof:

> In *Kumho Tire Co., Ltd.*, the Supreme Court explicitly extended the *Daubert* gatekeeping role to technical or other specialized expert testimony. The Supreme Court further noted that *Daubert*'s four factors for determining scientific reliability need not be applied in all cases. "Rather, we conclude that the trial judge must have considerable leeway in deciding in a particular case how to go about determining whether particular expert testimony is reliable." The Supreme Court also noted that the trial judge must have "discretionary authority . . . both to avoid unnecessary 'reliability' proceedings in ordinary cases where

and efficient standards to determine whether a litigant met his Rule 26(g)[107] reasonable inquiry obligations. Indeed, both the *Victor Stanley* and *William A. Gross Construction* decisions provide a primer for the multi-factor analysis that litigants should invoke to determine the reasonableness of a selected search and review process to meet the reasonable inquiry standard of Rule 26(f)[108]:

> 1. Explain how what was done was sufficient;
>
> 2. Show that it was reasonable and why;
>
> 3. Set forth the qualifications of the persons selected to design the search;
>
> 4. Carefully craft the appropriate keywords with input from the ESI's custodians as to the words and abbreviations they use; and
>
> 5. Use quality control tests on the methodology to assure accuracy in retrieval and the elimination of false positives.[109]

Interestingly, had any of the litigants in the above cases deployed the auto-categorization methodologies examined in the EDI study, the courts likely would have ruled differently because the study participants met the five-part test of reasonable inquiry distilled from *Victor Stanley* and *William A. Gross Construction*.[110]

CONCLUSION

The future of discovery and litigation rests heavily on a litigant's ability to respond to discovery requests accurately, inexpensively, and quickly. Lawyers must "wake-up" to Rule 1 of the Federal Rules of Civil Procedure: "to secure the just, speedy, and inexpensive determination of every action and proceeding."[111] The bar must empower itself to seek knowledge on reasonable discovery. Historically, lawyers would research and learn new areas of the law to better represent their clients and maintain a full understanding of their clients' case.[112] The nuances of search and retrieval to meet the Rule 26 reasonable inquiry standard must

> the reliability of an expert's methods is properly taken for granted, and to require appropriate proceedings in the less usual or more complex cases where cause for questioning the expert's reliability arises."

Bureau v. State Farm Fire & Cas. Co., 129 F. App'x. 972, 975 (6th Cir. 2005) (alteration in original) (citations omitted) (quoting Kumho Tire Co. v. Carmichael, 526 U.S. 137, 152 (1999)).

[107] FED. R. CIV. P. 26(g).

[108] *Id.* 26(f).

[109] *See Victor Stanley*, 250 F.R.D. at 262; *see also* William A. Gross Constr. Assocs., Inc., v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 136 (S.D.N.Y. 2009).

[110] *See Victor Stanley*, 250 F.R.D. at 262; *see also William A. Gross Constr.*, 256 F.R.D. at 136; *supra* Part II.B.1.

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

[112] If it is a medical malpractice case, an attorney must learn about medical procedures. If a case involves accounting issues, an attorney must learn how the accounting was done. If an attorney is hired for a construction case, he or she must learn construction techniques and agreements.

be understood if attorneys plan to effectively represent their clients. Moreover, litigators cannot continue to exempt themselves from information technology as discovery challenges continue to snowball—senior litigators can no longer assume that because a keyboard is attached, someone else, someone younger, will take care of it. The time to learn about technology is now. In the words of Master Yoda: "Try not. Do. Or do not. There is no try."[113] It is a matter of professional obligation.

[113] STAR WARS: EPISODE V—THE EMPIRE STRIKES BACK (Lucasfilm 1980).

Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review

Herbert L. Roitblat
Electronic Discovery Institute, OrcaTec LLC, PO Box 613, Ojai, CA 93024. E-mail: herb@orcatec.com

Anne Kershaw
Electronic Discovery Institute, A. Kershaw, P.C. Attorneys & Consultants, 303 South Broadway, Suite 430, Tarrytown, NY 10591. E-mail: anne.kershaw@akershaw.com

Patrick Oot
Electronic Discovery Institute, Verizon, 1320 North Courthouse Road, Arlington, VA 22201. E-mail: patrick.oot@verizon.com

In litigation in the US, the parties are obligated to produce to one another, when requested, those documents that are potentially relevant to issues and facts of the litigation (called "discovery"). As the volume of electronic documents continues to grow, the expense of dealing with this obligation threatens to surpass the amounts at issue and the time to identify these relevant documents can delay a case for months or years. The same holds true for government investigations and third-parties served with subpoenas. As a result, litigants are looking for ways to reduce the time and expense of discovery. One approach is to supplant or reduce the traditional means of having people, usually attorneys, read each document, with automated procedures that use information retrieval and machine categorization to identify the relevant documents. This study compared an original categorization, obtained as part of a response to a Department of Justice Request and produced by having one or more of 225 attorneys review each document with automated categorization systems provided by two legal service providers. The goal was to determine whether the automated systems could categorize documents at least as well as human reviewers could, thereby saving time and expense. The results support the idea that machine categorization is no less accurate at identifying relevant/responsive documents than employing a team of reviewers. Based on these results, it would appear that using machine categorization can be a reasonable substitute for human review.

Introduction

In litigation, particularly civil litigation in the US Federal Courts, the parties are required, when requested, to produce documents that are potentially relevant to the issues and facts of the matter. This is a part of the process called "discovery." When it involves electronic documents, or more formally, "electronically stored information (ESI)," it is called eDiscovery. The potentially relevant documents are said to be responsive.

The volume of electronically stored information that must be considered for relevance continues to grow and continues to present a challenge to the parties. The cost of eDiscovery can easily be in the millions of dollars. According to some commentators, these costs threaten to skew the justice system as the costs can easily exceed the amount at risk (Bace, 2007). Discovery is a major source of costs in litigation, sometimes accounting for as much as 25% of the total cost (e.g., Gruner, 2008). Overwhelmingly, the biggest single cost in eDiscovery is for attorney review time—the time spent considering whether each document is responsive (relevant) or not. Traditionally, each document or email was reviewed by an attorney who decided whether it was responsive or not. As the volume of material that needs to be considered continues to grow, it is becoming increasingly untenable to pursue that strategy.

Attorneys and their clients are looking for ways to minimize the cost of eDiscovery (Paul & Baron, 2007). One approach that holds promise for reducing costs while delivering appropriate results is the use of information retrieval tools.

Over the last several years, attorneys have come to rely increasingly on search tools, for example, Boolean queries,

Received June 5, 2009; revised July 29, 2009; accepted August 27, 2009

© 2009 ASIS&T • Published online 8 October 2009 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.21233

to limit the scope of what must be reviewed. The details of these queries may be negotiated between the parties. Here is an example of one such query in the case of U.S. v Philip Morris:

> (((master settlement agreement OR msa) AND NOT
> (medical savings account OR metropolitan standard area))
> OR s. 1415 OR
> (ets AND NOT educational testing service) OR
> (liggett AND NOT sharon a. liggett) OR atco OR lorillard
> OR
> (pmi AND NOT presidential management intern) OR pm usa
> OR rjr OR
> (b&w AND NOT photo* OR phillip morris OR batco OR ftc
> test method OR star scientific OR vector group OR joe
> camel OR
> (marlboro AND NOT upper marlboro)) AND NOT
> (tobacco* OR cigarette* OR smoking OR tar OR nicotine OR
> smokeless OR synar amendment OR philip morris OR r.j.
> reynolds OR
> ("brown and williamson") OR
> ("brown & williamson") OR bat industries OR liggett group)
> (Baron, 2008).

The information retrieval requirements of attorneys conducting eDiscovery are somewhat different from those in many information retrieval tasks. Document sets in eDiscovery tend to be very large with a large proportion of emails and a large number of requests that need to be translated into queries. The Philip Morris case, for example, involved over 1,726 requests from the tobacco companies and more than 32 million Clinton-era records that needed to be evaluated.

Information retrieval studies involving the World Wide Web, of course, have an even greater population of potentially relevant documents, but in those systems the user is usually interested in only a very tiny proportion of them, for example, between 1 and 50 documents out of billions. Getting the desired information within the first 10–50 results is generally the challenge in these studies.

Web searches are generally fairly specific, for example, "What are the best sites to visit in Paris?" In contrast, the information need in eDiscovery is generally much broader and more vague. Discovery requests include statements like "All documents constituting or reflecting discussions about unfair or discriminatory allocations of [Brand X] products or the fear of such unfair or discriminatory allocations." These requests will not typically be satisfied by one or a few documents.

Recall, the proportion of responsive documents actually retrieved, is arguably a more important measure of the success of information retrieval for the lawyers than is precision, the proportion of retrieved documents that are responsive. High precision will save the client money, because fewer documents will need to be reviewed. On the other hand, obviously low recall can lead to court sanctions, including an "adverse inference" instruction, where a jury is instructed that they may construe that the missing information was contrary to the interests of the party that failed to produce it. Obviously, low precision can also lead to accusations that the producing party is doing an inadequate job identifying responsive documents, but these sanctions are usually much less onerous than those for failing to produce.

This study is an investigation of methods that may be useful to reduce the expense and time needed to conduct electronic discovery. In addition to search techniques, these methods can include machine learning and other data mining techniques. In the present study, the categorization tools provided by two companies who are active eDiscovery service providers were used to categorize responsive documents. These providers' systems were taken to be representative of a broad range of similar systems that are available to litigators. The performance of these two systems was compared to the performance of a more traditional methodology—having attorneys read and categorize each document in the context of a substantial eDiscovery project.

Background: Related Work

Blair and Maron (1985) conducted one of the early studies on using computers to identify potentially responsive documents. They analyzed the search performance of attorneys working with experienced search professionals to find documents relevant to a case in which a computerized San Francisco Bay Area Rapid Transit (BART) train failed to stop at the end of the line. The case involved a collection which, at the time, seemed rather large, consisting of about 40,000 documents. Current cases often involve one to two orders of magnitude more documents.

Blair and Maron found that the attorney teams were relatively ineffective at using the search system to find responsive documents. Although they thought that their searches had retrieved 75% or more of the responsive documents, they had, in fact, found about 20% of them.

One reason for this difficulty is the variety of language used by the parties in the case. The parties on the BART side referred to "the unfortunate incident," but parties on the victim's side called it an "accident" or a "disaster." Some documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were sometimes left out. The limitation in this study was not the ability of the computer to find documents that met the attorneys' search criteria, but the inability to anticipate all of the possible ways that people could refer to the issues in the case.

Blair and Maron concluded that "It is impossibly difficult for users to predict the exact words, word combinations, and phrases that are used by all (or most) relevant documents and only (or primarily) by those documents" (p. 295). They advocated for the use of manually applied index terms, meaning that someone would have to read the documents, determine what they were about, and categorize them.

TREC (Text Retrieval Conference) is a multitrack project sponsored by the National Institute for Standards and Technology and others to conduct comparative research on text retrieval technologies. Since 2006 (Baron, Lewis, & Oard, 2007; Tomlinson, Oard, Baron, & Thompson, 2008, Oard, Hedin, Tomlinson, & Baron, 2009), TREC has included a

DOI: 10.1002/asi

legal track whose goal is to assess the ability of information retrieval technology to "meet the needs of the legal community for tools to help with retrieval of business records." In support of this goal, they seek to develop and apply collections and tasks that approximate the data, methods, and issues that real attorneys might use during civil litigation and to apply objective criteria by which to judge the success of various search methodologies. In 2008 (Oard et al., 2009), 15 research teams participated in at least one of the three types of task (ad hoc query, relevance feedback, and interactive search).

The searches were conducted against a collection (also used in 2006 and 2007) of tobacco-related documents released under the Tobacco Master Settlement Agreement (MSA) called the IIT Complex Document Information Processing Test Collection (CDIP) v. 1.0. The collection consists of 6,910,192 document records in the form of XML elements. Most of these documents were encoded from images using optical character recognition (OCR). Relying on OCR data for text presents its own challenges to these studies, because of the less than perfect accuracy of the process used to derive the text from the documents.

The performance of the various teams on each task was measured by having a pool of volunteer assessors evaluate a sample of documents for relevance. The assessors for the 2008 session were primarily second- and third-year law students, with a few recent law school graduates, experienced paralegals, and other litigation specialists. Each assessor was asked to evaluate a minimum of 500 documents. On average, an assessor managed about 21.5 documents per hour, so a block of 500 documents entailed a substantial level of effort from the volunteer assessors.

In the TREC ad hoc task, the highest recall achieved was 0.555 (i.e., 55.5% of the documents identified as relevant were retrieved; Run "wat7fuse"). The precision corresponding to that level of recall was 0.210, meaning that 21% of the retrieved documents were determined to be relevant.

The TREC interactive task allowed each team to interact with a topic authority and revise their queries based on this feedback. Each team was allowed 10 hours of access to the authority. The interactive task also allowed the teams to appeal reviewer decisions if they thought that the reviewers had made a mistake. Of the 13,339 documents that were assessed for the interactive task, 966 were appealed to the topic authority. This authority played the role, for example, of the senior litigator on the case, with the ultimate authority to overturn the decisions of the volunteer assessors. In about 80% of these appeals the topic authority supported the appeal and recategorized the document. In one case (Topic 103), the appeal allowed the team with the already highest recall rate to improve its recall by 47%, ending up with recall of 0.624 and precision of 0.810.

Some of the more interesting findings from the 2008 TREC legal track concern the levels of agreement seen between assessors. Some of the same topics were used in previous years of the TREC legal track, so it is possible to compare the judgments made during the current year with those made in previous years. For example, the level of agreement between assessors in the 2008 project and those from 2006 and 2007 were reported. Ten documents from each of the repeated topics that were previously judged to be relevant and 10 that were previously judged to be nonrelevant were assessed by the 2008 assessors. It turns out that "just 58% of previously judged relevant documents were judged relevant again this year." Conversely, "18% of previously judged non-relevant documents were judged relevant this year." Overall, the 2008 assessors agreed with the previous assessors 71.3% of the time.

Unfortunately, this is a fairly small sample, but it is consistent with other studies of inter-reviewer agreement. In 2006 the TREC coordinators gave a sample of 25 relevant and 25 nonrelevant documents from each topic to a second assessor and measured the agreement (http://cio.nist.gov/esd/emaildir/lists/ireval/msg00012.html, retrieved 23 July, 2009) between these two. Here they found about 76% agreement. Other studies outside of TREC Legal have found similar levels of (dis)agreement (e.g., Barnett, Godjevac, Renders, Privault, Schneider, & Wickstrom, 2009; Borko, 1964; Tonta, 1991; Voorhees, 1998).

Research Design: Methods

Research Questions

One solution to the problem of the exploding cost of eDiscovery is to use technology to reduce the effort required to identify responsive and privileged documents. Like the TREC legal track, the goal of the present research is to evaluate the ability of information retrieval technology to meet the needs of the legal community for tools to identify the responsive documents in a collection.

From a legal perspective, there is recognition that the processes used in discovery do not have to be absolutely perfect, but should be reasonable and not unduly burdensome (e.g., Rule 26(g) of the Federal Rules of Civil Procedure). The present study is intended to investigate whether the use of technology is reasonable in this sense.

The notion of "reasonable" is itself subject to interpretation. We have taken the approach that the current common practice of having trained reviewers examine each document does a reasonable job of identifying responsive documents, but at an often unreasonable cost. If information retrieval systems can be used to achieve the same level of performance as the current standard practice, then they too should be considered reasonable by this standard. Formally, the present study is intended to examine the hypothesis: *The rate of agreement between two independent reviewers of the same documents will be equal to or less than the agreement between a computer-aided system and the original review.*

Participants

The participants in this study were the original review teams, two re-review teams, and two electronic discovery service providers. The original review was conducted by two teams of attorneys, one focused on review for privilege, and one focused on review for relevance. A total of 225 attorneys participated in this initial review. The original purpose of this review was to meet the requirements of a US Department of Justice investigation of the acquisition of MCI by Verizon. It was not initially designed as a research study, but Verizon has made the outcome of this review available in support of the present study. For more details, see the Dataset section, below.

The two re-review teams were employees of a service provider specializing in conducting legal reviews of this sort. Each team consisted of five reviewers who were experienced in the subject matter of this collection. The two teams of re-reviewers (Team A and Team B) both reviewed the same 5,000 documents in preparation for one of the processes of one of the two service providers. Hence, there is a caveat that the decisions made by the service provider are not completely independent of the decisions made by the re-review teams. This issue will be discussed further in the Discussion section.

The two service providers volunteered their time, facilities, and processes to analyze the data. The two companies, one based in California and the other in Texas, each independently analyzed the data without knowledge of the original decisions made or of the decisions made by the other provider. Their systems are designated System C and System D. The identity of the two systems, that is, which company's is System C and which is System D, was determined by a coin flip in order to conceal the identity of the system yielding specific data. We did not cast this task as a competition between the two systems and do not wish to draw distinctions between them. Rather, we see these two systems as representative of a general analytic approach to information retrieval in electronic discovery.

Task

The task of the original review was to determine whether each document was responsive to the request of the Justice Department. The reviewers also made decisions about the privilege status of the documents, but these judgments were not used in the present study.

The task of the two systems was to replicate the classification of documents into the two categories of responsive and nonresponsive.

Dataset

The documents used in the present study were collected in response to a "Second Request" concerning Verizon's acquisition of MCI. Certain corporate mergers and acquisitions are subject to premerger antitrust review by the U.S. Department of Justice and the U.S. Federal Trade Commission under the Hart-Scott-Rodino Antitrust Improvements Act of 1976. The government may issue requests for additional information for material relevant to the proposed acquisition. These requests are commonly called "Second Requests."

The documents were collected from 83 employees in 10 US states. Together they consisted of 1.3 terabytes of electronic files in the form of 2,319,346 documents. The collection consisted of about 1.5 million email messages, 300,000 loose files, and 600,000 scanned documents. After eliminating duplicates, 1,600,047 items were submitted for review. The attorneys spent about 4 months, 7 days a week, and 16 hours per day on the review at a total cost of $13,598,872.61 or about $8.50 per document. After review, a total of 176,440 items were produced to the Justice Department.

During the course of managing the Second Request, these documents were put in a database, containing 2,224,023 records. From this database, 1,813,612 documents could be usably extracted for this study.

Analysis

This study compares two computer-assisted categorization processes with the more traditional process of having people, usually lawyers, read and categorize each document. This study uses agreement to assess the level of reliability of the human and computer processes.

A sample of 5,000 documents from the original review was randomly chosen for a second review $\sim$2 years after complying with the Second Request. Two teams, A and B, performed this re-review without explicit knowledge of the original classifications. Both teams reviewed the complete sample.

The goal of the analysis is to compare the level of agreement between the two re-review teams and the two computer systems, on the one hand, with the original review on the other. The level of agreement between the two re-review teams and the original review is an indication of the level of agreement one could expect from the traditional process. To the extent that the computer systems show similar levels of agreement, they could be thought of as a reasonable substitute for the traditional process.

Results

The original review identified 9.46% (SE $=$ 0.0002) of the documents in the entire collection and 9.8% (SE $=$ 0.0042) of the documents in the sample of 5,000 as responsive. The difference between these two proportions was not significant ($t = 0.719$, $p > 0.05$), supporting the notion that the sample was representative of the complete population.

The contingency tables showing how the different systems and reviewers classified documents are shown in Table 1. Each row represents the contingent relation shown in the first column. RR indicates the number of documents that were classified as responsive by both groups or systems. RN indicates the number of documents that were classified as responsive by the first group and as nonresponsive by the second group. NR indicates the documents that were classified as nonresponsive by the first group or system and as

DOI: 10.1002/asi

TABLE 1. The contingent frequencies between the designated sets of judgments.

	Contingent relation			
	RR	RN	NR	NN
1. Original vs. Team A	238	250	971	3,541
2. Original vs. Team B	263	225	1,175	3,337
3. Team A vs. Team B	580	629	858	2,933
4. Original vs. Teams A & B Nonadjudicated	349	139	1,718	2,794
5. Original vs. Teams A & B Adjudicated	216	272	739	3,773
6. Original vs. System C	78,617	92,908	211,403	1,430,684
7. Original vs. System C	90,416	81,109	216,359	1,425,728

Note. RR = Responsive/Responsive, RN = Responsive Nonresponsive, NR = Nonresponsive/Responsive, NN = Nonresponsive/Nonresponsive.

responsive by the second. NN indicates the documents that were classified as nonresponsive by both groups or systems.

Human Review

The contingency tables resulting from each of the two teams, compared with the original classifications, are shown in the first two rows of Table 1. Both contingency tables were significantly different from chance (independence) (Team A: $\chi^2 = 178.37$, Team B: $\chi^2 = 166.73$, both $p < 0.01$).

Row 3 of Table 1 shows the contingency table comparing Team B's classifications with those from Team A. The decisions made by the two teams were strongly related ($\chi^2 = 287.31$, $p < 0.01$).

The 1,487 documents on which Teams A and B disagreed were submitted to a senior Verizon litigator (P. Oot), who adjudicated between the two teams, again without knowledge of the specific decisions made about each document during the first review. This reviewer had knowledge of the specifics of the matter under review, but had not participated in the original review. This authoritative reviewer was charged with determining which of the two teams had made the correct decision. Row 4 of Table 1 contains the contingency table comparing the nonadjudicated decisions to the original classification and Row 5 contains the contingency table comparing the adjudicated decisions to the original classification. The adjudicated decisions, like those made independently by the two teams, were strongly related ($\chi^2 = 203.07$, $p < 0.01$) to the original review.

Team A identified 24.2% (SE = 0.006) and Team B identified 28.76% (SE = 0.006) of the sample as responsive. The difference between these two proportions was significant ($t = 5.20$, $p < 0.01$). After adjudication, the combined teams identified 955 or 19.1% (SE = 0.006) as responsive. Adjudication, in other words, reduced the overall number of documents that the new reviewers designated as responsive. Of the 1,487 documents on which Team A and Team B disagreed, the senior litigator chose Team A's classification on 796 documents, Team B's classification on 691 documents.

Team A agreed with the original review on 75.58% (SE = 0.006) of the documents and Team B agreed with the original review on 72.00% (SE = 0.006), both before adjudication. Team A agreed with Team B on 70.26% (SE = 0.006) of the documents. The adjudicated review agreed with the original classification on 79.8% (SE = 0.006) of the documents. Team A agreed with the original significantly more often ($t = 4.07$, $p < 0.01$) than did Team B. Because the adjudicated results included most of the decisions from Team A and Team B, it is not clear how to assess the difference in agreement between the adjudicated and nonadjudicated reviews—they are not independent.

Of the 488 documents in the sample identified as responsive by the original review team, Team A identified 238 or 48.78% (SE = 0.023) as responsive. Team B identified 263 or 53.89% (SE = 0.023) as responsive. Together, teams A and B identified as responsive 349 or 71.52% (SE = 0.02) of the documents classified as responsive by the original review. Conversely, of the 2067 documents identified as responsive by either Team A or Team B, the original review identified 349 or 16.88% (SE = 0.008) as responsive.

Of the 4,512 documents in the sample that were designated nonresponsive during the original review, Team A identified 971 or 21.52% (SE = 0.006) as responsive and Team B recognized 1,175 or 26.04% (SE = 0.007) as responsive. Together, Teams A and B recognized 1,718 or 38.07% (SE = 0.007) of these as responsive (before adjudication, i.e., if either team called it responsive, a document was counted for this purpose as responsive). After adjudication, the two teams combined recognized 739 or 16.38% (SE = 0.006) of the original review's nonresponsive documents as responsive.

Computer-Assisted Review and Comparison

In addition to the two review teams reexamining a sample of documents from the original review, two commercial electronic discovery systems were also used to classify documents as responsive vs. nonresponsive. One of these systems based its classifications in part on the adjudicated results of Teams A and B, but without any knowledge of how those teams' decisions were related to the decisions made by original review team. As a result, it is not reasonable to compare the classifications of these two systems to the classifications of the two re-review teams, but it is reasonable to compare them to the classifications of the original review.

The contingency table resulting from each of the two systems is shown in Rows 6 and 7 of Table 1.

DOI: 10.1002/asi

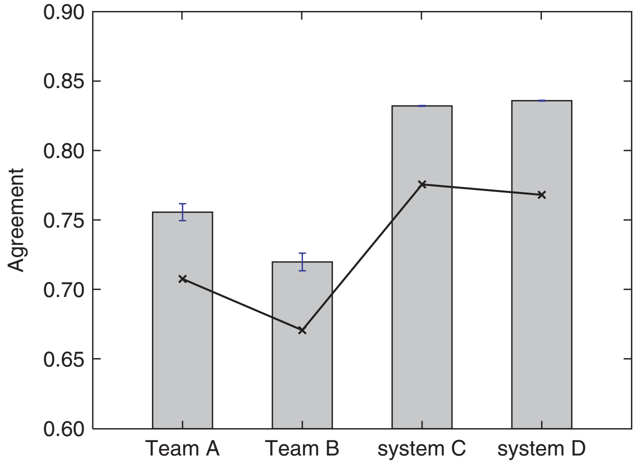

FIG. 1. The level of agreement with the original review and chance levels to be expected from the marginals for the two human teams and the two computer systems (the four reassessments). Error bars show standard error.

System C classified 15.99% (SE = 0.0003) of the documents and System D classified 16.92% (SE = 0.0003) of the documents as responsive, which were both higher than the proportion identified as responsive by the original team ($t = 187.6$, $p < 0.01$ and $t = 211.2$, $p < 0.01$, respectively). System C agreed with the original classification on 83.2% (SE = 0.00028) and System D agreed with the original classification on 83.6% (SE = 0.00028) of the documents.

Of the 171,525 documents identified as responsive by the original review team, System C identified 78,617 or 45.8% (SE = 0.001) as responsive. System D identified 90,416 or 52.7%% (SE = 0.001) as responsive. Together, Systems C and D identified as responsive (i.e., either C or D responsive), 123,750 or 72.1% (SE = 0.001) of the documents classified as responsive by the original review. Conversely, of the 493,004 documents identified as responsive by either System C or System D, the original review identified 123,750 or 25.1% (SE = 0.001) as responsive.

The percentage agreements between each of the two teams and each of the two systems and the original review are shown in Figure 1. The percentage agreements for each of the assessments shown in Figure 1 was significantly different from each other's assessment (A vs. B: $t = 4.07$, A vs. C: $t = 12.56$, A vs. D: 136.7, B vs. C: 17.65, B vs. D: 130.8, C vs. D: 2139.2, all $p < 0.01$). In addition, each assessment was significantly different from chance ($\chi^2 = 178.37$, 166.73, 125588.00, 172739.91, for A, B, C, and D, respectively, all $p < 0.01$).

Figure 2 breaks down overall agreement into positive agreement and negative agreement, proportions of specific agreement (Spitzer & Fleiss, 1974). When the base rates of the different categories are widely different, simple agreement is subject to chance-related bias. Positive and negative agreement remove that bias and allow one to look at each of these categories separately. Chance should affect only the more frequent category, in this case, the nonresponsive documents.

On positive agreement, assessments A and B did not differ significantly ($t = 0.702$, $p > 0.05$), but each of the other assessments differed from one another (t: A vs. C: 8.02, A vs. D: 50.10, B vs. C: 9.12, B vs. D: 50.79, C vs. D: 600.22, all $p < 0.01$). A similar pattern was seen for negative agreement. Assessments A and B did not differ significantly from one another ($p > 0.05$), but the comparisons did show significant differences in the degree to which they agreed with original review ($p < 0.01$) (t: A vs. B: 1.44, A vs. C: 3.89, A vs. D: 68.77, B vs. C: 6.00, B vs. D: 69.49, and C vs. D 905.68).

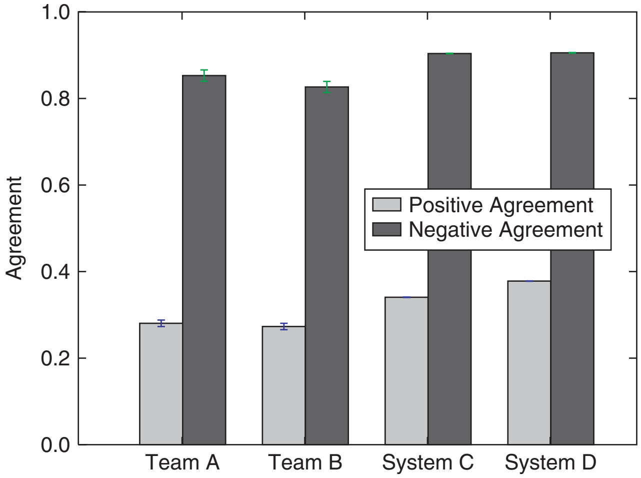

FIG. 2. Positive agreement (2*RR/(2*RR + RN + NR)) and negative agreement (2*NN/(2*NN + NR + RN)) for agreement between the original review and the four reassessments. NN, NR, etc. refer to the columns of Table 1. Error bars are standard error.

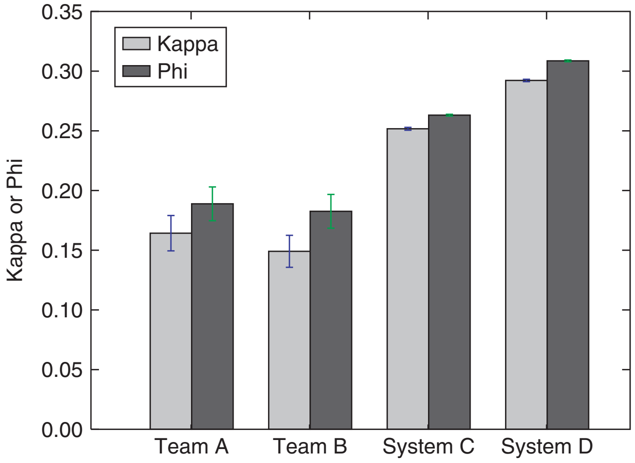

FIG. 3. Kappa and Phi for the agreement between the original review and the four reassessments. Kappa and Phi are "chance adjusted" measures of association or agreement. Error bars are standard error.

Another approach to characterizing the relationship between the latter assessments and the earlier reviews is to use "chance-corrected" measures of agreement. Figure 3 shows Cohen's kappa and phi, two measures that take into account the extent to which we might expect the assessments to agree based on chance. Cohen's kappa essentially subtracts out the level of agreement that one would expect by chance. Kappa is 1.0 if the two raters agree perfectly and is 0 if they agree exactly as often as expected by chance. Kappa less than 0 can be obtained if the raters agree less often than is expected by chance. Phi is derived from chi-squared and measures

DOI: 10.1002/asi

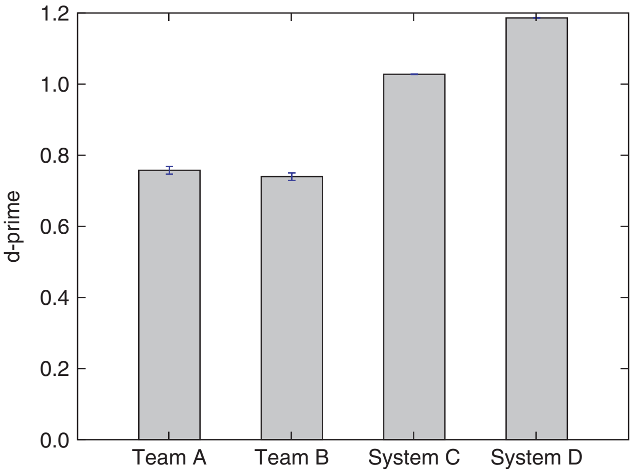

FIG. 4. The signal detection measure d′ comparing each of the re-reviews against the original review.

the deviation from the chance expectation. It has the value 0 only when there is complete independence between the two assessments. The pattern of results for both of these measures is the same as for agreement and for positive and negative agreement.

As with positive and negative agreement, Teams A and B did not differ significantly for either kappa ($t = 0.76$, $p > 0.05$) or phi ($t = 0.31$, $p > 0.05$). The other assessments did differ significantly from one another on kappa ($p < 0.01$) (t: A vs. C: 5.89, A vs. D: 8.62, B vs. C: 7.63, B vs. D: 10.65, C vs. D: 25.91) and on phi (t: A vs. C: 5.24, A vs. D: 8.45, B vs. C: 5.69, B vs. D: 8.90, C vs. D: 43.30). In addition to the data shown in Figure 3, we can also compute the corresponding measures comparing the decisions made by Team A with those made by Team B (kappa: 0.238, phi: 0.240).

The difference between the proportions identified as responsive by the original review and the re-reviews may indicate a difference in bias. Bias simply refers to an overall tendency to select one category over another, independent of the information in the documents. For example, one attorney might believe that it is more important to avoid missing a responsive document than another attorney does and so be more willing to classify documents as responsive. Recall increases and precision decreases when an assessor increases their willingness to call a document responsive; thus, these measures make it difficult to separate the discriminability of the classes from the bias. Signal detection theory (van Rijsbergen, 1979; Swets, 1969), on the other hand, offers a measure, d′, that is independent of bias. The more a system (or person) can separate two classes, the higher its d′ score will be. The value of d′ ranges from 0 when the responsive and nonresponsive documents are completely indistinguishable by the system to positive infinity when there is no overlap between the two.

With large numbers of trials (in our case documents), the binomial distribution is closely approximated by the normal distribution, so the use of the measure d′ is justified. Figure 4 shows the sensitivity measure, d′ for each of the four re-reviews.

TABLE 2. Standard information retrieval measures.

	Precision	Recall	F_1
Human Team A	0.196857	0.487705	0.280495
Human Team B	0.182893	0.538934	0.273105
System C	0.271074	0.458341	0.340669
System D	0.294731	0.52713	0.378072

The d′ values for Teams A and B did not differ significantly ($t = 1.19$, $p = > 0.05$). The other assessments did differ significantly from one another (t: A vs. C: 25.14, A vs. D: 39.85, B vs. C: 27.44, B vs. D: 42.51, C vs. D: 135.94). By comparison, the adjudicated reviews combining Team A and Team B judgments with that of a senior attorney showed a d′ of 0.835.

The use of precision and recall implies the availability of a stable ground truth against which to compare the assessments. Given the known variability of human judgments, we do not believe that we have a solid enough foundation to claim that we know which documents are truly relevant and which are not. Nevertheless, in the interest of comparison with existing studies (e.g., TREC Legal 2008), Table 2 shows the computed precision and recall of each of the four assessments using the original review as its baseline. F_1 is a summary measure combining precision and recall. It is calculated according to the formula used in TREC Legal 2008:

$$F_1 = \frac{2Pr \times R}{Pr + R}$$

where Pr = precision and R = recall.

These scores are comparable to those obtained in TREC Legal 2008. In that study, the median precision was 0.27 and median recall was 0.21.

Discussion

This study is an experimental investigation of how well computer-aided systems can do relative to traditional human review. It is an elaboration and extension of the kind of research done under the auspices of the TREC Legal Track. Both projects are concerned with identifying processes and methods that can help the legal community to meet its discovery obligations.

Although the volume of information that must be processed during litigation continues to grow, the legal profession's means for dealing with that information is on the verge of collapse. The same techniques that worked 20 years ago, when electronically stored information was relatively rare, do not continue to provide adequate or cost-effective results today, when electronic discovery matters can extend to many terabytes of data.

According to the Federal Rules of Civil Procedure (Rule 26(g)), each party must certify at the end of the discovery process that their production has been complete and accurate after a reasonable enquiry. There can be disagreement about what constitutes a reasonable enquiry, but it would seem that, all other things being equal, one that does as well as traditional practice would be likely to be considered reasonable.

Accuracy and Agreement

In the ideal case, we would like to know how accurate each classification is. Ultimately, measurement of accuracy implies that we have some reliable ground truth or gold standard against which to compare the classifier, but such a standard is generally lacking for measures of information retrieval in general and for legal discovery in particular. In place of a perfect standard, it is common to use an exhaustive set of judgments done by an expert set of reviewers as the standard (e.g., as is the practice in the TREC studies).

Under these circumstances, agreement with the standard is used as the best available measure of accuracy, but its acceptance should be tempered with the knowledge that this standard is not perfect.

Variability of Human Relevance Judgments

The level of agreement among human reviewers is not strikingly high. The two re-review teams agreed with the original review on about 76% and 72% of the documents. They agreed with one another on about 70% of the documents with corresponding kappa values in the low to fair range. Although low, these levels are realistic. They are comparable to those observed in the TREC studies and other studies of interrater agreement (e.g., Barnett et al., 2009, Borko, 1964; van Rijsbergen, 1979; Tonta, 1991; Voorhees, 1998).

There are two sources of this variability. Some variability is due to random factors, that is, factors that are unrelated to the material being judged or to any stable trait of the judges. For example, reviewers' attention may wander, they may be distracted, or fatigued. A document that they might have categorized as responsive when they were more attentive might then be categorized as nonresponsive or vice versa.

The second source of variability is systematic, which is due to the interaction between the content of the documents and stable properties of the reviewers, and to individual differences among reviewers.

Relevance judgments may be strategic. Reviewers may have different goals in mind when assessing documents and these goals may vary over time. Differences in strategic judgment may affect how likely two individuals are to call a certain document responsive. As noted by the TREC Legal 2008 Topic Authorities (http://trec-legal.umiacs.umd.edu/TAreflections2008.doc, retrieved May 7, 2009):

> While the ultimate determination of responsiveness (and whether or not to produce a given document) is a binary decision, the breadth or narrowness with which "responsiveness" is defined is often dependent on numerous subjective determinations involving, among other things, the nature of the risk posed by production, the party requesting the information, the willingness of the producing party to face a challenge for underproduction, and the level of knowledge that the producing party has about the matter at a particular point in time. Lawyers can and do draw these lines differently for different types of opponents, on different matters, and at different times on the same matter. This makes it exceedingly difficult to establish a "gold standard" against which to measure relevance/responsiveness and explains why document review cannot be completely automated.

Instead of "subjective," it may be more appropriate to say that discovery involves judgment about the situation as well as about the documents and their contents. Some judgments bias the reviewer to be more inclusive and some bias the reviewer to be less inclusive, but these judgments are not made willy-nilly. As opposed to pure errors, which are random, these judgment calls are based on a systematic interpretation of the evidence and the situation. To the extent that judgments are systematically related to the content of the documents, even if biased, they are capable of being mirrored by some automated system. The classifications made by an automated system can easily include the bias judgments of the attorneys managing a case, being either more or less inclusive as the situation warrants. Bias is not a barrier to automation, despite the implication drawn by the TREC Legal Topic Authorities.

Nevertheless, bias can change from case to case and individual to individual. It is not a stable property of the methods used to categorize the documents, so it is helpful to distinguish the power of the method from the bias to be more or less inclusive. Signal detection theory, by separating bias from discriminability, allows us to recognize the role of the information in the document contents and the sensitivity of the method. The d' values observed in the study showed that the human reviewers were no better at distinguishing responsive from nonresponsive documents than were the two automated systems.

Discovery cannot be wholly automated, not for the reason that it involves so-called subjective judgment, but because ultimately attorneys and parties in the case have to know what the data are about. They have to formulate and respond to arguments and develop a strategy for winning the case. They have to understand the evidence that they have available and be able to refute contrary evidence. All of this takes knowledge of the case, the law, and much more.

When judgments are made by review teams, they necessarily add to the variability of these judgments. Of the 225 attorneys conducting the review, few if any of them had much detailed knowledge of the business issues being considered, the case strategy, or the relative consequences of producing more or fewer documents before embarking on their review. There were certainly individual differences among them. Some of them were almost certainly better able to distinguish responsive from nonresponsive documents. And, moreover, the long arduous hours spent reviewing documents almost certainly resulted in fatigue and inattention. All of this variability does not lead to the creation of a very solid standard against which to compare other approaches to review. On the other side, the procedure of using many attorneys to conduct a review is current practice in large cases, so these results represent a realistic if not particularly reliable standard.

Anything that reduces this variability is likely to improve the level of agreement. One reason that recall rates are so low in the TREC Legal studies (and in the present study) is because of nonsystematic variability in the judgments that are being used as the ground truth. Reducing that variability, as the TREC Interactive Task did, improved recall by as much as 47% (Topic 103, H5). Similar factors are undoubtedly operating in this study. Adjudication, for example, improved the agreement between the combined judgments of Teams A and B with the original review. These differences again show the effect of bias. Teams A and B classified more documents as responsive than appeared in the adjudicated results. Using TREC methodology, this difference would show up as a decline in recall and an increase in precision with adjudication. Both the original review and the two human re-reviews reflected variable judgments.

Conversely, when we reduce the variability of one of the categorizers, in this case by using computer software to implement the judgments, then it may be possible to improve the measured level of agreement, even when compared to a variable standard. A given person may make different decisions about the same text at different times, while computer classifiers generally make consistent judgments. Comparing the decisions made by two variable processes is likely to lead to lower observed levels of agreement than would comparing a variable process to an invariant one. If the computer does not contribute its own variability to the agreement measure, then higher levels of agreement may be observed.

Effects of Base Rate

Because of the difference in base rates of responsive and nonresponsive documents, we used several measures to reduce the influence of simple chance on our measures. If high levels of agreement or accuracy were achieved simply because of base-rate differences, then separating the measures into positive and negative agreement would eliminate these differences. Even when eliminating differences in base rate by comparing within category, positive and negative agreement both show the same pattern of results.

As another approach to assessing agreement independent of base-rate differences, two chance-corrected measures, kappa and phi, were also used. Systems C and D showed at least as high a level of agreement on these measures as was found using Team A and Team B.

Blair and Maron Revisited

Blair and Maron (1985) found that their attorneys were able to find only about 20% of the responsive documents. They concluded that it was impossibly difficult to guess the right words to search for and instead advocated for using human indexers to develop a controlled vocabulary. Collections that seemed large to Blair and Maron, however, are dwarfed by the size of the present collection and many collections typical of modern electronic discovery. Employing human reviewers to manually categorize the documents can cost millions of dollars, an expense that litigants would prefer to reduce if possible.

Blair and Maron argued for using human readers to assign documents to specific categories because, they concluded, guessing the right terms to search for was too difficult to be practical. In contrast, with the size of modern collections, lawyers are finding that human categorization is too expensive to be practical.

The categorization systems used in the present study, and many others in current use, are more elaborate than the search system used by Blair and Maron. They employ more information about the documents and the collection as well as information from outside the collection (such as an ontology or the results of human classification). Many of these elaborations are designed to overcome the problem of guessing query terms.

Our best estimates from the present study suggest that both human review teams and computer systems identified a higher percentage of responsive documents than Blair and Maron's participants did. It is interesting to note that the human reviewers of Teams A and B were not more successful than the computer systems were at identifying responsive documents. One limitation may be the variability of the human judgments against which the computer systems are being compared.

Comparison With TREC Legal

The results of this study are generally congruent with those produced by TREC Legal. The methodology used in the present study has some advantages and some disadvantages relative to that used by TREC, but the differences typically are more indicative of the difficulty of doing this kind of research than of any flaw in design. They are predominantly responses to constraints, not errors.

By its charter, TREC is required to use publicly available datasets. Realistic litigation data, in contrast, are typically highly confidential and difficult to obtain for research purposes. For its first 3 years of investigations, TREC concentrated on a large collection of tobacco-related documents that were released as part of a legal settlement. These documents were mostly converted into electronic text using optical character recognition (OCR), which introduces errors. Because the documents in the collection were produced as part of a case, many of the irrelevant nonresponsive documents that are typical of actual electronic discovery collections were eliminated. Every document was deemed responsive to something. The TREC Legal designers have compensated for this by inventing issues/topics that might have been litigated. Their performance measures are based on sampling.

The present study, in contrast, used a real matter based on a Department of Justice request for information about a merger. Therefore, the responsiveness categorization is more naturalistic. It would be preferable, perhaps, if the matter were a litigation rather than a DOJ request, but these are the data that were made available. On the other hand, these data have not been made publicly available. Although some documents (600,000 out of 2.3 million) were scanned and OCRed, the majority were native electronic documents. Rather than sampling, the original collection was exhaustively reviewed at substantial expense in the context of a legal matter without any plans, at the time, for conducting a study. It would be very difficult to replicate this exhaustive review as part of a research project.

The reviews in the present study were performed by attorneys; in the TREC Legal studies the reviewers were predominantly law students. In the present study the reviewers spent hundreds of hours reviewing documents under some time pressure; in the TREC Legal study each reviewer spent about 21 hours reviewing documents at their own pace.

Another difference between the present study and the TREC Legal study is the use of documents that are more typical of modern electronic discovery situations than were many of the Tobacco documents. A majority of the documents in the present study (1.5 million) consisted of emails. The Tobacco collection contains a smaller proportion of emails, consisting rather of internal memos and other documents (Eichman & Chin, 2007).

In TREC Legal, many of the human assessor–assessor relations were computed on relatively small numbers of documents and typically involved equal numbers of responsive or nonresponsive documents. Decision bias is known to be affected by the proportion of positive events (e.g., Green & Swets, 1966). In contrast, the present study used naturalistic distributions of responsive and nonresponsive documents and larger sample sizes for the comparison of assessor–assessor relations. Still, both studies found similar levels of agreement.

Finally, the present study used two commercial electronic discovery service providers, whereas TREC is open to anyone who wants to contribute. These providers volunteered their processing time and effort to categorize the data. Although a few active service providers contributed to the TREC results, most of the contributors were academic institutions, so it is difficult to generalize from the overall performance of the TREC Legal participants to what one might expect in electronic discovery practice. Academic groups might be either more or less successful than commercial electronic discovery organizations.

The results from each service provider in the present study are displayed anonymously. These volunteers were intended to be representative of the many that are available. With the large number of documents involved, any slight difference between them is likely to be statistically significant, but small differences are not likely to be meaningful or replicable. The goal was to determine whether these tools could provide results comparable to those obtained through a complete manual review, and in that they have succeeded.

Conclusion

This study is an empirical assessment of two methods for identifying responsive documents. It set out to answer the question of whether there was a benefit to engaging a traditional human review or whether computer systems could be relied on to produce comparable results.

On every measure, the performance of the two computer systems was at least as accurate (measured against the original review) as that of a human re-review. Redoing the same review with more traditional methods as was done during the re-review had no discernible benefit.

There may be other factors at play in determining legal reasonableness, but all other things being equal, it would appear that employing a system like one of the two systems employed in this task will yield results that are comparable to the traditional practice in discovery and would therefore appear to be reasonable.

The use of the kind of processes employed by the two systems in the present study can help attorneys to meet the requirements of Rule 1 of the Federal Rules of Civil Procedure: “to secure the just, speedy, and inexpensive determination of every action and proceeding.”

References

Bace, J. (2007). Cost of e-discovery threatens to skew justice system. Gartner Report G00148170. Retrieved May 6, 2009, from http://www.akershaw.com/Documents/cost_of_ediscovery_threatens_148170.pdf

Baron, J.R. (2008). Beyond keywords: Emerging best practices in the area of search and information retrieval. New Mexico Digital Preservation Conference, June 5, 2008. Retrieved July 29, 2009, from http://www.archives.gov/rocky-mountain/records-mgmt/conferences/digital-preservation/beyond-keywords.pdf

Baron, J.R., Lewis, D.D., & Oard, D.W. (2007). TREC-2006 Legal Track Overview. In Proceedings of the 15th Text REtrieval Conference (TREC 2006) (pp. 79–99). Gaithersburg, MD: NIST. Retrieved September 21, 2009, from http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf

Barnett, T., Godjevac, S., Renders, J.-M., Privault, C., Schneider, J., & Wickstrom, R. (2009, June). Machine learning classification for document review. Paper presented at the ICAIL 2009 Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery. Retrieved July 24, 2009, from http://www.law.pitt.edu/DESI3_Workshop/Papers/DESI_III.Xerox_Barnett.Xerox.pdf

Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM, 28, 289–299.

Borko, H. (1964) Measuring the reliability of subject classification by men and machines. American Documentation, 15(4), 268–273.

Eichman, D., & Chin, S.-C. (2007, June). Concepts, semantics and syntax in e-Discovery. Paper presented at the ICAIL 2007 Workshop on Supporting Search and Sensemaking for Electronically Stored Information in Discovery. Retrieved July 24, 2009, from http://www.umiacs.umd.edu/~oard/desi-ws/papers/eichmann.pdf

Green, D.M., & Swets, J.A. (1966). Signal detection theory and psychophysics. New York: John Wiley & Sons.

Gruner, R.H. (2008). Anatomy of a lawsuit. Retrieved July 24, 2009, from http://www.vallexfund.com/download/AnatomyLawsuit.pdf

Oard, D.W., Hedin, B., Tomlinson, S., & Baron, J.R. (2009). Overview of the TREC 2008 Legal Track. In Proceedings of the 17th Text Retrieval Conference (TREC 2008). Retrieved September 21, 2009, from http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf

Paul. G.L., & Baron, J.R. (2007). Information inflation: Can the legal system adapt? Richmond Journal of Law and Technology, 13, Article 10, 1–41. Retrieved July 28, 2009, from http://law.richmond.edu/jolt/v13i3/article10.pdf

DOI: 10.1002/asi

Rijsbergen, C.J. van (1979). Information retrieval, 2nd ed. London: Butterworths.

Swets, J.A. (1969). Effectiveness of information retrieval methods. American Documentation, 20(1), 72–89.

Tomlinson, S., Oard, D.W., Baron, J.R., & Thompson P. (2008). Overview of the TREC 2007 Legal Track. In Proceedings of the 16th Text REtrieval Conference (TREC 2007). Retrieved September 21, 2009, from http://trec.nist.gov/pubs/trec16/papers/LEGAL.OVERVIEW16.pdf

Tonta, Y. (1991). A study of indexing consistency between Library of Congress and British Library catalogers, Library Resources & Technical Services, 35(2), 177–185.

Voorhees, E.M. (1998). Variations in relevance judgments and the measurement of retrieval effectiveness. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 315–323). New York: ACM Press.

DOI: 10.1002/asi