VIRGINIA:

IN THE CIRCUIT COURT FOR LOUDOUN COUNTY

GLOBAL AEROSPACE INC., et al.

Plaintiff,

v.

LANDOW AVIATION, L.P. d/b/a
Dulles Jet Center, et al.

Defendants.

CONSOLIDATED CASE NO. CL 61040

<u>CASES AFFECTED</u>

Global Aerospace Inc., et al. v. Landow Aviation, L.P. d/b/a Dulles Jet Center, et al. (Case No. CL 61040)

BAE Systems Survivability Systems, LLC v. Landow Aviation, L.P., et al. (Case No. CL 61991)

La Réunion Aérienne v. Landow Aviation, L.P. d/b/a Dulles Jet Center, et al. (Case No. CL 64475)

United States Aviation Underwriters, Inc. v. Landow Aviation, L.P., et al. (Case No. CL 63795)

Chartis Aerospace Adjustment Services, Inc. v. Landow Builders Inc., et al. (Case No. CL 63190)

Factory Mutual Insurance Company v. Landow Builders Inc., et al. (Case No. CL 63575)

The Travelers Indemnity Company, as subrogee of Landow Aviation Limited Partnership v. Bascon, Inc., et al. (Case No. CL 61909)

Global Aerospace, Inc. v. J. H. Brandt and Associates, Inc., et al. (Case No. CL 61712)

M.I.C. Industries, Inc. v. Landow Aviation, L.P., et al. (Case No. 71633)

MEMORANDUM IN SUPPORT OF MOTION FOR PROTECTIVE ORDER <u>APPROVING THE USE OF PREDICTIVE CODING</u>

I. INTRODUCTION

Landow Aviation Limited Partnership, Landow Aviation I, Inc., and Landow & Company Builders, Inc. (collectively "Landow") have moved the Court for a protective order because counsel for a number of parties have objected to Landow's proposed use of "predictive coding" to retrieve potentially relevant documents from a massive collection of electronically stored information ("ESI"). The ESI retrieved by predictive coding would be reviewed by lawyers or paralegals and, if responsive and not privileged or otherwise immune from discovery, produced to the parties. The use of predictive coding is a reasonable means of locating and retrieving documents that may be responsive to requests for production and, therefore, satisfies the Rules of the Supreme Court of Virginia and should be approved in this case to avoid the undue burden and expense associated with the alternative means of culling the Landow ESI collection.

Landow has an estimated 250 gigabytes (GB) of reviewable ESI from its computer systems, which could easily equate to more than two million documents. At average cost and rates of review and effectiveness, linear first-pass review would take 20,000 man hours, cost two million dollars, and locate only sixty percent of the potentially relevant documents.   As one alternative, keyword searching might be more cost-effective but likely would retrieve only twenty percent of the potentially relevant documents and would require Landow to incur substantial unnecessary costs for document review. Predictive coding, on the other hand, is capable of locating upwards of seventy-five percent of the potentially relevant documents and can be effectively implemented at a fraction of the cost and in a fraction of the time of linear review and keyword searching. Further, by including a statistically sound validation protocol, Landow's counsel will thoroughly discharge the "reasonable inquiry" obligations of Rule 4:1(g). Therefore, this Honorable Court should enter an Order approving the use of predictive coding as set forth herein, thereby allowing Landow to locate more potentially relevant documents while avoiding the undue burden and expense associated with other means of culling the Landow ESI.[1]

[1] Given their opposition to the implementation of a more economical and effective means of culling the ESI, Landow respectfully requests that, if it is not inclined to approve the use of predictive coding, this Honorable Court shift any incremental costs associated with a more expensive alternative to the opposing parties.

II. BACKGROUND

The Court is well aware of the genesis of this litigation, which stems from the collapse of three hangars at the Dulles Jet Center ("DJC") during a major snow storm on February 6, 2010. The parties have exchanged substantial discovery requests addressing both liability and damages. The liability discovery is directed largely at responsibility for the collapse and, more specifically, the existence of any design or construction deficiencies that may have contributed to the failure. Pursuant to the Rules of the Supreme Court of Virginia, the discovery includes requests for ESI.

A. The Landow ESI Collection

Landow took several steps to promptly collect and preserve ESI, resulting in a collection of more than eight terabytes (TB) (8,000 GB) of forensic electronic images within a few months of the collapse.[2] Landow then retained JurInnov, Ltd. to consolidate the images into a more manageable collection of reviewable ESI.

JurInnov first conformed and exported all of the email files. Then JurInnov removed all of the duplicate files and the common application/system files. Finally, JurInnov filtered the collection to segregate and eliminate any non-data file types from commonly

[2] Landow maintained computer systems that may contain ESI relating to this litigation at two locations – the corporate offices in Bethesda, Maryland, and the DJC location at the Dulles International Airport. Shortly after the collapse, Landow began to collect ESI. On February 17, 2010, Landow collected Norton Ghost backups from most of the operating computers at the Bethesda office. This resulted in the collection of 3.5 TB of data. On March 10, 2010, Simone Forensics Consulting, LLC collected forensic images of all of the personal computers at the DJC location, as well as a forensic image of the data located on the DJC server, resulting in an additional 1.05 TB of data. Then, on April 22 and 27, 2010, Simone returned to the Bethesda office to collect, as a complement to the Ghost backups, forensic images of the hard drives of all the personal computers, as well as the data located on the two operating servers. This effort resulted in the collection of an additional 3.6 TB of data, bringing the entire image collection to just over eight terabytes of data.

recognized data file types (such as Microsoft Word files, which are *doc* or *docx* file types). This processing step reduced the ESI images to a collection of roughly 200 GB of reviewable data – 128 GB of email files, and 71.5 GB of native data files.

In order to collect and preserve any subsequently generated ESI, Landow engaged JurInnov to do another data collection from both locations in April of 2011. Although this new ESI has not been fully processed, JurInnov estimates that it will generate an additional 32.5 GB of email files and 18 GB of native files.

Based on values typically seen in electronic discovery, this estimated 250 gigabyte collection of reviewable ESI could easily comprise more than two million documents, covering every aspect of the Landow operations for a period of several years.

In order to assess the extent to which the documents in the ESI collection might pertain to this case, Landow conducted a cursory preliminary review of the data, from which it became apparent that a significant portion of the Landow ESI is wholly unrelated to the DJC project. JurInnov loaded the ESI from three Landow personnel who were involved in the DJC project into an e-Discovery tool for review and analysis of the documents. The documents were automatically separated by the software into clusters, each cluster representing a relatively unique concept common among the constituent documents. It was readily apparent that the majority of the clusters reflected concepts that were entirely unrelated to the DJC Project, such as email virus warnings and disclaimers. Even when the broad concepts might be pertinent to the DJC, it was obvious that only a fraction of the documents in a cluster actually related to the construction, operation, or collapse of the DJC.

It became apparent through the preliminary review that it would be necessary to cull the collected ESI to reduce the burden of document review and improve accuracy by eliminating documents having nothing to do with the DJC and generating a much smaller set of documents potentially relevant to the case, which could then be reviewed for discovery purposes.

B. Alternatives For Culling The Landow ESI Collection

There generally are three ways to cull an ESI collection to locate potentially relevant documents, although they are not equally effective. Historically, documents were reviewed one-by-one by a team of human reviewers. This first-pass linear review, however, is very time-consuming and expensive, and it is not particularly effective. More recently, ESI collections were culled by searching for keywords designed to locate documents containing select words expected to be pertinent to the litigation. Keyword searching typically is less expensive than linear review, but it is generally not very effective at finding relevant documents. **Today, the most effective and economical means of reviewing large ESI collections is a technology known as predictive coding.**[3]

Given the size of the Landow ESI collection, first-pass linear review would be extraordinarily expensive and time-consuming. With more than 2 million documents to review, albeit for potential relevance alone, it would take reviewers more than 20,000 hours to review each document individually – that's ten (10) man-years of billable time. Even at modest contract review rates, a linear review of this magnitude would almost certainly cost more than two million

[3] Predictive coding uses direct input from an attorney reviewing a small subset of the collection to generate a mathematical model of relevant documents. The model is then used to identify documents in the balance of the collection that are relevant and segregate them from documents that are not relevant. Predictive coding can be orders of magnitude faster (and less expensive) than linear review, and it is much more effective than both linear review and keyword searching at both locating the relevant documents and eliminating the documents that are not relevant.

dollars just to identify potentially relevant documents to be reviewed by Landow during discovery.

Beyond the sheer magnitude of such a project, the inherent problem with linear review is that it is neither consistent nor particularly effective at identifying and segregating relevant documents from those that are not relevant. *See, e.g.*, Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII Rich. J.L. & Tech. 11 (2011).[4] There are two exemplary studies evaluating the consistency among human reviewers, one by Ellen Voorhees and another by the team of Dr. Herbert Roitblat, Anne Kershaw, and Patrick Oot. *Id*. at 10 -13. Voorhees asked three teams of *professional information retrieval experts* to identify relevant documents in response to several information requests. *Id*. at 10 -11. Voorhees found that even experienced information retrieval experts agreed, at most, only 49.4% of the time. *Id*. Roitblat, *et al.*, similarly asked two teams of lawyers to identify relevant documents in response to a Department of Justice Information Request concerning the Verizon acquisition of MCI. *Id*. at 13. In the Roitblat study, the agreement between the two teams of lawyers was only 28.1% of the responsive documents.. *Id*. Thus, "[i]t is well established that human assessors will disagree in a substantial number of cases as to whether a document is relevant, regardless of the information need or the assessors' expertise and diligence." *Id*. at 9.

Consistency aside, linear review simply is not very effective. There are two measures of the effectiveness of information retrieval – recall and precision. *Recall* is the

[4] A true and correct copy of the Grossman-Cormack article appearing in the Richmond Journal of Law and Technology (hereafter "*Technology-Assisted Review*") is attached hereto as Exhibit A.

percentage of the relevant documents in the collection that are found by the reviewer. *Technology-Assisted Review*, p. 8. Thus, a recall of 100% means that a reviewer retrieved every relevant document from a collection. *Precision*, on the other hand, is the percentage of documents pulled by the reviewer that are actually relevant. *Id.* The balance of the documents selected by the reviewer would be irrelevant. Therefore, a precision of 70% means that 30% of the documents pulled by the reviewer would be irrelevant. Across both studies, Voorhees and Roitblat, *et al*., determined that reviewer recall ranged from 52.8% to 83.6%, and precision ranged from 55.5% to 81.9%. *Id*. at 15 -17. Grossman and Cormack analyzed data from the Text Retrieval Conference (TREC), and found that recall ranged from 25% to 80% (59.3% on average), while precision varied from an abysmal 5% to an unusual 89% (31.7% on average). *Technology-Assisted Review*, p. 37, Table 7. In a discovery context, this means that linear review misses, on average, 40% of the relevant documents, and the documents pulled by human reviewers are nearly 70% irrelevant.

In general, keyword searching is a much less expensive means of culling a collection set. There will be technical costs associated with preparing the data and the indices necessary to conduct an effective keyword search, and costs will escalate as the complexity of the search increases. In addition, there will be legal costs associated with negotiating the appropriate keyword list to use, which often is not a simple, straightforward exercise. And legal costs will similarly increase if iterative keyword searches are used to refine the selection of relevant documents from an ESI collection, as increasing document review and negotiation will be necessary.

Keyword searching, however, is simply not an effective means of separating the wheat from the chaff in an effort to locate relevant documents. The preeminent discussion of the effectiveness of keyword searching was a study by Blair & Maron in 1985. *Technology-Assisted Review*, pp. 18 -19.[5] With no constraints on the ability to conduct keyword searches, the average recall was only twenty percent (20%), which means that keyword searches missed 80% of the relevant documents. *Technology-Assisted Review*, p. 18. Although the precision was reasonably high (at 79%), that is not always the case. Indeed, in one case before the United States District Court for the Western District of Pennsylvania only seven percent (7%) of the documents found using keyword searching were ultimately relevant. *Hodczak v. Latrobe Specialty Steel Co.*, 761 F. Supp. 2d 261, 279 (W.D. Pa. 2010) In other words, ninety-three percent (93%) of the documents identified using keyword searches were irrelevant to the litigation.

The preliminary review of the Landow ESI suggests that keyword searching would be similarly ineffective in this case. In the review discussed above, the documents were summarily classified as either relevant or irrelevant to get a sense of the likely distribution of electronic documents. JurInnov then compiled a list of dominant keywords contained within the documents, indicating the number of hits for both the relevant and irrelevant sets. In evaluating proposed keywords,[6] Landow determined that many of the keywords are likely to generate more documents that are not relevant than those that are relevant. For example, the terms Dulles Jet, Sergio, Plaza, and Curro (terms requested by non-Landow counsel) were not found to be

[5] Blair & Maron asked skilled reviewers to compose keyword searches to retrieve at least 75% of the relevant documents in response to a number of document requests derived from a BART train accident in the San Francisco Bay Area. *Technology-Assisted Review*, p. 18-19. *See also, The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery*, 8 The Sedona Conf. Journal, Fall 2007, at p. 189. A true and correct copy of the Sedona Conference materials (hereafter "*The Sedona Conference Best Practices Commentary*"), is attached hereto as Exhibit B.

[6] The most recent communication on proposed keywords was a letter from Jonathan Berman dated March 1, 2012, a true and correct copy of which is attached hereto as Exhibit C.

predominant words in the relevant document set but were found in the irrelevant set. Other terms showed a similar pattern with significant percentages of documents coming from the irrelevant set: Jet Center (64%), hangar (33%), Mickey (71%), column (53%) and inspection (85%). This, by no means, was an exhaustive review. Rather it is illustrative of two problems that would be encountered if keyword searches were used to cull the Landow ESI – (1) a significant number of irrelevant documents would be included in the result; and (2) it would be difficult to determine from the result why certain words were, or were not, contained in either the relevant or irrelevant set, and the implications of that distribution.

C. Predictive Coding

Predictive coding is an economical, efficient, and effective alternative to both linear review and keyword searching. Predictive coding will retrieve more of the relevant, and fewer of the irrelevant, documents than the other two culling methods, and it will do so more quickly and at a lower cost.

The technology underlying predictive coding has been in existence for many years. For example, some predictive coding technologies employ Bayesian probability systems that “set[ ] up a formula that places a value on words, their interrelationships, proximity and frequency.” *The Sedona Conference Best Practices Commentary*, p. 218. These Bayesian systems are based on a theorem that was developed by a British mathematician in the eighteenth century. *Id*.

Basic predictive coding technology is so prevalent that virtually everyone uses it today. It is the same technology that underlies spam filters, which are used to prevent unwanted emails from flooding our inboxes. *Technology-Assisted Review*, p. 22. Although technologies differ somewhat, the general operation of predictive coding tools is conceptually similar. First, ESI is loaded into the tool, just as it would be loaded into a review tool for either linear review or keyword searching. Then, an experienced reviewing attorney "trains" the tool to recognize relevant documents and differentiate them from documents that are not relevant.[7] Once the tool has stabilized, training ceases and the tool applies a developed mathematical model to segregate or prioritize (depending in the particular tool) relevant and irrelevant documents. Predictive coding tools can leverage an attorney's review of a fraction of the ESI documents into the categorization of millions of documents as either relevant or irrelevant.

Because a reviewing attorney has to review and code only a fraction of the collection set, **the cost associated with predictive coding can be orders of magnitude less than the cost of linear review, and it will take far less time** to identify the relevant documents from among the collected ESI. Indeed, if the predictive coding tool is stabilized by coding 3,000 or fewer documents, it would take less then two weeks to cull the relevant documents from a multi-million document set **at roughly 1/100th of the cost of linear review**. Similarly, predictive coding can be accomplished in less time and at less expense than an involved iterative keyword search program that requires an attorney to review each set of documents to derive improvements to the keyword search criteria.

[7] There are inherent differences in the manner of training predictive coding tools. Certain tools present the attorney with small random or nearly random sets of training documents from the collection set. Other tools depend on the ability to locate and introduce a "seed set" of clearly relevant documents to prompt the tool to select the training documents. The attorney decides whether each document is relevant or not and codes that decision into the tool. As coding decisions are made, the tool processes those decisions and, by developing an evolving mathematical model of the attorney's decisions, learns to recognize the difference between relevant and irrelevant documents. When the tool retrieves the next set of training documents, the tool "predicts" whether the document will be coded as relevant or irrelevant. When the attorney codes the document, the tool tracks the extent of agreement. As the iterations proceed, the tool becomes more effective at predicting relevance, until it has stabilized and can accurately predict relevance. With random training, the attorney typically only has to review 2,000 to 3,000 documents to stabilize the tool; other tools may require review of ten to thirty percent of the collection before stabilizing.

Perhaps even more importantly, **predictive coding will be more effective at retrieving relevant documents and eliminating irrelevant documents than either linear review or keyword searching**. As discussed above, linear review retrieves only an average of 59.3% of the relevant documents, while nearly 70% of the documents retrieved will be irrelevant. *Technology-Assisted Review*, p. 37, Table 7. Even worse, keyword searching misses 80% of the relevant documents. *Id*. at 18. Predictive coding, on the other hand, effectively identifies an average of 76.7% of the relevant documents, while only 15.5% of the documents retrieved will be irrelevant. *Id*. at 37, Table 7. Thus, for both applicable information retrieval measures, predictive coding is better than linear first-pass review or keyword searching.

D. Proposed ESI Protocol

Landow is proposing a predictive coding protocol that will effectively ensure that the tool is operated properly and also that a reasonable fraction of relevant documents will be identified from within the Landow ESI to be reviewed for discovery purposes. First, Landow will, with the exception of (1) privileged documents, and (2) sensitive documents coded as irrelevant, provide the full set of training documents to opposing counsel once the tool has stabilized but before it is used to separate the relevant from the irrelevant documents. The privileged and irrelevant sensitive documents will be logged sufficiently for opposing counsel to determine whether there is a need to review the documents to evaluate the coding decision and whether the coding decision appears to be correct. In the event counsel believe documents were improperly coded or are improperly being withheld, they can request a modification, failing which they may bring the matter before the Court.[8] Should it be agreed or determined that any

[8] In the event an agreement cannot be reached, the Court would need only to determine whether documents withheld by Landow must be given to opposing counsel for further consideration and

...Continued

coding decisions need to be modified, the changes will be made and the tool will be re-stabilized before being used to finally categorize the relevant and irrelevant documents. This provides other counsel with an effective means of monitoring the propriety of the coding and the operation of the predictive coding tool.

Once predictive coding has been completed and the documents within the Landow ESI have been categorized, Landow will implement a statistically valid sampling program to establish that the majority of the relevant documents have been retrieved, *i.e.*, that an acceptable level of recall has been achieved. Given that recall for linear review averages only 59.3%, Landow proposes an acceptable recall criterion of 75%. In other words, predictive coding will conclude once the sampling program establishes that at least 75% of the relevant documents have been retrieved from the Landow ESI and are available to Landow for discovery purposes (*e.g.*, privilege review, responsiveness review, etc.).

In order to determine the recall achieved by predictive coding, statistically valid random samples will be taken from both the relevant and irrelevant document set. The size of the sample sets will be determined at the conclusion of the predictive coding once the distribution of relevant documents becomes apparent and any disagreement concerning sample size may be raised with the Court. Landow will review both sample sets to examine the relevance of each document and thereby determine the number of relevant documents in both the relevant and irrelevant sets. Recall will be calculated as the percentage of the total number of relevant documents existing in the relevant set.

Continued from previous page
whether any relevant documents were improperly determined to be irrelevant – a determination well within the typical purview of the Court in addressing discovery disputes.

In order to permit opposing counsel to substantiate the calculation, Landow will make these documents available to opposing counsel after logging and removing the privileged and irrelevant sensitive documents. If counsel disagree with the position taken by Landow, the Court may determine the relevance of any document or the propriety of withholding any particular document.

If recall as calculated using this procedure does not equal or exceed 75%, the predictive coding process will be extended to achieve an acceptable recall. If recall is at least 75%, the relevant set will alone comprise the review set of documents for use by Landow for purposes of discovery.

E. The Dispute

Landow has proposed to opposing counsel its use of predictive coding to quickly, effectively, and economically cull the relevant documents from within the Landow ESI collection. Opposing counsel, for reasons that have not been fully articulated, will not consent. There have been a number of discussions among counsel, and Landow has attempted to respond to any reasonable inquiry from opposing counsel to explain the decision to use predictive coding.[9] Lacking consent, Landow chose to seek judicial approval of its use of predictive coding to protect it from the unnecessary expense and burden associated with other, less effective means of culling the Landow ESI.

III. DISCUSSION

There are two issues pertinent to the Court's evaluation of Landow's predictive coding protocol. The first is whether the burden and expense associated with alternative means

[9] A true and correct copy of Landow's response to the questions posed by opposing counsel is attached hereto as Exhibit D.

of culling the Landow ESI warrant the entry of a protective order. The second is whether the use of predictive coding complies with counsel's duty of reasonable inquiry in the discovery context. As set forth below, both questions should be answered in the affirmative.

A. The Entry Of An Order Approving The Use Of Predictive Coding Is Appropriate Under Rules 4:1(b)(1) And 4:1(c) To Protect Landow From The Unnecessary Burden And Expense Of Alternative Means Of Culling The Landow ESI

There are two Rules of the Supreme Court of Virginia that control the disposition of Landow's Motion for Protective Order, and both favor the entry of an Order approving the use of predictive coding. Rule 4:1(b)(1) establishes the scope of discovery and authorizes the Court to limit discovery when it is found to be unduly burdensome or expensive.[10] Rule 4:1(c) authorizes the Court to fashion an appropriate remedy to protect a party from such undue burden and expense.[11]

[10] That rule provides the following:

> (b) *Scope of Discovery*. In General. Parties may obtain discovery regarding any matter, not privileged, which is relevant to the subject matter involved in the pending action, whether it relates to the claim or defense of any other party.... Subject to the provisions of Rule 4:8(g) [regarding limitations on interrogatories], the frequency or extent of use of the discovery methods set forth in subdivision (a) shall be limited by the court if it determines that... (iii) the discovery is unduly burdensome or expensive, taking into account the needs of the case, the amount in controversy, limitations on the parties' resources, and the importance of the issues at stake in the litigation. The court may act upon its own initiative after reasonable notice to counsel of record or pursuant to a motion under subdivision (c).

Va. Supreme Ct. Rule 4:1(b).

[11] That rule provides the following:

> (c) *Protective Orders.* Upon motion by a party or by the person from whom discovery is sought, accompanied by a certification that the movant has in good faith conferred or attempted to confer with other affected parties in an effort to resolve the dispute without court action, and for good cause shown, the court in which the action is pending... *may make any order which justice requires* to

...Continued

Pursuant to Rule 4.1(c), the Court, upon a showing of good cause by the party seeking protection, may enter "any order which justice requires to protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense." *Shenandoah Pub. House, Inc. v. Fanning*, 235 Va. 253, 261-62 (1988); *Philip Morris Co., Inc. v. American Broadcasting Co., Inc.*, 36 Va. Cir. 1, 2 (Richmond 1994). What constitutes good cause "depends upon the circumstances of the individual case, and a finding of its existence lies largely in the discretion of the court." *Philip Morris Co.*, 36 Va. Cir. at 2. Indeed, "[t]he trial court is in the best position to weigh fairly the competing needs and interests of parties affected by discovery. The unique character of the discovery process requires that the trial court have substantial latitude to fashion protective orders." *Shenandoah Pub. House*, 235 Va. at 261 (quoting *Seattle Times Co. v. Rhinehart*, 467 U.S. 20, 34-6 (1984) (noting that "the discovery rules central to the Supreme Court's rationale and our Rules 4:1(b)(1) and (c) are essentially the same"); *accord DuPont v. Winchester Medial Ctr., Inc.*, 34 Va. Cir. 105, 109 (City of Winchester 1994) ("The Court has considerable discretion in the issuance of a protective order.").

Good cause exists to enter a protective order approving the use of predictive coding to retrieve the relevant documents from the Landow ESI collection. Predictive coding is the least expensive and most expeditious means of culling the Landow ESI. Linear review likely will cost one hundred times more than predictive coding, and predictive coding will generate a result in 1/100th the time. Effective keyword searching likely will be more expensive and time-consuming as well since an iterative keyword search will undoubtedly involve the review of

Continued from previous page

> protect a party or person from annoyance, embarrassment, oppression, or undue burden or expense....

Va. Supreme Ct. Rule 4:1(c) (emphasis added).

more documents than will be necessary to stabilize a predictive coding tool. Moreover, predictive coding will retrieve many more of the relevant documents, and fewer of the irrelevant documents, than either linear review or keyword searching. And this protocol benefits all parties because the relevant documents will be produced more quickly, allowing for expeditious discovery and a quicker resolution of this case. Moreover, employing the proposed ESI protocol, Landow will provide opposing counsel with a better result than would otherwise be available, and it will be much less expensive for Landow to do so. There is certainly sufficient good cause to support the entry of a protective order to avoid undue burden and expense.

Accordingly, this Court is authorized under Virginia Supreme Court Rule 4:1(c) to fashion a protective order approving the use of predictive coding by Landow as one "which justice requires." Va. Supreme Ct. Rule 4:1(c).

B. The Sampling Protocol Proposed By Landow Will More Than Satisfy The "Reasonable Inquiry" Obligations Of Rule 4:1(G)

Rule 4:1(g) of the Virginia Supreme Court Rules requires counsel to conduct a reasonable inquiry to determine that a response to a discovery request is "consistent with [the] Rules and warranted by existing law or a good faith argument for the extension, modification or reversal of existing law...." Va. Supreme Ct. Rules 4:1(g). Whether that obligation is met requires the application of "an objective standard of reasonableness as to whether a reasonable inquiry was made such that the attorney could have formed a reasonable belief that a response to a request for production of documents was well grounded in fact and warranted under existing law." *Lester v. Allied Concrete Co.*, 80 Va. Cir. 454, 460 (2010) (noting the paucity of law on Rule 4:1(g) and applying instead the analysis under Va. Code § 8.01-3). "Such an objective standard of reasonableness requires consideration of several factors." *Ford Motor Co. v. Benitez*, 273 Va. 242, 253 (2007) (similarly construing Va. Code § 8.01-3).

The reasonable inquiry standard in this context does not truly present a question of the application or extension of legal precedent. There is no judicial or other legal mandate requiring, or even advocating, the use of one method of document retrieval over another.

To the extent that judicial approval may be at issue, however, leading jurisprudence supports the use of predictive coding. *See Da Silva Moore v. Publicis Groupe,* 2012 U.S. Dist. LEXIS 23350 (S.D.N.Y. 2012) (Peck, J.).[12] The issue in *Da Silva Moore* related to the use of predictive coding to cull through an estimated three million electronic documents. *Id*. at *9. Judge Peck noted at the outset the lack of any judicial assessment of predictive coding and the limiting effect upon the bar.[13] Judge Peck concluded: "This judicial opinion now

[12] A true and correct copy of Magistrate Judge Peck's February 24, 2012 Opinion in *Da Silva Moore* is attached hereto as Exhibit E. Plaintiffs in the *Da Silva Moore* case have appealed Judge Peck's decision on the precise protocol; however, his Opinion currently remains the only definitive statement on the use of predictive coding.

[13] Judge Peck began his Opinion as follows:

> In my article *Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?,* I wrote:
>
> To my knowledge, no reported case (federal or state) has ruled on the use of computer-assisted coding. While anecdotally it appears that some lawyers are using predictive coding technology, it also appears that many lawyers (and their clients) are waiting for a judicial decision approving of computer-assisted review.
>
> Perhaps they are looking for an opinion concluding that: "It is the opinion of this court that the use of predictive coding is a proper and acceptable means of conducting searches under the Federal Rules of Civil Procedure, and furthermore that the software provided for this purpose by [insert name of your favorite vendor] is the software of choice in this court." If so, it will be a long wait.
>
> * * *
>
> Until there is a judicial opinion approving (or even critiquing) the use of predictive coding, counsel will just have to rely on this article as a sign of judicial approval. In my opinion, computer-assisted coding should be used in those cases where it will help "secure the just, speedy, and inexpensive" (Fed. R. Civ. P. 1) determination of cases in our e-discovery world.

...Continued

recognizes that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases." *Id*. at *3.

The reasonable inquiry standard, rather, poses the question of whether counsel has ensured that a sufficient search has been conducted to obtain documents relevant to the case from which thorough discovery responses may be prepared. In this case, the answer absolutely is yes. Landow has collected virtually all available ESI and winnowed that ESI to a collection set of reviewable data. Predictive coding will further refine the collection by identifying, with reasonable accuracy, only those documents that relate to this case. Analogizing to traditional discovery, predictive coding will find the relevant folders and documents in the filing cabinet from among the mass of files relating to all aspects of Landow's business – which can then be searched to respond to discovery requests. And to ensure that the relevant ESI has been retrieved, Landow will take the added step of implementing a statistically valid sampling program to evaluate the ESI left behind.

Ironically, what is being proposed in this case to ensure "reasonable inquiry" is far more than has ever been done with traditional discovery. Never has it been suggested that a producing party would be obligated to sample the documents determined to be irrelevant by first-pass reviewers to demonstrate the adequacy of that first-pass review. Nor is it a typical practice to review the documents left behind by a keyword search, even though upwards of 80% of the relevant documents typically are missed. The ESI protocol proposed by Landow goes well beyond what would otherwise be available to opposing counsel to ensure that relevant documents are being reviewed for production.

Continued from previous page

Da Silva Moore, at *2-3.

Thus, both relatively and absolutely, the ESI protocol proposed by Landow, premised upon the use of predictive coding technology, satisfies the reasonable inquiry obligations of Rule 4:1(g) of the Virginia Supreme Court Rules.

IV. CONCLUSION

For the foregoing reasons, Landow Aviation Limited Partnership, Landow Aviation I, Inc., and Landow & Company Builders, Inc. respectfully request that this Honorable Court enter an Order approving the use of predictive coding technology to cull the relevant documents from the Landow ESI for purposes of discovery in this litigation.

Dated: April 9, 2012

Respectfully submitted,

LANDOW AVIATION LIMITED PARTNERSHIP
LANDOW AVIATION I, INC.
By Counsel

SCHNADER HARRISON SEGAL & LEWIS LLP

By: ____________________________
Jonathan M. Stern (VA Bar No. 41930)
Gordon S. Woodward (VA Bar No. 42449)
750 Ninth Street, NW, Suite 550
Washington, DC 20001-4534
Telephone: (202) 419-4202
Facsimile: (202) 419-4252
E-mail: jstern@schnader.com

LANDOW & COMPANY BUILDERS, INC.
By Counsel

BAXTER, BAKER, SIDLE, CONN & JONES, P.A.

By: Danielle M. LaCoe/GRJ
Gary R. Jones
Danielle M. LaCoe (VA Bar # 78715)
120 E. Baltimore Street, Suite 2100
Baltimore, Maryland 21202
Telephone: (410) 230-3800
E-mail: grj@bbsclaw.com

CERTIFICATE OF SERVICE

I hereby certify that, on this 9th day of April 2012, I served the foregoing **MOTION AND MEMORANDUM IN SUPPORT OF MOTION FOR PROTECTIVE ORDER APPROVING THE USE OF PREDICTIVE CODING**, as follows. Counsel identified in Section I were served via electronic mail at the address as indicated below. Counsel and/or parties identified in Section II were served via U.S. Mail first-class, postage prepaid at the address indentified below.

SECTION I – ELECTRONIC MAIL

Randal R. Craft, Jr., Esq. randal.craft@hklaw.com Brandon H. Elledge, Esq. brandon.elledge@hklaw.com Michelle T. Hess, Esq. michelle.hess@hklaw.com ***Attorneys for Plaintiff, BAE Systems Survivability Systems, LLC***	James W. Walker, Esq. jwalker@vanblk.com Casaundra M. Maimone, Esq. cmaimone@vanblk.com ***Attorneys for Bascon, Inc.***
Mark A. Dombroff, Esq. mdombroff@dglitigators.com Morgan W. Campbell, Esq. mcampbell@dglitigators.com ***Attorneys for Plaintiffs, Chartis and Global Aerospace***	Christopher E. Hassell, Esq. chassell@bonnerkiernan.com Craig L. Sarner, Esq. csarner@bonnerkiernan.com ***Attorneys for DGS Construction, Inc.***
James N. Markels, Esq. jmarkels@jackscamp.com Robert N. Kelly, Esq. rkelly@jackscamp.com ***Attorneys for EagleSpan Entities and Pagemark 3, Inc.***	James E. Moore, Esq. jmoore@cblaw.com S. Perry Coburn, Esq. pcoburn@cblaw.com ***Attorneys for EagleSpan Steel Structures, LLC***

C. Jay Robbins, IV, Esq. crobbins@midkifflaw.com William G. Frey, Esq. WFrey@gibbonslaw.com Michael R. Spitzer, Esq. mspitzer@midkifflaw.com Robert Taylor Ross, Esq. rross@midkifflaw.com Susan Yukevich (Paralegal) syukevich@midkifflaw.com Lori Campbell (Assistant) lcampbell@midkifflaw.com ***Attorneys for Plaintiff, Factory Mutual Insurance Company***	John B. Mesirow, Esq. john@metrodclaw.com mesirowEsq@aol.com William N. Clark, Jr., Esq. wclark@cozen.com ***Attorneys for Plaintiff, Federal Insurance Company***
Jonathan M. Stern, Esq. jstern@schnader.com Michael Bock, Esq. mbock@schnader.com Eric Smith, Esq. esmith@schnader.com Gordon S. Woodward, Esq. gwoodward@schnader.com Jennifer Callery, Esq. jcallery@schnader.com Levi Jones, Esq. ljones@schnader.com ***Attorneys for Landow Aviation***	David Rust Clarke, Esq. dclarke@bklawva.com William B. Porter, Esq. wporter@bklawva.com Niccolo N. Donzella, Esq. nnd@bbsclaw.com Gary R. Jones, Esq. grj@bbsclaw.com Trace Krueger, Esq. tkk@bbsclaw.com Danielle M. LaCoe, Esq. dml@bbsclaw.com Michele J. Rogers (Paralegal) mjr@bbsclaw.com ***Attorneys for Landow & Company Builders, Inc.***
Bruce E. Titus, Esq. btitus@reesbroome.com Alison Mullins, Esq. amullins@reesbroome.com Mark P. Graham, Esq. mgraham@reesbroome.com ***Attorneys for Nangia Engineering of Texas, Ltd.; Chander P. Nangia; Nangia Engineering, L.P.; and Nangia Engineering I, LLC***	Stephan F. Andrews, Esq. sandrews@vanblk.com Sean M. Golden, Esq. sgolden@vanblk.com ***Attorneys for Pinnacle Engineering, Inc.***
Jennifer A. Mahar, Esq. jmahar@smithpachter.com Erica J. Geibel, Esq. egeibel@smithpachter.com Brian Vella, Esq. bvella@smithpachter.com Sharen Burton (Legal Assistant) sburton@smithpachter.com ***Attorneys for Schnabel Operations, Inc. & Schnabel Engineering Consultants, Inc.***	William H. Atwill, Jr. batwill@atandlpc.com Gertrude C. Bartel, Esq. gbartel@kg-law.com ***Attorneys for Plaintiff, The Travelers Indemnity Company, as subrogee of Landow Aviation, LP***

Jonathan Berman, Esq.
JBerman@JonesDay.com
William G. Laxton, Jr., Esq.
wglaxton@jonesday.com

Attorneys for Plaintiff, M.I.C. Industries, Inc.

SECTION II – U.S. MAIL

Alexa K. Mosley, Esq. Ms. Samantha Patterson Leffler & Mosley 10555 Main Street, Suite 600 Fairfax, VA 22030 ***Attorneys for Independent Testing & Inspection Services, Inc.***	Dominion Caisson Corp. c/o Richard L. Windham Registered Agent 4337 Ridgewood Center Drive Woodbridge, VA 22192
Pinnacle Erectors, Inc. 101 Pioneer Camp Lane Pinnacle, NC 27043	

Gordon S Woodward

Gordon S. Woodward

Exhibit A

TECHNOLOGY-ASSISTED REVIEW IN E-DISCOVERY CAN BE MORE EFFECTIVE AND MORE EFFICIENT THAN EXHAUSTIVE MANUAL REVIEW

By Maura R. Grossman[*] & Gordon V. Cormack[†] [**]

Cite as: Maura R. Grossman & Gordon V. Cormack, *Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review*, XVII RICH. J.L. & TECH. 11 (2011), http://jolt.richmond.edu/v17i3/article11.pdf.

[*] Maura R. Grossman is counsel at Wachtell, Lipton, Rosen & Katz. She is co-chair of the E-Discovery Working Group advising the New York State Unified Court System, and a member of the Discovery Subcommittee of the Attorney Advisory Group to the Judicial Improvements Committee of the U.S. District Court for the Southern District of New York. Ms. Grossman is a coordinator of the Legal Track of the National Institute of Standards and Technology's Text Retrieval Conference ("TREC"), and an adjunct faculty member at Rutgers School of Law–Newark and Pace Law School. Ms. Grossman holds a J.D. from Georgetown University Law Center, and an M.A. and Ph.D. in Clinical/School Psychology from Adelphi University. The views expressed herein are solely those of the Author and should not be attributed to her firm or its clients.

[†] Gordon V. Cormack is a Professor at the David R. Cheriton School of Computer Science, and co-director of the Information Retrieval Group, at the University of Waterloo. He is a coordinator of the TREC Legal Track, and Program Committee member of TREC at large. Professor Cormack is the co-author of *Information Retrieval: Implementing and Evaluating Search Engines* (MIT Press, 2010), as well as more than 100 scholarly articles. Professor Cormack holds a B.Sc., M.Sc., and Ph.D. in Computer Science from the University of Manitoba.

[**] The Authors would like to thank Ellen Voorhees and Ian Soboroff at NIST for providing access to the raw TREC 2009 data. The Authors gratefully acknowledge the helpful comments received from Hon. John M. Facciola (D.D.C.), Hon. Paul W. Grimm (D. Md.), and Hon. Andrew J. Peck (S.D.N.Y.) on an earlier draft of this paper.

ABSTRACT

E-discovery processes that use automated tools to prioritize and select documents for review are typically regarded as potential cost-savers – but inferior alternatives – to exhaustive manual review, in which a cadre of reviewers assesses every document for responsiveness to a production request, and for privilege. This Article offers evidence that such technology-assisted processes, while indeed more efficient, can also yield results superior to those of exhaustive manual review, as measured by recall and precision, as well as F_1, a summary measure combining both recall and precision. The evidence derives from an analysis of data collected from the TREC 2009 Legal Track Interactive Task, and shows that, at TREC 2009, technology-assisted review processes enabled two participating teams to achieve results superior to those that could have been achieved through a manual review of the entire document collection by the official TREC assessors.

I. INTRODUCTION

[1] *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery* cautions that:

> [T]here 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.[1]

While the word *myth* suggests disbelief, literature on the subject contains little scientific evidence to support or refute the notion that automated methods, while improving on the efficiency of manual review, yield inferior results.[2] This Article presents evidence supporting the position that a technology-assisted process, in which humans examine only a small fraction of the document collection, can yield higher recall and/or precision than an exhaustive manual review process, in which humans code and examine the entire document collection.

[2] A *technology-assisted review process* involves the interplay of humans and computers to identify the documents in a collection that are responsive to a production request, or to identify those documents that should be withheld on the basis of privilege.[3] A human examines and

[1] The Sedona Conference, *The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*, 8 SEDONA CONF. J. 189, 199 (2007) [hereinafter *Sedona Search Commentary*].

[2] *Id.* at 194 ("The comparative efficacy of the results of manual review versus the results of alternative forms of automated methods of review remains very much an open matter of debate.").

[3] *See* Douglas W. Oard et al., *Evaluation of information retrieval for E-discovery,* 18:4 ARTIFICIAL INTELLIGENCE & LAW 347, 365 (2010) ("In some cases . . . the end user will interact directly with the system, specifying the query, reviewing results, modifying the

codes only those documents the computer identifies – a tiny fraction of the entire collection.[4] Using the results of this human review, the computer codes the remaining documents in the collection for responsiveness (or privilege).[5] A technology-assisted review process may involve, in whole or in part, the use of one or more approaches including, but not limited to, keyword search, Boolean search, conceptual search, clustering, machine learning, relevance ranking, and sampling.[6] In contrast, *exhaustive manual review* requires one or more humans to examine each and every document in the collection, and to code them as responsive (or privileged) or not.[7]

[3] Relevant literature suggests that manual review is far from perfect.[8] Moreover, recent results from the Text Retrieval Conference ("TREC"), sponsored by the National Institute of Standards and Technology ("NIST"), show that technology-assisted processes can achieve high levels of recall and precision.[9] By analyzing data collected

query, and so on. In other cases, the end user's interaction with the system will be more indirect. . . .").

[4] *See Sedona Search Commentary supra* note 1, at 209.

[5] *See* Maura R. Grossman & Terry Sweeney, *What Lawyers Need to Know About Search Tools*, THE NAT'L L.J. (Aug. 23, 2010), *available at* http://www.law.com/jsp/lawtechnologynews/PubArticleLTN.jsp?id=1202470952987&slreturn=1&hbxlogin=1 ("'machine learning tools,' use 'seed sets' of documents previously identified as responsive or unresponsive to rank the remaining documents from most to least likely to be relevant, or to classify the documents as responsive or nonresponsive.").

[6] *See, e.g.*, *Sedona Search Commentary*, *supra* note 1, at 217–23; CORNELIS JOOST VAN RIJSBERGEN, INFORMATION RETRIEVAL 74-85 (2d ed. 1979). The specific technologies employed in the processes that are the subjects of this study are detailed *infra* Parts III.A. – III.B.

[7] *See, e.g.,* Herbert L. Roitblat et al., *Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review*, 61 J. AM. SOC'Y. FOR INFO. SCI. AND TECH. 70, 70 (2010).

[8] *See, e.g.*, *Sedona Search Commentary, supra* note 1.

[9] Bruce Hedin et al., *Overview of the TREC 2009 Legal Track*, *in* NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS 16 & tbl.5 (2009), *available at* http://trec-

during the course of the TREC 2009 Legal Track Interactive Task,[10] the Authors demonstrate that the levels of performance achieved by two technology-assisted processes exceed those that would have been achieved by the official TREC assessors – law students and lawyers employed by professional document-review companies – had they conducted a manual review of the entire document collection.

[4] Part II of this Article describes document review and production in the context of civil litigation, defines commonly used terms in the field of information retrieval, and provides an overview of recent studies. Part III details the TREC 2009 Legal Track Interactive Task, including the H5 and Waterloo efforts, as well as the TREC process for assessment and gold-standard creation. Part IV uses statistical inference to compare the recall, precision, and F_1 scores that H5 and Waterloo achieved to those the TREC assessors would have achieved had they reviewed all of the documents in the collection. Part V presents a qualitative analysis of the nature of manual review errors. Parts VI, VII, and VIII, respectively, discuss the results, limitations, and conclusions associated with this study. Ultimately, this Article addresses a fundamental uncertainty that arises in determining what is reasonable and proportional: Is it true that if a human examines every document from a particular source, that human will, as nearly as possible, correctly identify all and only the documents that should be produced? That is, does exhaustive manual review guarantee that production will be as complete and correct as possible? Or can technology-assisted review, in which a human examines only a fraction of the documents, do better?

II. CONTEXT

[5] Under Federal Rule of Civil Procedure 26(g)(1) ("Rule 26(g)(1)"), an attorney of record must certify "to the best of [his or her] knowledge,

legal.umiacs.umd.edu/LegalOverview09.pdf; *see also* Douglas W. Oard et al., *Overview of the TREC 2008 Legal Track*, *in* NIST SPECIAL PUBLICATION: SP 500-277, THE SEVENTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2008) PROCEEDINGS 8 (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf.

[10] *See* Hedin et al., *supra* note 9, at 2.

information, and belief formed after a reasonable inquiry," that every discovery request, response, or objection is

> consistent with [the Federal Rules of Civil Procedure] . . . not interposed for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation[, and is] neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues at stake in the action.[11]

Similarly, Federal Rule of Civil Procedure 26(b)(2)(C)(iii) ("Rule 26(b)(2)(C)(iii)") requires a court to limit discovery when it determines that "the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues."[12] Thus, Rules 26(g)(1) and 26(b)(2)(C)(iii) require that discovery requests and responses be *proportional*.[13] However, Federal Rule of Civil Procedure 37(a)(4) ("Rule 37(a)(4)") provides that "an evasive or incomplete disclosure, answer or response must be treated as a failure to disclose, answer, or respond[,]" and therefore requires that discovery responses be *complete*.[14] Together, Rules 26(g)(1), 26(b)(2)(C)(iii), and 37(a)(4) reflect the tension – between completeness on one hand, and burden and cost on the other – that exists in all electronic discovery ("e-discovery") processes.[15] In

[11] FED. R. CIV. P. 26(g)(1).

[12] FED. R. CIV. P. 26(b)(2)(C)(iii).

[13] The Sedona Conference, *The Sedona Conference Commentary on Proportionality in Electronic Discovery,* 11 SEDONA CONF. J. 289, 294 (2010) [hereinafter *Sedona Proportionality Commentary*].

[14] FED. R. CIV. P. 37(a)(4).

[15] Typically, a responding party will not only seek to produce *all* responsive documents, but to identify *only* the responsive documents, in order to guard against overproduction or waiver of privilege. *See, e.g.,* Mt. Hawley Ins. Co. v. Felman Prod., Inc., 271 F.R.D. 125, 136 (S.D.W. Va. 2010) (finding that plaintiff's over-production of documents by more than 30% was a factor in waiver of privilege).

assessing what is reasonable and proportional with respect to e-discovery, parties and courts must balance these competing considerations.[16]

[6] One of the greatest challenges facing legal stakeholders is determining whether or not the cost and burden of identifying and producing electronically stored information ("ESI") is commensurate with its importance in resolving the issues in dispute.[17] In current practice, the problem of identifying responsive (or privileged) ESI, once it has been collected, is almost always addressed, at least in part, by a manual review process, the cost of which dominates the e-discovery process.[18] A natural question to ask, then, is whether this manual review process is the most effective and efficient one for identifying and producing the ESI most likely to resolve a dispute.

A. Information Retrieval

[7] The task of finding all, and only, the documents that meet "some requirement" is one of information retrieval ("IR"), a subject of scholarly

[16] *See* Harkabi v. Sandisk Corp., No. 08 Civ. 8203 (WHP), 2010 WL 3377338, at *1 (S.D.N.Y Aug. 23, 2010) ("Electronic discovery requires litigants to scour disparate data storage mediums and formats for potentially relevant documents. That undertaking involves dueling considerations: thoroughness and cost.").

[17] *See id.* at *8 ("Integral to a court's inherent power is the power to ensure that the game is worth the candle—that commercial litigation makes economic sense. Electronic discovery in this case has already put that principle in jeopardy."); Hopson v. Mayor of Balt., 232 F.R.D. 228, 232 (D. Md. 2005) ("This case vividly illustrates one of the most challenging aspects of discovery of electronically stored information—how properly to conduct Rule 34 discovery within a reasonable pretrial schedule, while concomitantly insuring that requesting parties receive appropriate discovery, and that producing parties are not subjected to production timetables that create unreasonable burden, expense, and risk of waiver of attorney-client privilege and work product protection"). *See generally Sedona Proportionality Commentary*, *supra* note 13.

[18] Marisa Peacock, *The True Cost of eDiscovery*, CMSWIRE, http://www.cmswire.com/cms/enterprise-cms/the-true-cost-of-ediscovery-006060.php (2009) (citing *Sedona Search Commentary*, *supra* note 1, at 192); Ashish Prasad et al., *Cutting to the "Document Review" Chase: Managing a Document Review in Litigation and Investigations*, 18 BUS. LAW TODAY, 2, Nov.–Dec. 2008.

research for at least a century.[19] In IR terms, "some requirement" is referred to as an *information need*, and *relevance* is the property of whether or not a particular document meets the information need.[20] For e-discovery, the information need is typically specified by a production request (or by the rules governing privilege), and the definition of relevance follows.[21] Cast in IR terms, the objective of review in e-discovery is to identify as many *relevant* documents as possible, while simultaneously identifying as few *nonrelevant* documents as possible.[22] The fraction of relevant documents identified during a review is known as *recall*, while the fraction of identified documents that are relevant is known as *precision*.[23] That is, *recall* is a measure of completeness, while *precision* is a measure of accuracy, or correctness.[24]

[8] The notion of *relevance*, although central to information science, and the subject of much philosophical and scientific investigation, remains elusive.[25] While it is easy enough to write a document describing an

[19] The concepts and terminology outlined in Part II.A may be found in many information retrieval textbooks. For a historical perspective, see GERARD SALTON & MICHAEL J. MCGILL, INTRODUCTION TO MODERN INFORMATION RETRIEVAL (1983); VAN RIJSBERGEN, *supra* note 6. For a more modern treatment, see STEFAN BÜTTCHER ET AL., INFORMATION RETRIEVAL: IMPLEMENTING AND EVALUATING SEARCH ENGINES 33–75 (2010).

[20] *See* BÜTTCHER ET AL., *supra* note 19, at 5-6, 8.

[21] *See* Hedin et al., *supra* note 9, at 1.

[22] *See* VAN RIJSBERGEN, *supra* note 6, at 4.

[23] *See* David C. Blair & M. E. Maron, *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289, 290 (1985) ("Recall measures how well a system retrieves *all* the relevant documents; and Precision, how well the system retrieves *only* the relevant documents."); VAN RIJSBERGEN, *supra* note 6, at 112-13.

[24] *See* VAN RIJSBERGEN, *supra* note 6, at 113.

[25] *See* Tefko Saracevic, *Relevance: A Review of the Literature and a Framework for Thinking on the Notion in Information Science*. *Part II: Nature and Manifestations of Relevance*, 58 J. AM. SOC'Y FOR INFO. SCI. & TECH. 1915 (2007); Tefko Saracevic, *Relevance: A Review of the Literature and a Framework for Thinking on the Notion in*

information need and hence relevance, determining the relevance of any particular document requires human interpretation.[26] It is well established that human assessors will disagree in a substantial number of cases as to whether a document is relevant, regardless of the information need or the assessors' expertise and diligence.[27]

[9] A review resulting in higher recall and higher precision than another review is more nearly complete and correct, and therefore superior,[28] while a review with lower recall and lower precision is inferior.[29] If one result has higher recall while the other has higher precision, it is not immediately obvious which should be considered superior. To calculate a review's effectiveness, researchers often employ F_1 – the harmonic mean of recall and precision[30] – a commonly used summary measure that rewards results achieving both high recall and high precision, while penalizing those that have either low recall or low precision.[31] The value of F_1 is always intermediate between recall and precision, but is generally closer to the lesser of the two.[32] For example, a result with 40% recall and 60% precision has F_1 = 48%. Following

Information Science. *Part III: Behavior and Effects of Relevance*, 58:13 J. AM. SOC'Y FOR INFO. SCI. & TECH. 2126 (2007).

[26] *See* Peter Bailey et al., *Relevance Assessment: Are Judges Exchangeable and Does It Matter?*, *in* SIGIR '08 PROCEEDINGS OF THE 31ST ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 667 (2008); *see also* VAN RIJSBERGEN, *supra* note 6, at 112.

[27] *See* Bailey et al., *supra* note 26, at § 4.3.

[28] *See* Blair & Maron, *supra* note 23.

[29] *See id*.

[30] $F_1 = \frac{2}{\frac{1}{recall}+\frac{1}{precision}}$.

[31] *See* BÜTTCHER ET AL., *supra* note 19, at 68.

[32] *See id.*

TREC, this Article reports recall and precision, along with F_1 as a summary measure of overall review effectiveness.[33]

B. Assessor Overlap

[10] The level of agreement between independent assessors may be quantified by *overlap* – also known as the *Jaccard index* – the number of documents identified as relevant by two independent assessors, divided by the number identified as relevant by either or both assessors.[34] For example, suppose assessor A identifies documents {W,X,Y,Z} as relevant, while assessor B identifies documents {V,W,X}. Both assessors have identified two documents {W,X} as relevant, while either or both have identified five documents {V,W,X,Y,Z} as relevant. So the overlap is 2/5, or forty percent. Informally, overlap of less than fifty percent indicates that the assessors disagree on whether or not a document is relevant more often than when they agree that a document is relevant.[35]

[11] In her study, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, Ellen Voorhees measured overlap between primary, secondary, and tertiary reviewers who each made 14,968 assessments of relevance for 13,435 documents,[36] with respect to 49

[33] *See* Hedin et al., *supra* note 9, at 3.

[34] Ellen M. Voorhees, *Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness*, 36 INFO. PROCESSING & MGMT 697, 700 (2000), *available at* http://www.cs.cornell.edu/courses/cs430/2006fa/cache/Trec_8.pdf ("Overlap is defined as the size of the intersection of the relevant document sets divided by the size of the union of the relevant document sets."); *see* CHRISTOPHER D. MANNING ET AL., AN INTRODUCTION TO INFORMATION RETRIEVAL 61 (2009) (draft), *available at* nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf; *see also* Raimundo Real & Juan M. Vargas, *The Probabilistic Basis of Jaccard's Index of Similarity*, 45 SYSTEMATIC BIOLOGY 380, 381 (1996).

[35] *See* Ellen M. Voorhees, *The Philosophy of Information Retrieval Evaluation*, *in* EVALUATION OF CROSS-LANGUAGE INFORMATION RETRIEVAL SYSTEMS SECOND WORKSHOP OF THE CROSS-LANGUAGE EVALUATION FORUM, CLEF 2001 DARMSTADT, GERMANY, SEPTEMBER 3-4, 2001 REVISED PAPERS 355, 364 (Carol Peters et al. eds., 2002).

[36] E-mail from Ellen M. Voorhees to Gordon V. Cormack (Jul. 31, 2019 14:34 EDT) (on file with authors). The numbers in the text are derived from the file,

information needs (or "topics," in TREC parlance), in connection with Ad Hoc Task of the Fourth Text Retrieval Conference ("TREC 4").[37] As illustrated in Table 1, the overlap between primary and secondary assessors was 42.1%;[38] the overlap between primary and tertiary assessors was 49.4%;[39] and the overlap between secondary and tertiary assessors was 42.6%.[40]

[12] Perhaps due to the assessors' expertise,[41] Voorhees' overlap results are among the highest reported for pairs of human assessors. Her findings demonstrate that assessors disagree at least as often as they agree that a document is relevant.[42] Voorhees concluded:

> The scores for the [secondary and tertiary] judgments imply a practical upper bound on retrieval system performance is 65% precision at 65% recall since that is the level at which humans agree with one another.[43]

"threeWayJudgments," attached to Voorhees' e-mail. Some of the documents were assessed for relevance to more than one topic.

[37] Voorhees, *supra* note 34, at 708; *see also* Donna Harman, *Overview of the Fourth Text REtrieval Conference (TREC-4)*, *in* NIST SPECIAL PUBLICATION 500-236: THE FOURTH TEXT RETRIEVAL CONFERENCE (TREC-4) 2 (2004), *available at* http://trec.nist.gov/pubs/trec4/t4_proceedings.html (follow the first link under "PAPERS").

[38] *See infra* Table 1; *see also* Voorhees, *supra* note 34, at 701 tbl.1.

[39] *See infra* Table 1; *see also* Voorhees, *supra* note 34, at 701 tbl.1.

[40] *See infra* Table 1; *see also* Voorhees, *supra* note 34, at 701 tbl.1.

[41] All assessors were professional information retrieval experts. Voorhees, *supra* note 34, at 701.

[42] *See id*.

[43] *Id*.

[13] It is not widely accepted that these findings apply to e-discovery.[44] This "legal exceptionalism" appears to arise from common assumptions within the legal community:

1. that the information need (responsiveness or privilege) is more precisely defined for e-discovery than for classical information retrieval;[45]

2. that lawyers are better able to assess relevance and privilege than the non-lawyers typically employed for information retrieval tasks;[46] and

3. that the most defensible way to ensure that a production is accurate is to have a lawyer examine each and every document.[47]

[44] *See Sedona Search Commentary*, *supra* note 1 (noting the widespread perception that manual review is nearly perfect). If that perception were correct, manual reviewers would have close to 100% overlap, contrary to Voorhees' findings. Vorhees, *supra* note 34, at 701 tbl.1.

[45] Oard et al., *supra* note 3, at 362 ("It is important to recognize that the notion of relevance that is operative in E-discovery is, naturally, somewhat more focused than what has been studied in information seeking behavior studies generally . . . .").

[46] *Cf.* Alejandra P. Perez, *Assigning Non-Attorneys to First-Line Document Reviews Requires Safeguards*, THE E-DISCOVERY 4-1-1 (LeClairRyan), Jan. 2011, at 1, *available at* http://marketing.leclairryan.com/files/Uploads/Documents/the-e-discovery-4-1-1-01-21-2011.pdf (opining that non-attorney document reviewers typically require additional training, particularly regarding the legal concept of privilege).

[47] *See Sedona Search Commentary*, *supra* note 1, at 203 ("Some litigators continue to primarily rely upon manual review of information as part of their review process. Principal rationales [include] . . . the perception that there is a lack of scientific validity of search technologies necessary to defend against a court challenge . . . ."); *see also* Thomas E. Stevens & Wayne C. Matus, *A 'Comparative Advantage' To Cut E-Discovery Costs*, NAT'L L.J. (Sept. 4, 2008), http://www.law.com/jsp/nlj/PubArticle NLJ.jsp?id=1202424251053 (describing a "general reluctance by counsel to rely on anything but what they perceive to be the most defensible positions in electronic discovery, even if those solutions do not hold up any sort of honest analysis of cost or quality").

Assumptions (1) and (2) are amenable to scientific evaluation, as is the overarching question of whether technology-assisted review can improve upon exhaustive manual review. Assumption (3) – a legal opinion – should be informed by scientific evaluation of the first two assumptions.

Assessment	Primary	Secondary	Tertiary
Primary	100%		
Secondary	42.1%	100%	
Tertiary	49.4%	42.6%	100%

Table 1: Overlap in relevance assessments by primary, secondary, and tertiary assessors for the TREC 4 Ad Hoc Task.[48]

[14] Recently, Herbert Roitblat, Anne Kershaw, and Patrick Oot studied the level of agreement among review teams using data produced to the Department of Justice ("DOJ") in response to a Second Request that stemmed from MCI's acquisition of Verizon.[49] In their study, two independent teams of professional assessors, Teams A and B, reviewed a random sample of 5,000 documents.[50] Roitblat and his colleagues reported the level of agreement and disagreement between the original production, Team A, and Team B, as a contingency matrix,[51] from which the Authors calculated overlap, as shown in Table 2.[52] The overlap between Team A and the original production was 16.3%;[53] the overlap between Team B and the original production was 15.8%;[54] and the overlap between Teams A and B was 28.1%.[55] These and other studies of overlap

[48] Voorhees, *supra* note 34, at 701 tbl.1.

[49] *See* Roitblat et al., *supra* note 7, at 73.

[50] *See id*. at 73-74.

[51] *Id.* at 74 tbl.1.

[52] *See infra* Table 2.

[53] *Id*.

[54] *Id*.

[55] *Id*.

indicate that relevance is not a concept that can be applied consistently by independent assessors, even if the information need is specified by a production request and the assessors are lawyers.[56]

Assessment	Production	Team A	Team B
Production	100%		
Team A	16.3%	100%	
Team B	15.8%	28.1%	100%

Table 2: Overlap in relevance assessments between original production in a Second Request, and two subsequent manual reviews.[57]

C. Assessor Accuracy

[15] Measurements of overlap provide little information regarding the accuracy of particular assessors because there is no "gold standard" against which to compare them.[58] One way to resolve this problem is to deem one assessor's judgments correct by definition, and to use those judgments as the gold standard for the purpose of evaluating the other assessor(s).[59]

[16] In the Voorhees study, the primary assessor composed the information need specification for each topic.[60] It may therefore be reasonable to take the primary assessor's coding decisions to be the gold standard. In the Roitblat, Kershaw, and Oot study, a senior attorney familiar with the case adjudicated all instances of disagreement between Teams A and B.[61] Although Roitblat and his colleagues sought to

[56] *See* Roitblat et al., *supra* note 7, at 73; Voorhees, *supra* note 34.

[57] The Authors derived the information in Table 2 from the Roitblat, Kershaw, and Oot study. Roitblat et al., *supra* note 7, at 74; *see supra* para. 13.

[58] Roitblat et al., *supra* note 7, at 77.

[59] *See* Voorhees, *supra* note 34, at 700.

[60] *Id.*

[61] Roitblat et al., *supra* note 7, at 74.

measure agreement,[62] it may be reasonable to use their "adjudicated results" as the gold standard. These adjudicated results deemed the senior attorney's opinion correct in cases where Teams A and B disagreed, and deemed the consensus correct in cases where Teams A and B agreed.[63] Assuming these gold standards, Table 3 shows the effectiveness of the various assessors in terms of recall, precision, and F_1.[64] Note that recall ranges from 52.8% to 83.6%, while precision ranges from 55.5% to 81.9%, and F_1 ranges from 64.0% to 70.4%.[65] All in all, these results appear to be reasonable, but hardly perfect. Can technology-assisted review improve on them?

D. Technology-Assisted Review Accuracy

[17] In addition to the two manual review groups, Roitblat, Kershaw, and Oot had two service providers (Teams C and D) use technology-assisted review processes to classify each document in the dataset as

[62] *Id.* at 72 ("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*.").

[63] *Id.* at 74.

> 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.

Id.

[64] *See infra* Table 3. Recall and precision for the secondary and tertiary assessors, using the primary assessor as the gold standard, are provided by Voorhees, *supra* note 34, at 701 tbl.2; recall and precision for Teams A and B, using the adjudicated results as the gold standard, were derived from Roitblat et al., *supra* note 7, at 74 tbl.1; F_1 was calculated from recall and precision using the formula at *supra* note 30.

[65] *See infra* Table 3.

relevant or not.[66] Unfortunately, the adjudicated results described in Part II.C. were made available to one of the two service providers, and therefore, cannot be used as a gold standard to evaluate the accuracy of the providers' efforts.[67]

Study	Review	Recall	Precision	F_1
Voorhees	Secondary	52.8%	81.3%	64.0%
Voorhees	Tertiary	61.8%	81.9%	70.4%
Roitblat et al.	Team A	77.1%	60.9%	68.0%
Roitblat et al.	Team B	83.6%	55.5%	66.7%

Table 3: Recall, precision, and F_1 of manual assessments in studies by Voorhees, and Roitblat et al. Voorhees evaluated secondary and tertiary assessors with respect to a primary assessor, who was deemed correct. The Authors computed recall, precision, and F_1 from the results reported by Roitblat et al. for Teams A and B, using their adjudicated results as the gold standard.[68]

[18] Instead, Roitblat and his colleagues reported recall, precision, and F_1 using, as an alternate gold standard, the set of documents originally produced to, and accepted by, the DOJ.[69] There is little reason to believe that this original production, and hence the alternate gold standard, was perfect.[70] The first two rows of Table 4 show the recall and precision of manual review Teams A and B when evaluated with respect to this

[66] Roitblat et al., *supra* note 7, at 74-75.

[67] *Id.* at 74 ("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 [the] 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.").

[68] Voorhees, *supra* note 34, at 701 tbl.2; Roitblat et al. *supra* note 7, at 74 tbl.1.

[69] Roitblat et al., *supra* note 7, at 74.

[70] *Id.* at 76 ("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.").

alternate gold standard.[71] These results are much worse than those in Table 3.[72] Team A achieved 48.8% recall and 19.7% precision, while Team B achieved 52.9% recall and 18.3% precision.[73] The corresponding F_1 scores were 28.1% and 27.2%, respectively – less than half of the F_1 scores achieved with respect to the gold standard derived using the senior attorney's opinion.[74]

[19] The recall and precision Roitblat, Kershaw, and Oot reported were computed using the original production as the gold standard, and are dramatically different from those shown in Table 3, which were computed using their adjudicated results as the gold standard.[75] Nevertheless, both sets of results appear to suggest the *relative* accuracy between Teams A and B: Team B has higher recall, while Team A has higher precision and higher F_1, regardless of which gold standard is applied.[76]

[20] The last two rows of Table 4 show the effectiveness of the technology-assisted reviews conducted by teams C and D, as reported by Roitblat, Kershaw, and Oot using the original production as the gold standard.[77] The results suggest that technology-assisted review Teams C and D achieved about the same recall as manual review Teams A and B, and somewhat better precision and F_1.[78] However, due to the use of the alternate gold standard, the result is inconclusive.[79] Because the

[71] *See id.* at 76 tbl.2; *infra* Table 4.

[72] *Compare supra* Table 3, *with infra* Table 4.

[73] *See infra* Table 4; *see also* Roitblat et al., *supra* note 7, at 74-76.

[74] *Compare supra* Table 3, *with infra* Table 4.

[75] *Compare supra* Table 3, *with infra* Table 4. *See generally* Roitblat et al., *supra* note 7, at 76 tbl.2.

[76] *See supra* Table 3; *infra* Table 4; Roitblat et al., *supra* note 7, at 76 tbl.2.

[77] *See infra* Table 4; *see also* Roitblat et al., *supra* note 7, at 74-75.

[78] *See infra* Table 4.

[79] *See* Roitblat et al., *supra* note 7, at 76 ("The use of precision and recall implies the availability of a stable ground truth against which to compare the assessments. Given the

improvement from using technology-assisted review, as reported by Roitblat and his colleagues, is small compared to the difference between the results observed using the two different gold standards, it is difficult to determine whether the improvement represents a real difference in effectiveness as compared to manual review.

Study	Review	Method	Recall	Precision	F_1
Roitblat et al.	Team A	Manual	48.8%	19.7%	28.1%
Roitblat et al.	Team B	Manual	52.9%	18.3%	27.2%
Roitblat et al.	Team C	Tech. Asst.	45.8%	27.1%	34.1%
Roitblat et al.	Team D	Tech. Asst.	52.7%	29.5%	37.8%

Table 4: Recall, precision, and F_1 of manual and technology-assisted review teams, evaluated with respect to the original production to the DOJ. The first two rows of this table differ from the last two rows of Table 3 only in the gold standard used for evaluation.[80]

[21] In a heavily cited study by David C. Blair and M.E. Maron, skilled paralegal searchers were instructed to retrieve at least 75% of all documents relevant to 51 requests for information pertaining to a legal matter.[81] For each request, the searchers composed keyword searches using an interactive search system, retrieving and printing documents for further review.[82] This process was repeated until the searcher was satisfied that 75% of the relevant documents had been retrieved.[83] Although the searchers believed they had found 75% of the relevant documents, their average recall was only 20.0%.[84] Despite this low rate of

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.").

[80] *Id.* at 73-76.

[81] *See* Blair & Maron, *supra* note 23, at 291.

[82] *Id.*

[83] *Id.*

[84] *Id.* at 293; *see also* Maureen Dostert & Diane Kelly, *Users' Stopping Behaviors and Estimates of Recall*, *in* SIGIR '09 PROCEEDINGS OF THE 32ND ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 820–21 (2009) (showing that most subjects in an interactive information

recall, the searchers achieved a high average precision of 79.0%.[85] From the published data,[86] the Authors calculated the average F_1 score to be 28.0% – remarkably similar to that observed by Roitblat and his colleagues for manual review.[87]

[22] Blair and Maron argue that the searchers would have been unable to achieve higher recall even if they had known there were many relevant documents that were not retrieved.[88] Researcher Gerald Salton disagrees.[89] He claims that it would have been possible for the searchers to achieve higher recall at the expense of lower precision, either by broadening their queries or by taking advantage of the relevance ranking capability of the search system.[90]

[23] Overall, the literature offers little reason to believe that manual review is perfect. But is it as complete and accurate as possible, or can it be improved upon by technology-assisted approaches invented since Blair and Maron's study?

[24] As previously noted, recent results from TREC suggest that technology-assisted approaches may indeed be able to improve on manual review.[91] In the TREC 2008 Legal Track Interactive Task, H5, a San

retrieval experiment reported they had found about 51-60% of the relevant documents when, on average, recall was only 7%).

[85] *See* Blair & Maron, *supra* note 23, at 293.

[86] *Id*.

[87] *See* Roitblat et al., *supra* note 7 at 76.

[88] *See* Blair & Maron, *supra* note 23, at 295-96.

[89] *See* Gerard Salton, *Another Look at Automatic Text-Retrieval Systems*, 29:7 COMMC'NS ACM 648, 650 (1986).

[90] *Id.* at 648-49.

[91] *See generally* Hedin et al., *supra* note 9; Oard et al., *supra* note 9.

Francisco-based legal information retrieval firm,[92] employed a user-modeling approach[93] to achieve recall, precision, and F_1 of 62.4%, 81.0%, and 70.5%, respectively, in response to a mock request to produce documents from a 6,910,192-document collection released under the tobacco Master Settlement Agreement.[94] In the course of this effort, H5 examined only 7,992 documents[95] – roughly 860 times fewer than the 6,910,192 it would have been necessary to examine in an exhaustive manual review. Yet the results compare favorably with those previously reported for manual review or keyword search, exceeding what Voorhees characterizes as a "practical upper bound" on what may be achieved, given uncertainties in assessment.[96]

[92] *See Contact Us*, H5, http://www.h5.com/about/contact.php (last visited Mar. 22, 2011); *Who We Are*, H5, http://www.h5.com/about/who_we_are.html (last visited Apr. 11, 2011).

[93] Christopher Hogan et al., *H5 at TREC 2008 Legal Interactive: User Modeling, Assessment & Measurement*, *in* NIST SPECIAL PUBLICATION: SP 500-277, THE SEVENTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2008) PROCEEDINGS (2008), *available at* http://trec.nist.gov/pubs/trec17/papers/ h5.legal.rev.pdf (last visited Mar. 23, 2011).

[94] Oard et al., *supra* note 9, at 30 tbl.15; *see also Complex Document Image Processing (CDIP)*, ILL. INST. TECH., http://ir.iit.edu/projects/ CDIP.html (last visited Apr. 11, 2011); *Master Settlement Agreement*, NAT'L ASS'N ATTORNEYS GEN. (Nov. 1998), *available at* http://www.naag.org/backpages/naag/tobacco/msa/msa-pdf/MSA%20with%20Sig%20 Pages%20and%20Exhibits.pdf; TREC 2008, *Complaint for Violation of the Federal Securities Laws, Mellon v. Echinoderm Cigarettes, Inc.*, (2008), *available at* http://trec-legal.umiacs.umd.edu/topics/8I.pdf.

[95] Hogan et al., *supra* note 92, at 8.

[96] Voorhees, *supra* note 34, at 701.

Topic	Production Request
201	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in structured commodity transactions known as "prepay transactions."
202	All documents or communications that describe, discuss, refer to, report on, or relate to the Company's engagement in transactions that the Company characterized as compliant with FAS 140 (or its predecessor FAS 125).
203	All documents or communications that describe, discuss, refer to, report on, or relate to whether the Company had met, or could, would, or might meet its financial forecasts, models, projections, or plans at any time after January 1, 1999.
204	All documents or communications that describe, discuss, refer to, report on, or relate to any intentions, plans, efforts, or activities involving the alteration, destruction, retention, lack of retention, deletion, or shredding of documents or other evidence, whether in hard-copy or electronic form.
205	All documents or communications that describe, discuss, refer to, report on, or relate to energy schedules and bids, including but not limited to, estimates, forecasts, descriptions, characterizations, analyses, evaluations, projections, plans, and reports on the volume(s) or geographic location(s) of energy loads.
206	All documents or communications that describe, discuss, refer to, report on, or relate to any discussion(s), communication(s), or contact(s) with financial analyst(s), or with the firm(s) that employ them, regarding (i) the Company's financial condition, (ii) analysts' coverage of the Company and/or its financial condition, (iii) analysts' rating of the Company's stock, or (iv) the impact of an analyst's coverage of the Company on the business relationship between the Company and the firm that employs the analyst.
207	All documents or communications that describe, discuss, refer to, report on, or relate to fantasy football, gambling on football, and related activities, including but not limited to, football teams, football players, football games, football statistics, and football performance.

Table 5: Mock production requests ("topics") composed for the TREC 2009 Legal Track Interactive Task.[97]

[97] TREC 2009, *Complaint, Grumby v. Volteron Corp.*, 14 (2009) *available at* http://*trec-legal.umiacs.umd.edu/LT09_Complaint _J_final.pdf*; *see also* Hedin et al., *supra* note 9, at 5-6.

[25] One of the Authors was inspired to try to reproduce these results at TREC 2009 using an entirely different approach: statistical active learning, originally developed for e-mail spam filtering.[98] At the same time, H5 reprised its approach for TREC 2009.[99] The TREC 2009 Legal Track Interactive Task used the same design as TREC 2008, but employed a different complaint[100] and seven new mock requests to produce documents (see Table 5) from a new collection of 836,165 e-mail messages and attachments captured from Enron at the time of its collapse.[101] Each participating team was permitted to request as many topics as they wished, however, due to resource constraints, the most topics that any team was assigned was four of the seven.[102]

[98] *See generally* Gordon V. Cormack & Mona Mojdeh, *Machine Learning for Information Retrieval: TREC 2009 Web, Relevance Feedback and Legal Tracks*, *in* NIST SPECIAL PUBLICATION: SP 500-278, THE EIGHTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2009) PROCEEDINGS (2009), *available at* http://trec.nist.gov/pubs/trec18/papers/uwaterloo-cormack.WEB.RF.LEGAL.pdf.

[99] Hedin et al., *supra* note 9, at 6.

[100] *See generally* TREC 2009, *Complaint, supra* note 97.

[101] Hedin et al., *supra* note 9, at 4; *see Information Released in Enron Investigation*, FED. ENERGY REG. COMM'N, http://www.ferc.gov/industries/electric/indus-act/wec/enron/info-release.asp (last visited Apr. 11, 2011) [hereinafter FERC]; E-mail from Bruce Hedin to Gordon V. Cormack (Aug. 31, 2009 20:33 EDT) (on file with authors) ("I have attached full list of the 836,165 document-level IDs . . . ."). The collection is available at *Practice Topic and Assessments for TREC 2010 Legal Learning Task*, U. WATERLOO, http://plg1.uwaterloo.ca/~gvcormac/treclegal09/ (follow "The TREC 2009 dataset") (last visited Apr. 18, 2011).

[102] Hedin et al., *supra* note 9, at 7; E-mail from Bruce Hedin to Gordon V. Cormack & Maura R. Grossman (Mar. 24, 2011 02:46 EDT) (on file with authors).

Team	Topic	Reviewed	Produced	Recall	Precision	F_1
Waterloo	201	6,145	2,154	77.8%	91.2%	84.0%
Waterloo	202	12,646	8,746	67.3%	88.4%	76.4%
Waterloo	203	4,369	2,719	86.5%	69.2%	76.9%
H5	204	20,000	2,994	76.2%	84.4%	80.1%
Waterloo	207	34,446	23,252	76.1%	90.7%	82.8%
	Average:	15,521	7,973	76.7%	84.7%	80.0%

Table 6: Effectiveness of H5 and Waterloo submissions to the TREC 2009 Legal Track Interactive Task.[103]

[26] Together, H5 and Waterloo produced documents for five distinct TREC 2009 topics;[104] the results of their efforts are summarized in Table 6. The five efforts employed technology-assisted processes, with the number of manually reviewed documents for each topic ranging from 4,369 to 34,446[105] (or 0.5% to 4.1% of the collection). That is, the total human effort for the technology-assisted processes – measured by the number of documents reviewed – was between 0.5% and 4.1% of that which would have been necessary for an exhaustive manual review of all 836,165 documents in the collection.[106] The number of documents produced for each topic ranged from 2,154 to 23,252[107] (or 0.3% to 2.8% of the collection; about half the number of documents reviewed). Over the five efforts, the average recall and precision were 76.7% and 84.7%,

[103] *See infra*, para. 25.

[104] *See* Hedin et al., *supra* note 9, at 7.

[105] Cormack & Mojdeh, *supra* note 98, at 6 tbl.2 (showing that Waterloo reviewed between 4,369 documents (for Topic 203) and 34,446 documents (for Topic 207); *see* E-mail from Dan Brassil to Maura R. Grossman (Dec. 17, 2010 15:21 EST) (on file with authors) ("[H5] sampled and reviewed 20,000 documents").

[106] *See* sources cited *supra* note 101.

[107] NIST Special Publication 500-277: The Seventeenth Text REtrieval Conference Proceedings (TREC 2008) http://trec.nist.gov/pubs/trec17/t17_proceedings.html Appendix: Per Topic Scores: TREC 2009 Legal Track, Interactive Task, 3 tbl.4, 4 tbl.8, 5 tbl.12, 6 tbl.16, 9 tbl.26 http://trec.nist.gov/pubs/trec18/appendices/ app09int2.pdf.

respectively; no recall was lower than 67.3%, and no precision was lower than 69.2%,[108] placing all five efforts above what Voorhees characterized as a "practical upper bound" on what may be achieved, given uncertainties in assessment.[109]

[27] Although it appears that the TREC results are better than those previously reported in the literature, either for manual or technology-assisted review, they do not include any direct comparison between manual and technology-assisted review.[110] To draw any firm conclusion that one is superior to the other, one must compare manual and technology-assisted review efforts using the same information needs, the same dataset, and the same evaluation standard.[111] The Roitblat, Kershaw, and Oot study is the only peer-reviewed study known to the Authors suggesting that technology-assisted review *may be* superior to manual review – if only in terms of precision, and only by a small amount – based on a common information need, a common dataset, and a common gold standard, albeit one of questionable accuracy.[112]

[28] This Article shows conclusively that the H5 and Waterloo efforts *are* superior to manual reviews conducted contemporaneously by TREC assessors, using the same topics, the same datasets, and the same gold standard. The manual reviews considered for this Article were the "First-Pass Assessments" undertaken at the request of the TREC coordinators for

[108] *See* Hedin et al, *supra* note 9, at 17.

[109] Voorhees, *supra* note 34, at 701.

[110] *See e.g.*, Oard et al., *supra* note 9, at 1-2.

[111] *See* Voorhees, *supra* note 35, at 356 ("The [Cranfield] experimental design called for the same set of documents and same set of information needs to be used for each [search method], and for the use of both precision and recall to evaluate the effectiveness of the search.").

[112] *See* Roitblat et al., *supra* note 7, at 76 ("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.").

the purpose of evaluating the participating teams' submissions.[113] In comparing the manual and technology-assisted reviews, the Authors used exactly the same adjudicated gold standard as TREC.[114]

III. TREC Legal Track Interactive Task

[29] TREC is an annual event hosted by NIST, with the following objectives:

- to encourage research in information retrieval based on large test collections;
- to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas;
- to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and
- to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems.[115]

Since its inception in 2006,[116] the TREC Legal Track has had the goal "to develop search technology that meets the needs of lawyers to engage in effective discovery in digital document collections."[117]

[113] Hedin et al., *supra* note 9, at 3 (describing the "First-Pass Assessment" process).

[114] *See id*. at 3-4.

[115] Text REtrieval Conference (TREC), *Overview*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/overview.html (last updated Aug. 10, 2010).

[116] *See* Jason R. Baron, *The TREC Legal Track: Origins and Reflections on the First Year*, 8 SEDONA CONF. J. 251, 253 (2007); *see also* Jason R. Baron et al., *TREC-2006 Legal Track Overview*, *in* NIST SPECIAL PUBLICATION: SP 500-272, THE FIFTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2006) PROCEEDINGS 1-2 (2006), *available at* http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf.

[30] Within the TREC Legal Track, the Interactive Task simulates the process of review of a large population of documents for responsiveness to one or more discovery requests in a civil litigation.[118] In 2008, the first year of the Interactive Task,[119] the population of documents used was the "Illinois Institute of Technology Complex Document Information Processing Test Collection, version 1.0" ("IIT CDIP"),[120] consisting of about seven million documents that were released in connection with various lawsuits filed against certain U.S. tobacco companies and affiliated research institutes.[121] A mock complaint and three associated requests for production (or topics) were composed for the purposes of the Interactive Task.[122] Participating teams were required to produce the responsive documents for one or more of the three requests.[123]

[31] The population of documents used for TREC 2009 consisted of e-mail messages and attachments that Enron produced in response to requests by FERC.[124] A mock complaint and seven associated requests for production were composed for the purposes of TREC 2009.[125] Participating teams requested as many topics as they desired to undertake, but time and cost constraints limited the number of topics that any team was assigned to a maximum of four.[126]

[117] Text Retrieval Conference (TREC), *TREC Tracks*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/tracks.html (last updated Feb. 24, 2011).

[118] *See* Oard et al., *supra* note 9, at 20.

[119] *See id.* at 2.

[120] *Id.* at 3; *see Complex Document Image Processing (CDIP)*, *supra* note 94.

[121] *See* Oard et al., *supra* note 9, at 3; *Complex Document Image Processing (CDIP)*, *supra* note 93.

[122] *See* Oard et al., *supra* note 9 at 3, 24.

[123] *Id.* at 24.

[124] *See* Hedin et al., *supra* note 9, at 4; *see also* FERC, *supra* note 101.

[125] *See* Hedin et al., *supra* note 9, at 5-6.

[126] *See id.* at 7 tbl.1.

[32] Aside from the document collections, the mock complaints, and the production requests, the conduct of the 2008 and 2009 Interactive Tasks was identical.[127] Participating teams were given the document collection, the complaint, and the production requests several weeks before production was due.[128] Teams were allowed to use any combination of technology and human input; the exact combination differed from team to team.[129] However, the size of the document population, along with time and cost constraints, rendered it infeasible for any team to conduct an exhaustive review of every document.[130] To the Authors' knowledge, no team examined more than a small percentage of the document population; H5 and Waterloo, in particular, used various combinations of computer search, knowledge engineering, machine learning, and sampling to select documents for manual review.[131]

[33] To aid the teams in their efforts, as well as to render an authoritative interpretation of responsiveness (or relevance, within the context of TREC), a volunteer *Topic Authority* ("TA") – a senior attorney familiar with the subject matter – was assigned for each topic.[132] The TA played three critical roles:

- to consult with the participating teams to clarify the notion of relevance, in a manner chosen by the teams;

[127] *See id.* at 1-2.

[128] *See* Text Retrieval Conference (TREC), *TREC-2008 Legal Track Interactive Task: Guidelines*, 8, 17 (2008), trec-legal.umiacs.umd.edu/2008InteractiveGuidelines.pdf [hereinafter *TREC-2008 Guidelines*]; *see also* E-mail from Dan Brassil to Maura R. Grossman, *supra* note 105.

[129] *TREC-2008 Guidelines*, *supra* note 128, at 4, 7; *see also* E-mail from Bruce Hedin to Gordon V. Cormack (Apr. 07, 2011 00:56 EDT) (confirming that teams were permitted to use any combination of technology and human input).

[130] *See TREC-2008 Legal Track Interactive Task: Guidelines*, *supra* note 128, at 8.

[131] *See* Hogan et al., *supra* note 9, at 5; Cormack & Mojdeh, *supra* note 98, at 6.

[132] *See* Hedin et al., *supra* note 9, at 2.

- to prepare a set of written guidelines used by the human reviewers to evaluate, after the fact, the relevance of documents produced by the teams; and

- to act as a final arbiter of relevance in the adjudication process.[133]

[34] The TREC coordinators evaluated the various participant efforts using estimates of recall, precision, and F_1 based on a two-pass human assessment process.[134] In the first pass, human reviewers assessed a stratified sample of about 7,000 documents for relevance.[135] For some topics (Topics 201, 202, 205, and 206), the reviewers were primarily volunteer law students supervised by the TREC coordinators; for others (Topics 203, 204, and 207), the reviewers were lawyers employed and supervised by professional document-review companies, who volunteered their services.[136]

[35] The TREC coordinators released the first-pass assessments to participating teams, which were invited to appeal relevance determinations with which they disagreed.[137] For each topic, the TA adjudicated the appeals, and the TA's opinion was deemed to be correct and final.[138] The gold standard of relevance for the documents in each sample was therefore:

- The same as the first-pass assessment, for any document that participants did not appeal; or

[133] *Id.* at 2-3; *see* Oard et al., *supra* note 9, at 20.

[134] Hedin et al., *supra* note 9, at 3-4.

[135] *See id.* at 12-14.

[136] *Id.* at 8.

[137] *Id.* at 3.

[138] *Id.*

- The TA's opinion, for any document that participants did appeal.

The TREC coordinators used statistical inference to estimate recall, precision, and F_1 for the results each participating team produced.[139]

[36] Assuming participants diligently appealed the first-pass assessments with which they disagreed, it is reasonable to conclude that TREC's two-pass assessment process yields a reasonably accurate gold standard. Moreover, that same gold standard is suitable to evaluate not only the participants' submissions, but also the first-pass assessments of the human reviewers.[140]

[37] Parts III.A and III.B briefly describe the processes employed by the two participants whose results this Article compares to manual review. Notably, the methods the two participants used differ substantially from those typically described in the industry as "clustering" or "concept search."[141]

A. H5 Participation

[38] At TREC 2009, H5 completed one topic (Topic 204).[142] According to Dan Brassil of H5, the H5 process involves three steps: (i) "definition of relevance," (ii) "partly-automated design of deterministic queries," and (iii) "measurement of precision and recall."[143] "Once relevance is defined, the two remaining processes of (1) sampling and query design and (2) measurement of precision and recall are conducted

[139] *Id.* at 3, 11-16.

[140] *See* Hedin et al., *supra* note 9, at 13 (describing the construction of the gold standard).

[141] *Sedona Search Commentary*, *supra* note 1, at 202-03.

[142] Hedin et al., *supra* note 9, at 6-7.

[143] E-mail from Dan Brassil to Maura R. Grossman, *supra* note 105.

iteratively – 'allowing for query refinement and correction' – until the clients' accuracy requirements are met."[144]

[39] H5 describes how its approach differs from other information retrieval methods as follows:

> It utilizes an iterative issue-focusing and data-focusing methodology that defines relevancy in detail; most alternative processes provide a reductionist view of relevance (e.g.: a traditional coding manual), or assume that different individuals share a common understanding of relevance.
>
> [H5's approach] is deterministic: each document is assessed against the relevance criteria and a relevant / not relevant determination is made. . . .
>
> [The approach] is built on precision: whereas many alternative approaches start with a small number [of] keywords intended to be broad so as to capture a lot of relevant data (with the consequence of many false positives), H5's approach is focused on developing in an automated or semi-automated fashion large numbers of deterministic queries that are very precise: each string may capture just a few documents, but nearly all documents so captured will be relevant; and all the strings together will capture most relevant documents in the collection.[145]

In the course of its TREC 2009 effort, H5 sampled and reviewed a total of 20,000 documents.[146] H5 declined to quantify the number of person-hours

[144] *Id.*

[145] *Id.* (citing Dan Brassil et al., *The Centrality of User Modeling to High Recall with High Precision Search, in* 2009 IEEE Int'l Conf. on Systems, Man, and Cybernetics, 91, 91-96.

[146] *Id*.

it expended during the seven to eight week time period between the assignment of the topic and the final submission date.[147]

B. Waterloo Participation

[40] The University of Waterloo ("Waterloo") completed four topics (Topics 201, 202, 203, and 207).[148] Waterloo's approach consisted of three phases: (i) "interactive search and judging," (ii) "active learning," and (iii) recall estimation.[149] The interactive search and judging phase "used essentially the same tools and approach [Waterloo] used in TREC 6."[150] Waterloo coupled the Wumpus search engine[151] to a custom web interface that provided document excerpts and permitted assessments to be coded with a single mouse click.[152] Over the four topics, roughly 12,500 documents were retrieved and reviewed, at an average rate of about 3 documents per minute (about 22 seconds per document; 76 hours in

[147] *Id.*; E-mail from Dan Brassil to Maura R. Grossman (Feb. 16, 2011 15:58 EST) (on file with authors).

[148] Cormack & Mojdeh, *supra* 98, at 2.

[149] *Id.* at 1-3.

[150] *Id.* at 2. *See generally*, Gordon V. Cormack et al., *Efficient Construction of Large Test Collections*, *in* SIGIR '98 PROCEEDINGS OF THE 21ST ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 282, 284 (1998).

[151] *Welcome to the Wumpus Search Engine!*, WUMPUS, http://www.wumpussearch.org/ (last visited Apr. 11, 2011).

[152] *See* Cormack & Mojdeh, *supra* note 98, at 3 & fig.2; *see also infra* Figure 1. "We used the Wumpus search engine and a custom html interface that showed hits-in-context and radio buttons for adjudication . . . . Available for reference were links to the full text of the document and to the full email message containing the document, including attachments in their native format." Cormack & Mojdeh, *supra* note 98, at 3.

total).[153] Waterloo used the resulting assessments to train an on-line active learning system, previously developed for spam filtering.[154]

[41] The active learning system "yields an estimate of the [probability] that each document is relevant."[155] Waterloo developed an "efficient user interface to review documents selected by this relevance score" (see Figure 2).[156] "The primary approach was to examine unjudged documents in decreasing order of score, skipping previously adjudicated documents."[157] The process displayed each document as text and, using a single keystroke, coded each document as relevant or not relevant.[158] Among the four topics, "[a]bout 50,000 documents were reviewed, at an average rate of 20 documents per minute (3 seconds per document)" or 42 hours in total.[159] "From time to time, [Waterloo] revisited the interactive search and judging system, to augment or correct the relevance assessments as new information came to light."[160]

[153] E-mail from Gordon V. Cormack to K. Krasnow Waterman (Feb. 24, 2010 08:25 EST) (on file with authors) (indicating that 12,508 documents were reviewed at a rate of 22 seconds per document, *i.e.,* 76.44 hours in total).

[154] Cormack & Mojdeh, *supra* note 98, at 3.

[155] *Id.* at 3.

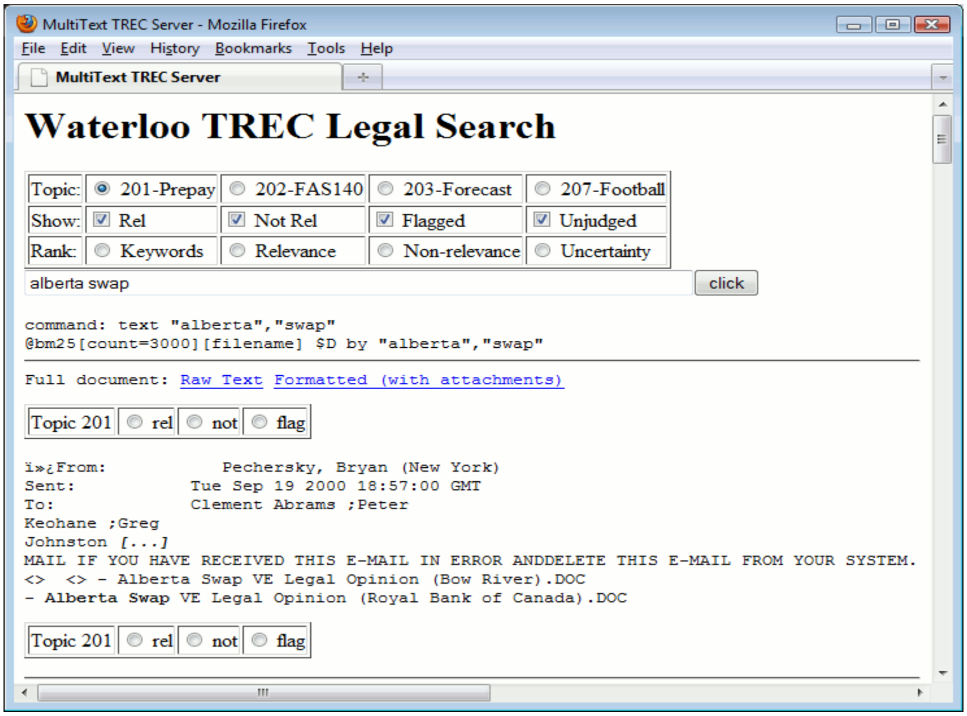
[156] *Id.*

[157] *Id*.

[158] *Id.*

[159] Cormack & Mojdeh, *supra* note 98, at 3.

[160] *Id.*


MultiText TREC Server - Mozilla Firefox
File Edit View History Bookmarks Tools Help
MultiText TREC Server
click

Waterloo TREC Legal Search

Topic:	● 201-Prepay	○ 202-FAS140	○ 203-Forecast	○ 207-Football
Show:	☑ Rel	☑ Not Rel	☑ Flagged	☑ Unjudged
Rank:	○ Keywords	○ Relevance	○ Non-relevance	○ Uncertainty

alberta swap

```
command: text "alberta","swap"
@bm25[count=3000][filename] $D by "alberta","swap"
```

Full document: Raw Text Formatted (with attachments)

Topic 201 ○ rel ○ not ○ flag

```
ï»¿From:            Pechersky, Bryan (New York)
Sent:           Tue Sep 19 2000 18:57:00 GMT
To:             Clement Abrams ;Peter
Keohane ;Greg
Johnston [...]
MAIL IF YOU HAVE RECEIVED THIS E-MAIL IN ERROR ANDDELETE THIS E-MAIL FROM YOUR SYSTEM.
<>  <> - Alberta Swap VE Legal Opinion (Bow River).DOC
- Alberta Swap VE Legal Opinion (Royal Bank of Canada).DOC
```

Topic 201 ○ rel ○ not ○ flag

Figure 1: Waterloo's interactive search and judging interface.[161]

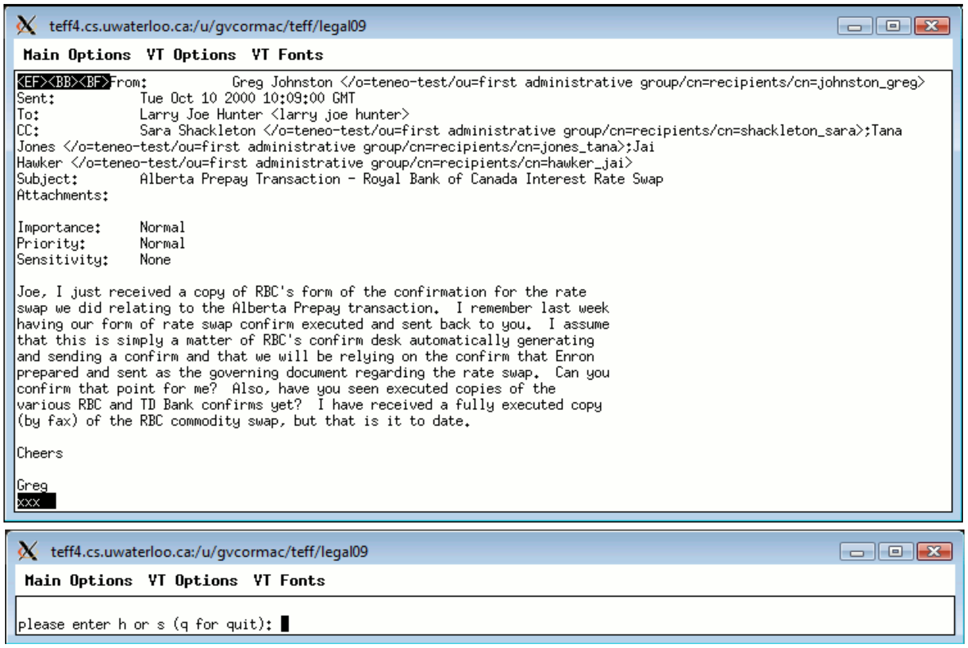
[42] The third and final phase estimated the density of relevant documents as a function of the score assigned by the active learning system, based on the assessments rendered during the active learning phase.[162] Waterloo used this estimate to gauge the tradeoff between recall and precision, and to determine the number of documents to produce so as to optimize F_1, as required by the task guidelines.[163]

[161] *Id.* at 3 & fig.2.

[162] *See id.* at 6.

[163] *Id.* at 3, 6; *see* Hedin et al., *supra* note 9, at 3.

[43] For Waterloo's TREC 2009 effort, the end result was that a human reviewed every document produced;[164] however, the number of documents reviewed was a small fraction of the entire document population (14,396 of the 836,165 documents were reviewed, on average, per topic).[165] Total review time for all phases was about 118 hours; 30 hours per topic, on average.[166]


teff4.cs.uwaterloo.ca:/u/gvcormac/teff/legal09
Main Options VT Options VT Fonts
teff4.cs.uwaterloo.ca:/u/gvcormac/teff/legal09
Main Options VT Options VT Fonts

```
<EF><BB><BF>From:          Greg Johnston </o=teneo-test/ou=first administrative group/cn=recipients/cn=johnston_greg>
Sent:           Tue Oct 10 2000 10:09:00 GMT
To:             Larry Joe Hunter <larry joe hunter>
CC:             Sara Shackleton </o=teneo-test/ou=first administrative group/cn=recipients/cn=shackleton_sara>;Tana
Jones </o=teneo-test/ou=first administrative group/cn=recipients/cn=jones_tana>;Jai
Hawker </o=teneo-test/ou=first administrative group/cn=recipients/cn=hawker_jai>
Subject:        Alberta Prepay Transaction - Royal Bank of Canada Interest Rate Swap
Attachments:

Importance:     Normal
Priority:       Normal
Sensitivity:    None

Joe, I just received a copy of RBC's form of the confirmation for the rate
swap we did relating to the Alberta Prepay transaction.  I remember last week
having our form of rate swap confirm executed and sent back to you.  I assume
that this is simply a matter of RBC's confirm desk automatically generating
and sending a confirm and that we will be relying on the confirm that Enron
prepared and sent as the governing document regarding the rate swap.  Can you
confirm that point for me?  Also, have you seen executed copies of the
various RBC and TD Bank confirms yet?  I have received a fully executed copy
(by fax) of the RBC commodity swap, but that is it to date.

Cheers

Greg
xxx
```

```
please enter h or s (q for quit): █
```

Figure 2: Waterloo's minimalist review interface.[167]

[164] *See* Cormack & Mojdeh *supra* note 98, at 6 ("the optimal strategy was to include *no* unassessed documents").

[165] *Id.*, at 6 tbl.2; E-mail from Bruce Hedin to Gordon V. Cormack, *supra* note 101 ("I have attached full list of the 836,165 document-level IDs").

[166] 118 hours is the sum of 76 hours for the interactive search and judging phase (*supra* para. 39) and 42 hours for the active learning phase (*supra* para. 41). Since Waterloo did four topics, the average effort per topic was 29.5 hours.

[167] Cormack & Mojdeh, *supra* note 98, at 4 fig.3.

IV. QUANTITATIVE ANALYSIS

[44] This Article's purpose is to refute the hypothesis that manual review is the best approach by showing that technology-assisted review can yield results that are more nearly complete and more accurate than exhaustive manual review, as measured by recall, precision, and F_1. To compare technology-assisted to manual review, the study required:

1. The results of one or more technology-assisted reviews. For this purpose, the Authors used the H5 review and the four Waterloo reviews conducted during the course of their participation in the TREC 2009 Legal Track Interactive Task.[168]

2. The results of manual reviews for the same topics and datasets as the technology-assisted reviews. For this purpose, the Authors used the manual reviews that TREC conducted on document samples for the purpose of evaluating the results that the participating teams submitted.[169]

3. A gold standard determination of relevance or nonrelevance. For this purpose, the Authors used the TREC final adjudicated assessments, for which the TA was the ultimate arbiter.[170]

[45] The Authors evaluated the results of the technology-assisted reviews and the manual reviews in exactly the same manner, using the

[168] The TREC results are available online, but use, dissemination and publication of the material is limited. Text REtrieval Conference (TREC), *Past Results*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/results.html (last visited Apr. 11, 2011) ("Individuals may request access to the protected area containing the raw results by contacting the TREC Program Manager. Before receiving access, individuals will be asked to sign an agreement that acknowledges the limited uses for which the data can be used.").

[169] Text REtrieval Conference (TREC), *Relevance Judgments and Evaluation Tools for the Interactive Task*, NAT'L INST. STANDARDS & TECH., http://trec.nist.gov/data/legal/09/evalInt09.zip (last visited Apr. 11, 2011).

[170] *Id.*; *see* Hedin et al., *supra* note 9, at 2-3.

TREC methodology and the TREC gold standard.[171] To compare the effectiveness of the reviews, this Article reports, for each topic:

1. Recall, precision, and F_1 for both the technology-assisted and manual reviews.[172]

2. The *difference* in recall, the difference in precision, and the difference in F_1 between the technology-assisted and manual reviews.[173]

3. The *significance of the difference* for each measure, expressed as P.[174] Traditionally, $P < 0.05$ is interpreted to mean that the difference is statistically significant; $P > 0.1$ is interpreted to mean that the measured difference is not statistically significant. Smaller values of P imply stronger significance; $P < 0.001$ indicates overwhelming significance.[175] The Authors used 100 bootstrap samples of paired differences to estimate the standard error of measurement, assuming a two-tailed normal distribution, to compute P.[176]

Table 7 shows recall, precision, and F_1 for the technology-assisted and manual reviews for each of the five topics, as well as the overall average for the five technology-assisted reviews and the five manual reviews. For brevity, the difference in each measure is not shown, but is easily

[171] *See* Hedin et al., *supra* note 9, at 2-5.

[172] *See id.* at 3 (reporting recall, precision, and F_1 for TREC participants); *infra* Table 7 (reporting recall, precision, and F_1 for the TREC manual reviews).

[173] *See infra* Table 7. A positive difference in some measure indicates that the technology-assisted review is superior in that measure, while a negative difference indicates that it is inferior.

[174] BÜTTCHER ET AL., *supra* note 19, at 426.

[175] *See id.*

[176] *See id.* at 412-31. "The *bootstrap* . . . is a method for simulating an empirical distribution modeling *f* (S) by sampling the sample *s*."). *Id.* at 424.

computed from the table. For example, for Topic 201, the difference in recall between Waterloo and TREC is 77.8% – 75.6% = +2.2%.

Topic	Team	Recall	Precision	F_1
201	Waterloo	(†) 77.8%	(*) 90.8%	(*) 83.8%
	TREC (Law Students)	75.6%	5.0%	9.5%
202	Waterloo	67.3%	(*) 88.0%	(*) 76.2%
	TREC (Law Students)	(†) 79.9%	26.7%	40.0%
203	Waterloo	(*) 86.5%	(*) 68.6%	(*) 76.5%
	TREC (Professionals)	25.2%	12.5%	16.7%
204	H5	(*) 76.2%	(*) 84.4%	(*) 80.1%
	TREC (Professionals)	36.9%	25.5%	30.2%
207	Waterloo	76.1%	(†) 90.7%	82.8%
	TREC (Professionals)	(†) 79.0%	89.0%	(†) 83.7%
Avg.	H5/Waterloo	(†) 76.7%	(*) 84.5%	(*) 79.8%
	TREC	59.3%	31.7%	36.0%

Table 7: Effectiveness of TREC 2009 Legal Track technology-assisted approaches (H5 and Waterloo) compared to exhaustive manual reviews (TREC). Results marked (*) are superior and overwhelmingly significant (P < 0.001). Results marked (†) are superior but not statistically significant (P > 0.1).[177]

[46] For each topic and each measure, the larger value is marked with either (*) or (†); (*) indicates that the measured difference is overwhelmingly significant (P < 0.001), while (†) indicates that it is not statistically significant (P > 0.1). As Table 7 illustrates, all of the measured differences are either overwhelmingly significant or not statistically significant.[178]

V. Qualitative Analysis

[47] The quantitative results show that the recall of the manual reviews varies from about 25% (Topic 203) to about 80% (Topic 202). That is, human assessors missed between 20% and 75% of all relevant documents.[179] Is this shortfall the result of clerical error, a

[177] For the information contained in this table, see *Past Results*, *supra* note 168; *Relevance Judgments and Evaluation Tools for the Interactive Task*, *supra* note 169. For details on the calculation and meaning of *P*, see s*upra* para. 43.

[178] *Supra* Table 7.

[179] *See supra* Table 7.

misinterpretation of relevance, or disagreement over marginal documents whose responsiveness is debatable? If the missed documents are marginal, the shortfall may be of little consequence; but if the missed documents are clearly responsive, production may be inadequate, and under Rule 37(a)(4), such a production could constitute a failure to respond.[180]

[48] To address this question, the Authors examined the documents that the TREC assessors coded as nonresponsive to Topics 204 and 207, but H5 or Waterloo coded as responsive, and the TA adjudicated as responsive. Recall from Table 5 that Topic 204 concerned shredding and destruction of documents, while Topic 207 concerned football and gambling. The Authors chose these topics because they were more likely to be easily accessible to the reader, as opposed to other topics, which were more technical in nature. In addition, lawyers employed by professional review companies assessed these two topics using accepted practices for manual review.[181]

[49] For Topic 204, 160 of the assessed documents were coded as nonresponsive by the manual reviewers and responsive by H5 and the TA;[182] Topic 207, 51 documents met these same criteria except that Waterloo and the TA made the responsiveness determinations.[183] From these numbers, the Authors extrapolated that the manual reviewers would

[180] *See* FED. R. CIV. P. 37(a)(4).

[181] *See* Hedin et al., *supra* note 9, at 8 ("The review of the samples for three of the seven Interactive topics (203, 204, and 207) was carried out by two firms that include professional document-review services among their offerings.").

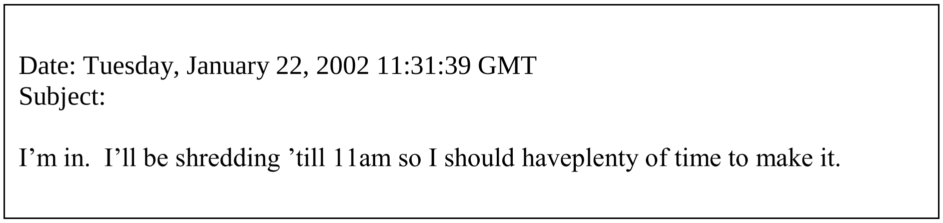
[182] The Authors identified these documents by comparing the submitted results, *see* Past *Results*, *supra* note 168 (file input.H52009.gz), the first-pass assessments, *see Relevance Judgments and Evaluation Tools for the Interactive Task*, *supra* note 169 (file qrels_doc_pre_all.txt), and the final adjudicated results, *see id.* (file qrels_doc_post_all.txt).

[183] The Authors identified these documents by comparing the submitted results, *see Past Results*, *supra* note 168 (file input.watlint.gz), the first-pass assessments, *see Relevance Judgments and Evaluation Tools for the Interactive Task*, *supra* note 169 (file qrels_doc_pre_all.txt), and the final adjudicated results, *see id.* (file qrels_doc_post_all.txt).

have missed 1,918 and 1,273 responsive documents (for Topics 204 and 207, respectively), had they reviewed the entire document collection.

[50] For each of these documents, the Authors used their judgment to assess whether the document had been miscoded due to:

- *Inarguable error*: Under any reasonable interpretation of relevance, the reviewer should have coded the document as responsive, but did not. Possible reasons for such error include fatigue or inattention, overlooking part of the document, poor comprehension, or data entry mistakes in coding the document.[184] For example, a document about "shredding" (see Figure 3) is responsive on its face to Topic 204; similarly "Fantasy Football" (see Figure 4) is responsive on its face to Topic 207.



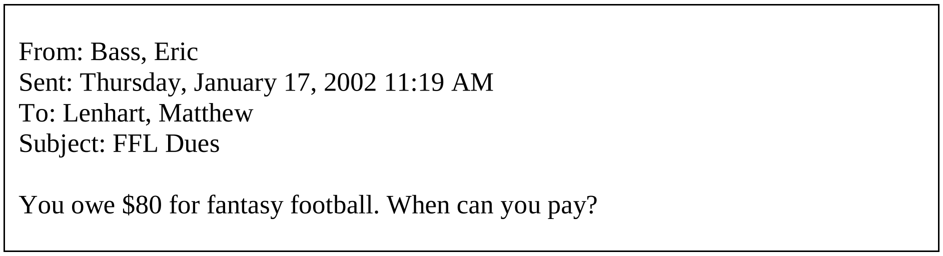
Date: Tuesday, January 22, 2002 11:31:39 GMT
Subject:

I'm in. I'll be shredding 'till 11am so I should haveplenty of time to make it.

Figure 3: Topic 204 Inarguable error. A professional reviewer coded this document as nonresponsive, although it clearly pertains to document shredding, as specified in the production request.[185]

[184] *Cf.* Jeremy M. Wolfe et al., *Low Target Prevalence Is a Stubborn Source of Errors in Visual Search Tasks*, 136 J. EXPERIMENTAL PSYCH. 623, 623-24 (2007) (showing that in visual search tasks, humans have much higher error rates when the prevalence of target items is low).

[185] *See supra* Table 5. Figure 3 is an excerpt from document 0.7.47.1449689 in the TREC 2009 dataset, *supra* note 101.



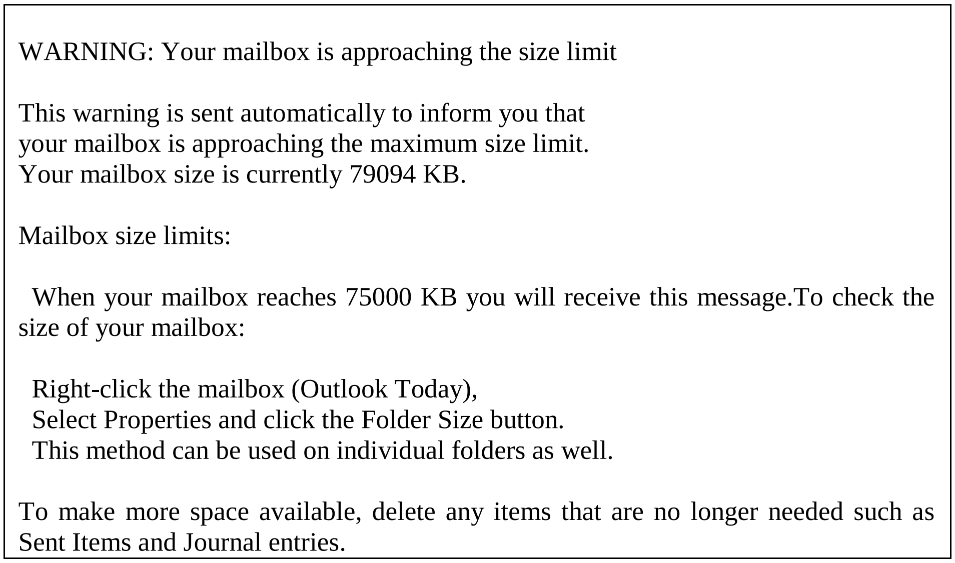
From: Bass, Eric
Sent: Thursday, January 17, 2002 11:19 AM
To: Lenhart, Matthew
Subject: FFL Dues

You owe $80 for fantasy football. When can you pay?

Figure 4: Topic 207 Inarguable error. A professional reviewer coded this document as nonresponsive, although it clearly pertains to fantasy football, as specified in the production request.[186]

- *Interpretive error*: Under some reasonable interpretation of relevance – but not the TA's interpretation as provided in the topic guidelines – an assessor might consider the document as nonresponsive. For example, a reviewer might have construed an automated message stating, "your mailbox is nearly full; please delete unwanted messages" (see Figure 5) as nonresponsive to Topic 204, although the TA defined it as responsive. Similarly, an assessor might have construed a message concerning children's football (see Figure 6) as nonresponsive to Topic 207, although the TA defined it as responsive.

[186] *See supra* Table 5. Figure 4 is an excerpt from document 0.7.47.320807 from the TREC 2009 dataset, *supra* note 101.



WARNING: Your mailbox is approaching the size limit

This warning is sent automatically to inform you that
your mailbox is approaching the maximum size limit.
Your mailbox size is currently 79094 KB.

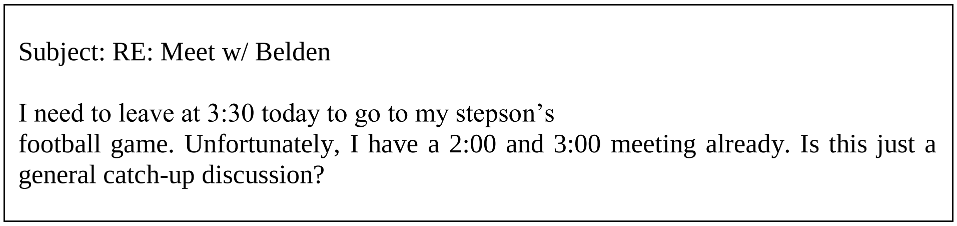
Mailbox size limits:

When your mailbox reaches 75000 KB you will receive this message.To check the size of your mailbox:

Right-click the mailbox (Outlook Today),
Select Properties and click the Folder Size button.
This method can be used on individual folders as well.

To make more space available, delete any items that are no longer needed such as Sent Items and Journal entries.

Figure 5: Topic 204 Interpretive error. A professional reviewer coded this automated message as nonresponsive, although the TA construed such messages to be responsive to Topic 204.[187]



Subject: RE: Meet w/ Belden

I need to leave at 3:30 today to go to my stepson's
football game. Unfortunately, I have a 2:00 and 3:00 meeting already. Is this just a general catch-up discussion?

Figure 6: Topic 207 Interpretive error. The reviewer may have construed a children's league football game to be outside of the scope of "gambling on football." The TA deemed otherwise.[188]

- *Arguable error*: Reasonable, informed assessors might disagree or find it difficult to determine whether or not the document met the TA's conception of responsiveness (e.g., Figures 7 and 8).

[187] *See supra* Table 5. Figure 5 is an excerpt from document 0.7.47.1048852 in the TREC 2009 dataset, *supra* note 101.

[188] *See supra* Table 5. Figure 6 is an excerpt from document 0.7.47.668065 in the TREC 2009 dataset, *supra* note 101.

Subject: Original Guarantees
Just a followup note:
We are still unclear as to whether we should continue to send original incoming and outgoing guarantees to Global Contracts (which is what we have been doing for about 4 years, since the Corp. Secretary kicked us out of using their vault on 48 for originals because we had too many documents). I think it would be good practice if Legal and Credit sent the originals to the same place, so we will be able to find them when we want them. So my question to y'all is, do you think we should send them to Global Contracts, to you, or directly the the 48th floor vault (if they let us!).

Figure 7: Topic 204 Arguable error. This message concerns *where* to store particular documents, not specifically their destruction or retention. Applying the TA's conception of relevance, reasonable, informed assessors might disagree as to its responsiveness.[189]

Subject: RE: How good is Temptation Island 2
They have some cute guy lawyers this year-but I bet you probably watch that manly Monday night Football.

Figure 8: Topic 207 Arguable error. This message mentions football, but not a specific football team, player, or game. Reasonable, informed reviewers might disagree about whether or not it is responsive according to the TA's conception of relevance.[190]

[51] When rendering assessments for the qualitative analysis, the Authors considered the mock complaint,[191] the topics,[192] and the topic-specific assessment guidelines memorializing the TA's conception of relevance, which were given to the human reviewers for reference

[189] *See supra* Table 5. Figure 7 is an excerpt from document 0.7.47.1304583 in the TREC 2009 dataset, *supra* note 101.

[190] *See supra* Table 5. Figure 8 shows an excerpt from document 0.7.6.179483 in the TREC 2009 dataset, *supra* note 101.

[191] *See generally Complaint, Grumby v. Volteron Corp.*, *supra* note 97.

[192] *Id.* at 14; Hedin et al, *supra* note 9, at 5-6.

purposes.[193] Table 8 summarizes the findings: The vast majority of missed documents are attributable either to inarguable error or to misinterpretation of the definition of relevance (interpretive error). Remarkably, the findings identify only 4% of all errors as arguable.

	Error Type			
Topic	Inarguable	Interpretive	Arguable	Total
204	98	56	6	160
207	39	11	1	51
Total	137	67	7	211
Fraction	65%	31%	4%	100%

Table 8: Number of responsive documents that human reviewers missed, categorized by the nature of the error. 65% of missed documents are relevant on their face. 31% of missed documents are clearly relevant, when the topic-specific guidelines are considered. Only 4% of missed documents, in the opinion of the Authors, have debatable responsiveness, according to the topic-specific guidelines.[194]

VI. Results and Discussion

[52] Tables 6 and 7 show that, by all measures, the average efficiency and effectiveness of the five technology-assisted reviews surpasses that of the five manual reviews. The technology-assisted reviews require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review. For F_1 and precision, the measured difference is overwhelmingly statistically significant ($P < 0.001$);[195] for recall the measured difference is not significant ($P > 0.1$).[196] These measurements provide strong evidence that the technology-assisted

[193] Text REtrieval Conference (TREC), *TREC-2009 Legal Track – Interactive Task, Topic-Specific Guidelines – Topic 204*, U. Waterloo, http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_204.pdf (last updated Oct. 22, 2009); Text REtrieval Conference (TREC), *TREC-2009 Legal Track – Interactive Task, Topic-Specific Guidelines – Topic 207*, U. Waterloo, http://plg1.cs.uwaterloo.ca/trec-assess/TopicGuidelines_207_.pdf (last updated Oct. 22, 2009).

[194] *See* sources cited *supra* note 193.

[195] *See supra* Tables 6, 7.

[196] *Id*.

processes studied here yield better overall results, and better precision, in particular, than the TREC manual review process. The measurements also suggest that the technology-assisted processes may yield better recall, but the statistical evidence is insufficiently strong to support a firm conclusion to this effect.

[53] It should be noted that the objective of TREC participants was to maximize F_1, not recall or precision, per se.[197] It happens that they achieved, on average, higher precision.[198] Had the participants considered recall to be more important, they might have traded off precision (and possibly F_1) for recall, by using a broader interpretation of relevance, or by adjusting a sensitivity parameter in their software.

[54] Table 7 shows that, for four of the five topics, the technology-assisted processes achieve substantially higher F_1 scores, largely due to their high precision. Nonetheless, for a majority of the topics, the technology-assisted processes achieve higher recall as well; for two topics, substantially higher.[199] For Topic 207, there is no meaningful difference in effectiveness between the technology-assisted and manual reviews, for any of the three measures. *There is not one single measure for which manual review is significantly better than technology-assisted review.*

[55] For three of the five topics (Topics 201, 202, and 207) the results show no significant difference in recall between the technology-assisted and manual reviews. This result is perhaps not surprising, since the recall scores are all on the order of 70% – the best that might be reasonably achieved, given the level of agreement among human assessors. As such, the results support the conclusion that technology-assisted review can achieve at least as high recall as manual review, and higher precision, at a fraction of the review effort, and hence, a fraction of the cost.

[197] *See* Hedin et al., *supra* note 9, at 15.

[198] *See supra* Tables 6, 7.

[199] *See supra* Table 7.

VII. LIMITATIONS

[56] The 2009 TREC effort used a mock complaint and production requests composed by lawyers to be as realistic as possible.[200] Furthermore, the role of the TA was intended to simulate that of a senior attorney overseeing a real document review.[201] Finally, the dataset consisted of real e-mail messages captured within the context of an actual investigation.[202] These components of the study are perhaps as realistic as might reasonably be achieved outside of an actual legal setting.[203] One possible limitation is that the Enron story, and the Enron dataset, are both well known, particularly since the Enron documents are frequently used in vendor product demonstrations.[204] Both participants and TAs may have had prior knowledge of both the story and dataset, affecting their strategies and assessments. In addition, there is a tremendous body of extrinsic information that may have influenced participants and assessors alike, including the results of the actual proceedings, commentaries,[205] books,[206]

[200] Hedin et al., *supra* note 9, at 2.

[201] *See id.*; *see also* Oard et al., *supra* note 9, at 20.

[202] *See* Hedin et al., *supra* note 9, at 4.

[203] *See id.*

[204] *See, e.g.*, John Markoff, Armies of Expensive Lawyers Replaced by Cheaper Software, N.Y. TIMES, Mar. 5, 2011, A1, available at http://www.nytimes.com/2011/03/05/science/05legal.html; *see also* E-mail from Jonathan Nystrom to Maura R. Grossman (Apr. 5, 2011 19:12 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations); E-mail from Jim Renehan to Maura R. Grossman (Apr. 5, 2011 20:06 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations); E-mail from Lisa Schofield to Maura R. Grossman (Apr. 5, 2011 18:27 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations); E-mail from Edward Stroz to Maura R. Grossman (Apr. 5, 2011 18:32 EDT) (on file with authors) (confirming use of the Enron data set for product demonstrations).

[205] *See, e.g.*, John C. Coffee Jr., *What Caused Enron?: A Capsule Social and Economic History of the 1990's*, 89 CORNELL L. REV. 269 (2004); Paul M. Healy & Krishna G. Palepu, *The Fall of Enron*, 17 J. ECON. PERSP. 3 (2003).

and even a popular movie.[207] It is unclear what effect, if any, these factors may have had on the results.

[57] In general, the TREC teams were privy to less detailed guidance than the manual reviewers, placing the technology-assisted processes at a disadvantage. For example, Topic 202 required the production of documents related to "transactions that the Company characterized as compliant with FAS 140."[208] Participating teams were required to undertake research to identify the relevant transactions, as well as the names of the parties, counterparties, and entities involved.[209] Manual reviewers, on the other hand, were given detailed guidelines specifying these elements.[210]

[58] Moreover, TREC conducted manual review on a stratified sample containing a higher proportion of relevant documents than the collection as a whole,[211] and used statistical inference to evaluate the result of reviewing every document in the collection.[212] Beyond the statistical uncertainty, there also is uncertainty as to whether manual reviewers would have had the same error rate had they reviewed the entire collection. It is not unreasonable to think that, because the proportion of relevant documents would have been lower in the collection than it was in the sample, reviewer recall and precision might have been even lower, because reviewers would have tended to miss the needles in the haystacks due to fatigue, inattention, boredom, and related human factors. This

[206] *See, e.g.*, LOREN FOX, ENRON: THE RISE AND FALL (2002); BETHANY MCLEAN AND PETER ELKIND, THE SMARTEST GUYS IN THE ROOM: THE AMAZING RISE AND SCANDALOUS FALL OF ENRON (2003).

[207] ENRON: THE SMARTEST GUYS IN THE ROOM (Magnolia Pictures 2005).

[208] Hedin et al., *supra* note 9, at 5.

[209] *See id*. at 8.

[210] *See id*. at 3.

[211] *See id*. at 12, tbl.3.

[212] *See generally id*.

sampling effect, combined with the greater guidance provided to the human reviewers, may have resulted in an overestimate of the effectiveness of manual review, and thus understated the results of the study.

[59] Of note is the fact that the appeals process involved reconsideration – and potential reversal – *only* of manual coding decisions that one or more participating teams appealed, presumably because their results disagreed with the manual reviewers' decisions.[213] The appeals process depended on participants exercising due diligence in identifying the assessments with which they disagreed.[214] And while it appears that H5 and Waterloo exercised such diligence, it became apparent to the Authors during the course of their analysis that a few assessor errors were overlooked.[215] These erroneous assessments were deemed correct under the gold standard, with the net effect of overstating the effectiveness of manual reviews, while understating the effectiveness of technology-assisted review.[216] It is also likely that the manual review and technology-assisted processes incorrectly coded some documents that were not appealed.[217] The impact of the resulting errors on the gold standard would be to overstate both recall and precision for manual review, as well as for technology-assisted review, with no net advantage to either.

[213] *See* Hedin et al., *supra* note 9 at 3, 13-14. There is no benefit, and therefore no incentive, for participating teams to appeal coding decisions with which they agree.

[214] *See id.* If participating teams do not appeal the manual reviewers' incorrect decisions, those incorrect decisions will be incorporated into the gold standard, compromising its accuracy and usefulness.

[215] Hedin et al., *supra* note 9 at 14, tbl.4 (showing that for every topic, H5 and Waterloo appealed the majority of disagreements between their results and the manual assessments).

[216] *See supra* note 214. If the manual review is incorrect, and the technology-assisted review is correct, the results will overstate the effectiveness of manual review at the expense of technology-assisted review.

[217] Given that neither the manual reviewers nor the technology-assisted processes are infallible, it stands to reason that they may occasionally agree on coding decisions that are incorrect.

[60] In designing this study, the Authors considered only the results of two of the eleven teams participating in TREC 2009, because they were considered most likely to demonstrate that technology-assisted review can improve upon exhaustive manual review. The study considered all submissions by these two teams, which happened to be the most effective submissions for five of the seven topics. The study did not consider Topics 205 and 206, because neither H5 nor Waterloo submitted results for them. Furthermore, due to a dearth of appeals, there was no reliable gold standard for Topic 206.[218] The Authors were aware before conducting their analysis that the H5 and Waterloo submissions were the most effective for their respective topics. To show that the results are significant in spite of this prior knowledge, the Authors applied Bonferroni correction,[219] which multiplies *P* by 11, the number of participating teams. Even under Bonferroni correction, the results are overwhelmingly significant.

VIII. CONCLUSION

[61] Overall, the myth that exhaustive manual review is the most effective – and therefore, the most defensible – approach to document review is strongly refuted. Technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort. Of course, not all technology-assisted reviews (and not all manual reviews) are created equal. The particular processes found to be superior in this study are both interactive, employing a combination of computer and human input. While these processes require the review of orders of magnitude fewer documents than exhaustive manual review, neither entails the naïve application of technology absent human judgment. Future work may address *which* technology-assisted review process(es) will improve *most* on manual review, not *whether* technology-assisted review *can* improve on manual review.

[218] Hedin et al., *supra* note 9, at 17-18 ("Topic 206 represents the one topic, out of the seven featured in the 2009 exercise, for which we believe the post-adjudication results are not reliable. . . . We do not believe, therefore, that any valid conclusions can be drawn from the scores recorded for this topic . . . .").

[219] *See* BÜTTCHER ET AL., *supra* note 19, at 428.

Exhibit B

The Sedona Conference Journal®

Volume 8 ❖ Fall 2007

Articles

ESI Symposium

The Sedona Conference®

Antitrust Law, Complex Litigation and Intellectual Property Rights

THE SEDONA CONFERENCE® BEST PRACTICES COMMENTARY ON THE USE OF SEARCH AND INFORMATION RETRIEVAL METHODS IN E-DISCOVERY

*A Project of The Sedona Conference® Working Group on Best Practices for Document Retention and Production (WG1), Search & Retrieval Sciences Special Project Team**

August 2007 Public Comment Version

Editor-in-Chief:
Jason R. Baron

Executive Editors:
Richard G. Braman
Kenneth J. Withers

Senior Editors:
Thomas Y. Allman
M. James Daley
George L. Paul

* With valuable input from many other WG1 members and the RFP+ Vendor Panel. This document is for educational purposes only and is not a substitute for legal advice. The opinions expressed herein are consensus views of the editors and authors, and do not necessarily represent the views of any individual participants or authors or any of the organizations to which they belong or clients they represent, nor do they necessarily represent official views of The Sedona Conference®.

Table of Contents

Preface and Acknowledgements

Welcome to another major publication in The Sedona Conference Working Group Series (the "WGS"), *Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery*. This effort is an outgrowth of our Working Group on Electronic Document Retention and Production (WG1) and represents the work of its Search and Retrieval Sciences Special Project Team, consisting of a diverse group of lawyers and representatives of firms providing consulting and legal services to the legal tech community.

The mission of the Search and Retrieval Sciences Special Project Team has been to explore the nature of the search and retrieval process in the context of civil litigation and regulatory compliance in the digital age. The goal of this Best Practices Commentary is to provide the bench and bar with an educational guide to an area of e-discovery law that we believe will only become more important over time, given the need to accurately and efficiently search for relevant evidence contained within the exponentially increasing volumes of electronically stored information (ESI) that are stored and made subject to litigation, investigations, and regulatory activities. We also understand that the subject of what constitutes best practices in this area will necessarily be subject to change, given the accelerating pace of technological developments that the law is struggling to keep up with. We hope that our efforts will assist the legal profession in this area, and we welcome all feedback at tsc@sedona.net.

This Commentary was originally conceived at the Fourth Annual Meeting of WG1 in Vancouver, B.C., in the fall of 2005. Through the efforts of many individual contributors and editors, several successive drafts were prepared for comment by the full WG1 membership in successive midyear and annual meetings. I especially want to acknowledge the contributions to the overall success of this project made by Jason R Baron, who took the lead role in editing the Commentary, along with all of the special contributions of his fellow Co-chairs of the Search and Retrieval Sciences Team, M. James Daley and Ariana J. Tadler. I also wish to acknowledge the invaluable editorial assistance provided on one or more successive drafts by senior contributing editors Thomas Y. Allman, M. James Daley, and George L. Paul, as well as the drafting contributions provided along the way by Macyl Burke, Christopher Cotton, Matthew Cohen, Conor Crowley, Sherry Harris, William Herr, Joe Looby, Stephanie Mendelsohn, Dan Regard, Herbert Roitblat, Sonya Sigler, and Stephen Whetstone. Lastly, I wish to acknowledge that many other individuals in WG1, including on the Search and Retrieval Sciences Special Project Team and the RFP+ Vendor Panel, spent time in collaborating on earlier proposals for material to be included in the Commentary. On behalf of Richard Braman, Executive Director of The Sedona Conference, I wish to thank everyone involved in devoting their time and attention during the drafting and editing process.

Kenneth J. Withers
Director, Judicial Education and Content
The Sedona Conference
June 2007

Traditional Approaches To Searching For Relevant Evidence Are No Longer Practical Or Financially Feasible

Discovery of the relevant information gathered about a topic in dispute is at the core of the litigation process.[1] However, the advent of "e-discovery" is causing a rapid transformation in how that information is gathered. While discovery disputes are not new, the huge volume of available electronically stored information poses unique challenges. Just a few years ago, a party seeking to review information for production to the other side in a "large" document review case might have been concerned with hundreds of "banker's" boxes of documents.

Today, that same amount of data might be found on a single computer hard drive.[2] Moreover, as the ability to create and store massive volumes of electronic information mushrooms, the cost to store that information inversely plummets. In 1990, a typical gigabyte of storage cost about $20,000; today it costs less than $1 dollar. As a result, more individuals and companies are generating, receiving and storing more data, which means more information must be gathered, considered, reviewed and produced in litigation. But, with billable rates for junior associates at many law firms now starting at over $200 per hour, the cost to review just one gigabyte of data can easily exceed $30,000.[3] These economic realities – *i.e.*, the huge cost differential between the $1 to store a gigabyte of data and the $30,000 to review it – act as a driver in changing the traditional attitudes and approaches of lawyers, clients, courts and litigation support providers about how to search for relevant evidence during discovery and investigations. Escalating data volumes into the billions of ESI objects, review costs, and shrinking discovery timetables, all add up to equaling the need for profound change.

As discussed below in this Commentary, just as technology has given rise to these new litigation challenges, technology can help solve them, too. The emergence of new discovery strategies, best practices and processes, as well as new search and retrieval technologies, are transforming the way lawyers litigate and, collectively, offer real promise that huge volumes of information can be reviewed faster, more accurately, and more affordably than ever before. The good news is that search and retrieval systems are improving and expanding, buoyed by a huge economic wave of activity aimed at improving the "search" experience for users generally.[4] For example, advanced forms of search techniques, including various forms of fuzzy logic, text mining and machine learning all automatically organize electronically stored information in new ways not achieved by past more familiar methods, including the simple use of "keywords" as the only automated aid to conducting manual searches. Although we are at the dawn of a new era, these new techniques hold the potential to increase both accuracy and efficiency. Through statistical sampling and validation techniques we can then confirm the accuracy of the results of either traditional or alternative forms of search, retrieval, and review.

New challenges require new solutions. This Commentary aspires to serve as a guide to enable both the bench and the bar to become more familiar with the new challenges presented by needing to search and retrieve electronically stored information. The Commentary seeks to identify ways to address those challenges, and select the best solution to maximize the just, speedy, and inexpensive determination of every action, consistent with Federal Rule of Civil Procedure 1.

[1] *Hickman v. Taylor*, 329 U. S. 495, 507 (1947)("Mutual knowledge of all the relevant facts gathered by both parties is essential to proper litigation").

[2] Here's why: One gigabyte of electronic information can generate approximately 70,000-80,000 of text pages, or 35 to 40 banker's boxes of documents (at 2,000 pages per box). Thus, a 100-gigabtye storage device (*e.g.*, a personal computer hard drive), theoretically, could hold as much as the equivalent of 3,500 to 4,000 banker's boxes of documents. By contrast, in 1990, a typical personal computer held just 200 megabytes of data - 1/500 the capacity of a typical hard drive today. Even if only 10% of a computer's available capacity today contains useful or "useable" information (as distinguished from application programs, operating systems, utilities, etc.), attorneys still would need to consider and potentially review 700,000 to 800,000 pages per each device.

[3] See Commentary, *infra*, n.13.

[4] One indication of the amount of ongoing effort and investment generally to improve search and retrieval capabilities is evidenced by the research and development spending of internet giants Google, Yahoo!, and eBay. According to published reports, Google spent $ 1.23 billion, Yahoo! spent $883 million, and eBay spent $495 million on core research and development activities in 2006. *See* Robert Hertzberg, "I.T.'s Top 84 R&D Spenders," *Baseline* (April 17, 2007), www.baselinemag.com/article2/0,1540,2114821,00.asp.

Executive Summary

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

A perfect review of the resulting volume of information is not possible. Nor is it economic. The governing legal principles and best practices do not require perfection in making disclosures or in responding to discovery requests.

The Sedona Conference® has helped establish the benchmarks governing the evolution and refinement of reasonable, good faith practices for searching intimidating amounts of data. Principle 6 of The Sedona Principles, Second Edition (2007) notes that "[r]esponding parties are best situated to evaluate the procedures, methodologies and technologies appropriate for preserving and producing their own electronically stored information," and Principle 11 amplifies the point by stating that "[a] responding party may satisfy its good faith obligation to preserve and produce relevant electronically stored information by using electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonably likely to contain relevant information."

This Commentary discusses the existing and evolutionary methods by which a party may choose to search unprecedented volumes of information. As the practice of using these "search and retrieval" technologies – the generic term we will utilize in this Commentary – continues to advance, a new understanding will evolve about what is "reasonable" under the particular circumstances of those technologies. Thus, the challenges addressed by this Commentary go beyond litigation and encompass all aspects of the search and retrieval of information from large volumes of data.

The Revolution in Discovery

Just a few years ago all information was stored on physical records such as paper. There was typically only one original document, and the number of duplicative copies and their location was generally limited. Administrative assistants, file clerks, records managers and archivists developed expertise in managing the storage, generally pursuant to pre-existing file systems. It was reasonable, and indeed relatively easy in all but the exceptional case, for the legal profession to gather and then manually review all the individual items collected as part of the discovery process prior to their production.

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

Finally, added to the search and retrieval challenge is the fact that a large percentage of the records being searched are expressed in *human language*, not just numbers. Human language is an inherently elastic, ambiguous "living" tool of enormous power. Its elasticity allows for private codes and vocabularies to exist in different subcultures in any enterprise, thus making the identification of the "words" to be searched much more challenging.

Essential Conclusions of this Commentary

This Sedona Conference® "Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery" strives to set forth state-of-the-art knowledge about meeting the challenge of searching enormous databases for relevant information, and then retrieving that information with a minimum of wasted effort.

By way of summary, we set forth our conclusions about the Problems and their Solutions, and summarize our Practical Advice which the balance of the paper articulates.

Problems

- Exponential growth in informational records is a critical challenge to the justice system.

- Electronically stored information contains human language, which challenges computer search tools. These challenges lie in the ambiguity inherent in human language and tendency of people within organizations or networks to invent their own words or communicate in code.

- The comparative efficacy of the results of manual review versus the results of alternative forms of automated methods of review remains very much an open matter of debate. Moreover, simple keyword searching, while itself a valuable tool, has certain known deficiencies.

Solutions

- Much that is useful in selecting information for production in discovery can be learned from other disciplines, including: information retrieval science; the study of linguistics; and implementation of effective management processes, to name just a few.

- Alternative search tools are available to supplement simple keyword searching and Boolean search techniques. These include using fuzzy logic to capture variations on words; using conceptual searching, which makes use of taxonomies and ontologies assembled by linguists; and using other machine learning and text mining tools that employ mathematical probabilities.

- It may be useful and appropriate to seek agreement on ways to measure and evaluate the effectiveness of the search and retrieval process. The metrics currently used in information science, such as "precision" and "recall," as well as more involved concepts are worth studying.

Practical Advice

Practice Point 1. *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.*

Practice Point 2. *Success 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.*

Practice Point 3. *The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed.*

Practice Point 4. *Parties should perform due diligence in choosing a particular information retrieval product or service from a vendor.*

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.*

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).*

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).*

Practice Point 8. *Parties and the courts should be alert to new and evolving search and information retrieval methods.*

How The Legal Community Can Contribute to The Growth of Knowledge

A consensus is forming in the legal community that human review of documents in discovery is expensive, time consuming, and error-prone. There is growing consensus that the application of linguistic and mathematic-based content analysis, embodied in new forms of search and retrieval technologies, tools, techniques and process in support of the review function can effectively reduce litigation cost, time, and error rates.

Recommendations

1. *The legal community should support collaborative research with the scientific and academic sectors aimed at establishing the efficacy of a range of automated search and information retrieval methods.*

2. *The legal community should encourage the establishment of objective benchmarking criteria, for use in assisting lawyers in evaluating the competitive legal and regulatory search and retrieval services market.*

Members of The Sedona Conference® community have and will continue to participate in collaborative workshops and other fora focused on issues involving information retrieval. The Sedona Conference® intends to remain in the forefront of the efforts of the legal community in seeking out centers of excellence in this area, including the possibility of fostering private-public partnerships aimed at focused research.

I. INTRODUCTION

The exponential growth in the volume of electronically stored information or "ESI" found in modern enterprises poses a substantial challenge to the justice system. Today, even routine discovery requests can require searches of the storage devices found on mainframes, servers, networked workstations, desktops and laptops, home computers, removable media (such as CDs, DVDs and USB flash drives), and handheld devices (such as PDAs, cell phones and iPods). Complicating things, such information is now almost always flowing robustly throughout a "network," in which it has likely been replicated, distributed, modified, linked, attached, accessed, backed-up, overwritten, deleted, undeleted, fragmented, de-fragmented, morphed and multiplied. Discovery requests for e-mail, as one common example of ESI, often require searching and retrieving information from thousands to millions or even tens of millions of individual messages, with attachments in various file formats.

The volume and complexity of this electronically stored information highlights several issues: First, whether automated search and information retrieval methods are reliable and accurate? Second, whether the legal profession has developed the skills, know-how and processes to use such automated search and retrieval methods intelligently, when applied to huge data sets, in ways that are defensible under the rules governing discovery? Yet another issue is what impact, if any, the changes to the Federal Rules governing e-discovery will have on the search and retrieval process?

The Sedona Principles, Second Edition (2007) issued by The Sedona Conference® have endorsed several highly pragmatic and relevant consensus best practices relevant to this discussion.[5]

First, Principle 6 provides that responding parties are in the best position "to evaluate the procedures, methodologies, and technologies appropriate or preserving and producing their own electronically stored information." Principle 11 expands this concept to include the use of "electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonablylikely to contain relevant information."

Second, the Commentary to Principle 11 provides that the "selective use of keyword searches can be a reasonable approach when dealing with large amounts of electronic data," and goes on to state that it "is also possible to use technology to search for 'concepts,' which can be based on ontologies, taxonomies, or data clustering approaches, for example."[6] This exploits a unique feature of electronic information – the ability to conduct fast, iterative searches for the presence of patterns of words and concepts in large document populations. The Commentary to Principle 11 also states that "[c]ourts should encourage and promote the use of search and retrieval techniques in appropriate circumstances," and suggests that "[i]deally, the parties should agree on the search methods, including search terms or concepts, to be used as early as practicable. Such agreement should take account of the iterative nature of the discovery process and allow for refinement as the parties' understanding of the relevant issues develops."[7]

Third, the Sedona Conference® has recognized that "there are now hundreds of companies offering electronic discovery services."[8] This is also true of search and information retrieval products and services for use in legal contexts – which form a subset of a burgeoning sector of the economy devoted to improving users' "search" experience. However, there remains substantial confusion as to the strengths and weaknesses of such tools. Legal practitioners have a need for guidance as to the appropriate use of search and information retrieval technologies. Such guidance can help practitioners judge the relative costs and benefits of such tools in specific cases.

This Commentary is designed to help educate the justice system – attorneys, judges and litigants alike – about "state of the art" search and retrieval tools, techniques, and methodologies, and

[5] *The Sedona Principles, Second Edition: Best Practices Recommendations & Principles for Addressing Electronic Document Production (The Sedona Conference® Working Group Series, 2007) ("The Sedona Principles, Second Edition, 2007"), available at* www.thesedonaconference.org.

[6] *Id.*, Comment 11.a.

[7] *Id.*

[8] *The Sedona Conference® Best Practices for the Selection of Electronic Discovery Vendors: Navigating the Vendor Proposal Process* (2007), *available at* http://www.thesedonaconference.org/content/miscFiles/RFP_Paper.pdf.

how they can best be used as part of an overall process to more efficiently manage discovery. This discussion includes the critically important concept of an integrated process of search and retrieval; the ability to differentiate among different search methods; how to evaluate such differences; and what questions to ask before using any particular method or product in a specific legal setting.

The legal community is familiar with keyword and natural language searches on Westlaw® and Lexis® in the context of legal research, and to a lesser extent the use of "Boolean" logic to combine keywords and "operators" (such as "AND," "OR" and "AND NOT" or "BUT NOT") that produce broader or narrower searches. However, the use of keyword, Boolean, and other search and retrieval tools to narrow information to be reviewed for production in discovery is relatively recent.[9] Moreover, to date, the relative efficacy of competing search and retrieval tools used to accomplish production review simply have not been measured. The field is wide open for the development of search and information retrieval best practices that take into account various alternative search and retrieval methods. These methods extend from improvements in basic keyword searching, to more sophisticated systems that use mathematical algorithms and various forms of linguistic techniques to help find, group and present related content.

What follows is an in-depth analysis of the problems lawyers confront in managing massive amounts of data in discovery, including how search and retrieval techniques are used in everyday practice and the key element of "process." This Commentary also provides background on the field of information retrieval and describes the world of search tools, techniques and methodologies that are currently commercially available. It also includes a "practice pointers" guide on the factors to consider in making an overall legal evaluation among different search methods, both on a conceptual and practical level. In a concluding section, the future of search and retrieval efforts is discussed. A more technical discussion of various search methodologies is included in an Appendix. Where appropriate, reference will be made to technical definitions found in the updated Sedona Glossary.

II. THE SEARCH AND INFORMATION RETRIEVAL PROBLEM CONFRONTING LAWYERS

The discovery process of today is drowning in potential sources of information. The exponential increase in volume, especially since the mid-1990s, is principally due to the impact of the PC revolution, the widespread use of email and the growth of networks. Indeed, the implication of this growth in volume is that it places at severe risk the justice system's ability to achieve the "just, speedy and inexpensive" resolution of disputes, as contemplated by Rule 1 of the Federal Rules of Civil Procedure.

The Rise of a Crushing Volume of Information in the Digital Realm

A history of the computer and information technology advances occurring since the mid-1970s is beyond the scope of this Commentary. Suffice it to say that over the last 30 years, there has been a fast-paced and widespread shift from civilization's original physical information storage technologies to new, digital information storage technologies. This "digital realm" was created by an accretion of technological advances, each built on preceding advances, which together have resulted in as fundamental a shift in the way information is shared as that which occurred in 1450 when Johannes Guttenberg invented the printing press. Included among the advances contributing to the new "digital realm" are the invention of the microchip, the development and diffusion of the

[9] There may be a role for use of some type of search and retrieval technology in discharging obligations to preserve ESI, as well as during the initial pre-review data culling or "collection" phase, in anticipation of complying with specific ESI and document requests. During the collection phase, for example, the goal is to maximize the amount of potentially relevant evidence in a subset of the greater universe of available ESI, without necessarily selecting only the more relevant information that might be the focus of a production phase review. Accordingly, parties may well end up using (and agreeing to use) differing search methods in the initial collection and later review phases of litigation. While we acknowledge that use of advanced search tools during earlier phases of litigation is truly cutting edge and worthy of future discussion, the primary focus of this Commentary will be on search tools as they are used in the review process. *See generally* Mia Mazza, Emmalena K. Quesada, and Ashley L. Sternberg, "In Pursuit of FRCP 1: Creative Approaches To Cutting and Shifting the Costs of Discovery of Electronically Stored Information," 13 RICHMOND J. LAW & TECHNOLOGY 11 (2007), at Paragraphs 53, 60, http://law.richmond.edu/jolt/v13i3/article11.pdf (discussing the use of concept searching in regard to preservation); *The Sedona Principles, Second Edition*, 2007, Comment 11.a ("Organizations should internally address search terms and other filtering criteria as soon as possible so that they can begin a dialogue on search methods as early as the initial discovery conference.").

personal computer, the spread of various types of networks linking together both computers and other networks, the rise of e-mail and its dominant use in the business world, the plunging cost of computing power and storage, and of course, the spread of the Internet and with it, the World Wide Web.[10]

By the mid-1990s, networked computers and their storage devices had created a true information-based society, with a constant flow of messages in all forms happening on a 24/7 basis. For example, studies reflect that the average U.S. worker sends and receives 100 e-mails per day. The size and nature of the attachments to these emails is also growing, with increased integration of image, audio and video files. Most recently, there has been a similar explosion in the use of instant messaging throughout business enterprises. In many organizations, the average worker maintains several gigabytes of stored data.[11] At the same time, the costs of storage have plummeted from $20,000 per gigabyte in 1990 to less than $ 1 per gigabyte today.[12] Existing technologies are only beginning to grapple with providing a viable automated means for applying records retention requirements, including the ability to implement legal holds, in the new ESI world.

Companies have continued to aggressively leverage technology to increase productivity. No one really controls how, where, how many times, and in how many forms information is stored. For example, the same Word documents can be found on e-mail attachments, local hard drives, network drives, document management systems, websites, and on all manner of removable media, such as USB flash drives, CDs, DVDs, and so on.

Discovery During the Recent Past: Manageable Amounts of Physically Stored Information

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. 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.

The cost of manual review of such volumes is prohibitive, often exceeding the damages at stake. Anecdotal reports indicate that the cost of reviewing information can easily exceed thousands of dollars per custodian, *per event*, for collection and attorney review. Litigants often cannot afford to review all available electronically stored information in the time permitted for discovery.[13] Moreover, efforts to reduce time and cost by use of "claw back"[14] provisions are problematic because of the risk of disclosure of sensitive proprietary and privileged information, as well as the risk of privilege waiver that can be imposed by substantive law, irrespective of new changes in procedural rules.

Accordingly, the conventional discovery review process is poorly adapted to much of today's litigation.[15] Lawyers of all stripes therefore have a vital interest in utilizing automated search and

[10] *See* George L. Paul and Jason R. Baron, "Information Inflation: Can the Legal System Adapt?," 13 RICHMOND J. LAW & TECHNOLOGY 10 (2007), at Paragraph 1, n.2, http://law.richmond.edu/jolt/v13i3/article10.pdf ("Organizations have thousands if not tens of thousands of times as much information within their boundaries as they did 20 years ago."); Peter Lyman and Hal R. Varian, "How Much Information," 2003, http://www.sims.berkeley.edu/how-much-info-2003.

[11] As noted *supra*, n.2, one gigabyte is equivalent in volume to between 70,000 and 80,000 pages of material. At 2000 pages per box, one gigabyte is therefore equivalent to 35 to 40 boxes of documents.

[12] Michelle Kessler, "Days of officially drowning in data almost upon us," *USA Today*, Mar. 5, 2007, *available at* www.usatoday.com/tech/news/2007-03-05-data_N.htm.

[13] Compare $1 to store a gigabyte of data with $32,000 to review it (*i.e.*, assuming one gigabyte equals 80,000 pages, and assuming that an associate billing $200 per hour can review 50 documents per hour at 10 pages in length, such a review would take 160 hours at $200/hr, or approximately $32,000).

[14] "Clawback" and "quick peek" provisions in case management agreements seek to permit large productions of electronic data little or no review, and without waiver of any claim of privilege, work product, etc. *See The Sedona Principles, Second Edition, 2007*, Comment 10.d. *See also* amended Fed. R. Civ. P. 26(f)(4), effective December 1, 2006, and accompanying Committee Note.

[15] Not all cases are equally heavy in involving electronic discovery and, from time to time, counsel may forgo production of electronically stored information and rely solely on hard copy documents.

retrieval tools where appropriate. The plaintiff's bar has a particular interest in being able to efficiently extract key information received in mammoth "document" productions, and in automated tools that facilitate the process. The defense bar has an obvious interest in reducing attendant costs, increasing efficiency, and in better risk-management of litigation (including reducing surprises). All lawyers, clients, and judges have an interest in maximizing the quality of discovery, by means of using automated tools that produce a reliable, reproducible and consistent product.

Ideally, then, judges and litigants should strive to increase their awareness of search and retrieval sciences generally, and of their appropriate application in discovery. Some technologies have been used for years to produce documents from large litigant document databases, but often without much critical analysis. The legal system may benefit from the rich body of research available through the information retrieval and library science disciplines. The discussion that follows is designed to provide a common framework and vocabulary for proper application of search and retrieval technologies in this new "age of information complexity" in the legal environment.

The Reigning Myth of "Perfect" Retrieval Using Traditional Means

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. Moreover, past research demonstrates the gap between lawyers' expectations and the true efficacy of certain types of searches. The Blair and Maron study (discussed below) reflects that human beings are less than 20% to 25% accurate and complete in searching and retrieving information from a heterogeneous set of documents (*i.e.,* in many data types and formats). The importance of this point cannot be overstated, as it provides a critical frame of reference in evaluating how new and enhanced forms of automated search methods and tools may yet be of benefit in litigation.

The Intelligent Use of Tools

Although the continued use of manual search and review methods may be indefensible in discovery involving significant amounts of electronically stored information, merely adopting sophisticated automated search tools, alone, will not necessarily lead to successful results. Lawyers must recognize that, just as important as utilizing the automated tools, is tuning the *process* in and by which a legal team uses such tools, including a close involvement of lead counsel. This may require an iterative process which importantly utilizes feedback and learning as tools, and allows for measurement of results. The time and effort spent on the front end designing a sophisticated discovery process that targets the real needs of the client must be viewed as a condition precedent to deploying automated methods of search and retrieval.

III. LAWYERS' CURRENT USE OF SEARCH AND RETRIEVAL METHODOLOGIES

Attorneys across all disciplines are generally familiar with search and retrieval methodologies based on their exposure over the past thirty years to using the automated means of searching provided by LexisNexis® and Westlaw® databases. More recently, lawyers have begun to use Google® and other Web-based search engines to hunt down information relevant to their practice. Additionally, law firms and corporate legal departments use search methods for administrative matters, such as searching data on available personnel, to support billing functions, to manage conflicts of interest, and for purposes of contact management. Many products employing search methods of various kinds exist in the legal marketplace to assist lawyers in these functions.

Current Database Tools in The Practice of Law

Litigators use automated search and retrieval tools at many stages of the litigation process. PACER and other automated means are used to uncover data on their opposing counsels' pleadings, motions, and pretrial filings in similar litigation, as well as showing how a judge has ruled in similar issues even if unreported in legal reporting services. Lawyers also use a variety of search methods with online and CD-ROM databases to dig up facts on opposing parties, witnesses, and even jury pools. At later stages of litigation, lawyers use various litigation management software applications to search through potential exhibits in connection with proceedings held in "electronic courtrooms." But until recently, litigators seldom used automated search and retrieval methods with their clients' or their opponents' growing collections of unstructured ESI.

"De-duplication" in the Processing of ESI

With the exponential increase in the amount of data subject to e-discovery, lawyers have begun to take steps towards employing automated search tools to manage the discovery process. One example of this is "de-duplication" software used to find duplicate electronic files, since ESI often consists of a massively redundant universe. For example, the same email can be copied tens or even hundreds of times in different file locations on a network or on backup media. Such de-duplication software reduces the time attorneys must spend reviewing a large document set and helps to ensure consistent classification of documents for responsiveness or privilege.[16] Increasingly, "near de-duplication" tools also are being used to assist in organizing and expediting overall document reviews, even if the technique is not used to reduce the actual number of unique documents subject to review.[17]

The Use of "Keywords"

By far the most commonly used search methodology today is the use of "keyword searches" of full text and metadata as a means of filtering data for producing responsive documents in civil discovery. For the purpose of this commentary, the use of the term "keyword searches" refers to set-based searching using simple words or word combinations, with or without Boolean and related operators (see below and Appendix for definitions). The ability to perform keyword searches against large quantities of evidence has represented a significant advance in using automated technologies, as increasingly recognized by the courts. As one United States Magistrate Judge stated, "the glory of electronic information is not merely that it saves space but that it permits the computer to search for words or 'strings' of text in seconds."[18]

Courts have not only accepted, but in some cases have ordered, the use of keyword searching to define discovery parameters and resolve discovery disputes. One court has also suggested that a party might satisfy its duty to preserve documents in anticipation of litigation by conducting system-wide keyword searching and preserving a copy of each "'hit."[19]

Because of the costs and burdens (if not impossibility) of reviewing increasingly vast volumes of electronic data, it makes sense for producing parties to negotiate with requesting parties in advance to define the parameters of discoverable information. For example, parties could agree on

[16] "De-duplication" services work to tag identical documents as duplicates by means of a "binary hash function" (which simply is a mathematical way of comparing the text of two documents -- represented in the underlying digital 1's and O's actually stored on the computer, to see if the documents are in fact perfectly alike). De-duplication by binary hash has been widely used without much notice in court opinions to date. *See Wiginton v. CB Richard Ellis, Inc.*, 229 F.R.D. 568, 571 (N.D. Ill. 2004) (referring to de-duplication process); *Medtronic Sofamor Danek Inc. v. Michelson*, 229 F.R.D. 550, 561 (W.D.Tenn., 2003) (same).

[17] "Near de-duplication" involves files that "are not hash value duplicates but are materially similar." *See* http://www.law.com/jsp/legaltechnology/roadmapArticle.jsp?id=1158014995345&hubpage=Processing.

[18] *In re Lorazepam & Clorazepate*, 300 F. Supp. 2d 43, 46 (D.D.C. 2004). *See also In re CV Therapeutics, Inc.*, 2006 WL 2458720 (N.D. Cal. Aug. 22, 2006) (court endorses employment of search terms as reasonable means of narrowing production); *J.C. Associates v. Fidelity & Guaranty Ins. Co.*, 2006 WL 1445173 (D.D.C. 2006) (requiring search of files using four specified keywords); *FTC v. Ameridebt, Inc.*, 2006 WL 6188563 (N.D. Cal. Mar. 13, 2006) ("e-mail could likely be screened efficiently through the use of electronic search terms that the parties agree upon"); *Windy City Innovations, LLC v. American Online, Inc.*, 2006 WL 2224057 (N.D. Ill. July 31, 2006) ("[k]eyword searching permits a party to search a document for a specific word more efficiently"); *Reino de Espana v. Am. Bureau of Shipping*, 2006 WL 3208579 (S.D.N.Y. Nov. 3, 2006) (court approves of e-keyword search for names and email addresses as a "targeted and focused discovery search"); *U.S. ex rel. Tyson v. Amerigroup Ill., Inc.*, 2005 WL 3111972 (N.D. Ill. Oct. 21, 2005) (referencing agreement by parties to search terms); *Medtronic Sofamor Danek, Inc., v. Michelson*, 229 F.R.D. 550 (W.D. Tenn. 2003) (court orders defendant to conduct searches using the keyword search terms provided by plaintiff); *Alexander v. FBI*, 194 F.R.D. 316 (D.D.C. 2000) (court places limitations on the scope of plaintiffs' proposed keywords to be used to search White House email).

[19] *Zubulake v.UBS Warburg, LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004); *cf. Cache La Poudre Feeds, LLC v. Land O'Lakes, Inc.*, 2007 WL 684001 (D. Colo. Mar. 2, 2007) (where court denied motion for sanctions based on an allegation that the opposing party failed to properly monitor compliance with its discovery obligations by not conducting keyword searches, court also stated that *The Sedona Principles, 2004 Edition* and *Zubulake* were not to the contrary). *See also Zakre v. Norddeutsche Landesbank Girozentrale*, 2004 WL 764895 (S.D.N.Y. Apr. 9, 2004) (court denies plaintiff's request for additional indexing of e-records, holding that defendant's production of CD-ROMS in a text searchable form was sufficient, citing to Guideline 11 of *The Sedona Principles, 2004 Edition*).

conducting a search of only files maintained by relevant or key witnesses, and/or for certain date ranges. They often can also agree to a set of key words relevant to the case. Both sides can often see the advantage to using such protocols or filters to reduce the volume of extraneous information, such as spam, routine listserv notifications, and personal correspondence, which comes with the territory of searching through electronic realms.[20]

In *Treppel v. Biovail Corp.*,[21] the defendant refused to produce documents because the plaintiff would not agree to keyword search terms. Citing to Principle 11 of the *Sedona Principles for Electronic Document Production*, the court held that the defendant was justified in using keyword search terms to find responsive documents and should have proceeded unilaterally to use its list of terms when the plaintiff refused to endorse the list. The Court held that plaintiff's "recalcitrance" did not excuse defendant's failure to produce any records and ordered the company immediately to conduct the automated search, produce the results, and explain its search protocol. Another recent case emphasized the need to confer after plaintiff was successful in obtaining a "mirror image" of data on all of defendant's computers.[22]

Issues With Keywords

Keyword searches work best when the legal inquiry is focused on finding particular documents and when the use of language is relatively predictable. For example, keyword searches work well to find all documents that mention a specific individual or date, regardless of context. However, although basic keyword searching techniques have been widely accepted both by courts and parties as sufficient to define the scope of their obligation to perform a search for responsive documents, the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages).[23]

Keyword searches identify all documents containing a specified term regardless of context, and so they can possibly capture many documents irrelevant to the user's query. For example, the term "strike" could be found in documents relating to a labor union tactic, a military action, options trading, or baseball, to name just a few (illustrating "polysemy," or *ambiguity* in the use of language). The problem of the relative percentage of "false positive" hits or noise in the data is potentially huge, amounting in some cases to huge numbers of files which must be searched to find responsive documents.[24]

On the other hand, keyword searches have the potential to miss documents that contain a word that has the same meaning as the term used in the query, but is not specified. For example, a user making queries about labor actions might miss an email referring to a "boycott" if that particular word was not included as a keyword, and a lawyer investigating tax fraud via options trading might miss an email referring to "exercise price" if that term was not specifically searched (illustrating

[20] *See generally* Kenneth J. Withers, *Computer-Based Discovery in Federal Court Litigation*, 2000 FEDERAL COURTS L. REV. 2, http://www.fclr.org/articles/2000fedctslrev2.pdf (suggesting parties adopt collaborative strategies on search protocols); *see also* R. Brownstone, *Collaborative Navigation of the Stormy e-discovery Seas*, 10 RICHMOND J. LAW & TECHNOLOGY. 53 (2004), http://law.richmond.edu/jolt/v10i5/article53.pdf (arguing that parties must agree to search terms and other selection criteria to narrow the scope to manageable data sets); *see also The Sedona Principles, Second Edition, 2007*, Comment 11.a ("For example, use of search terms could reveal that a very low percentage of files (such as emails and attachments) on a data tape contain terms that are responsive to 'key' terms. This may weigh heavily against a need to further search that source, or it may be a factor in a cost-shifting analysis. Such techniques may also reveal substantial redundancy between sources (*i.e.*, duplicate data is found in both locations) such that it is reasonable for the organization to preserve and produce data from only one of the sources.").

[21] 233 F.R.D. 363 (S.D.N.Y. 2006).

[22] *Balboa Threadworks v. Stucky*, 2006 WL 763668 (D. Kan. Mar. 24, 2006) (court orders parties to meet and confer on the use of a search protocol, including key word searching).

[23] Some case law has held that keyword searches were either incomplete or overinclusive, *see Alexander v FBI, supra*, n.18; *Quinby v. WestLB*, AG, 2006 WL 2597900 (S.D.N.Y. Sept. 5, 2006) (court narrows party's demand for 170 proposed search terms in part due to the inclusion of commonly used words).

[24] *See, e.g.*, G. Paul and J. Baron, "Information Inflation," *supra*, n.10, at Paragraph 20 (discussing potential time and cost of searching through 1 billion emails); Craig Ball, "Unlocking Keywords: How you frame your search words will shape your success," 14 No. 1 *Law Technology News* 56 (January 2007) (discussing how to improve keyword search results by use of various techniques, including eliminating "noise words" such as "law" and "legal"). *See also* Steven C. Bennett, "E-Discovery by Keyword Search," 15 No. 3 *Prac. Litigator 7* (2004).

"synonymy" or *variation* in the use of language). And of course, if authors of records are inventing words "on the fly," as they have done through history, and now are doing with increasing frequency in electronic communications, such problems are compounded.[25]

Keyword searches can also exclude common or inadvertently misspelled instances of the term (*e.g.*, "Phillip" for "Philip," or "strik" for "strike") or variations on "stems" of words (*e.g.* "striking"). So too, it is well known that even the best of optical character recognition (OCR) scanning processes introduce a certain rate of random error into document texts, potentially transforming would-be keywords into something else. Finally, using keywords alone results in a return set of potentially responsive documents that are not weighted and ranked based upon their potential importance or relevance. In other words, each document is considered to have an equal probability of being responsive upon further manual review.

More advanced keyword searches using "Boolean" operators and techniques borrowed from "fuzzy logic" may increase the number of relevant documents and decrease the number of irrelevant documents retrieved. These searches attempt to emulate the way humans use language to describe concepts. In essence, however, they simply translate ordinary words and phrases into a Boolean search argument. Thus, a natural language search for "all birds that live in Africa" is translated to something like ("bird* + liv* + Africa").

At the present time, it would appear that the majority of automated litigation support providers and software continue to rely on keyword searching. Such methods are limited by their dependence on matching a specific, sometimes arbitrary choice of language to describe the targeted topic of interest.[26] The issue of whether there is room for improvement in the rate of "recall" (as defined in the next section) of relevant documents in a given collection is something lawyers must consider when relying on simple and traditional input of keywords alone.

Use of Alternative Search Tools and Methods

Lawyers are beginning to feel more comfortable using alternative search tools to identify potentially relevant electronically stored information. These more advanced text mining tools include "conceptual search methods" which rely on semantic relations between words, and/or which use "thesauri" to capture documents that would be missed in keyword searching. Specific types of alternate search methods are set out in detail in the Appendix.

"Concept" search and retrieval technologies attempt to locate information that relates to a desired concept, without the presence of a particular word or phrase. The classic example is the concept search that will recognize that documents about Eskimos and igloos are related to Alaska, even if they do not specifically mention the word "Alaska." At least one reported case has referenced the possible use of "concept searching" as an alternative to strict reliance on keyword searching.[27]

Other automated tools rely on "taxonomies" and "ontologies" to help find documents conceptually related to the topic being searched, based on commercially available data or on specifically compiled information. This information is provided by attorneys or developed for the business function or specific industry (*e.g.*, the concept of "strike" in labor law *vs.* "strike" in options trading). These tools rely on the information that linguists collect from the lawyers and witnesses about the key factual issues in the case – the people, organization, and key concepts relating to the business as well as the idiosyncratic communications that might be lurking in documents, files, and emails. For example, a linguist would want to know how union organizers or company officials might

[25] Philosophers use colorful imagery to describe the dynamism and complexity of human language. *See, e.g.* Ludwig Wittgenstein, THE PHILOSOPHICAL INVESTIGATIONS, Section 18 (G.E.M. Anscombe, trans., The Macmillan Co., 1953 ("[T]o imagine a language is to imagine a form of life…. [L]anguage can be seen as an ancient city; a maze of little streets and squares, of old and new houses, and of houses with additions from various periods; and this surrounded by a multitude of new boroughs with straight regular streets and uniform houses").

[26] See Part IV, *infra*; see *generally*, S.I. Hayakawa, LANGUAGE IN THOUGHT AND ACTION (Harcourt 1990) (5th ed.) (such methods are inherently limited by their specific choice of language to describe a specific object or reality).

[27] *Disability Rights Council of Greater Washington v. Washington Metropolitan Transit Authority*, 2007 WL 1585452 (D.D.C. June 1, 2007) (citing to G. Paul and J. Baron, *supra*, n.10); *see generally* M. Mazza, E. Quesada, and A. Sternberg, "In Pursuit of FRCP 1: Creative Approaches To Cutting and Shifting the Costs of Discovery of Electronically Stored Information," *supra* n.9, at Paragraph 54 (discussing concept searching).

communicate plans, any special code words used in the industry, the relationships of collective bargaining units, company management structure, and other issues and concepts.

Another type of search tool relies on mathematical probabilities that a certain text is associated with a particular conceptual category. These types of machine learning tools, which include "clustering" and "latent semantic indexing," are arguably helpful in addressing cultural biases of taxonomies because they do not depend on linguistic analysis, but on mathematical probabilities. They can also help to find communications in code language and neologisms. For example, if the labor lawyer were searching for evidence that management was targeting neophytes in the union, she might miss the term "n00b" (a neologism for "newbie"). This technology, used in government intelligence, is particularly apt in helping lawyers find information when they don't know exactly what to look for. For example, when a lawyer is looking for evidence that key players conspired to violate the labor union laws, she will usually not know the "code words" or expressions the players may have used to disguise their communications.

Anecdotal information suggests that a small number of companies and law firms – particularly those that have gained significant experience in e-discovery – are using alternative search methods to either identify responsive documents (reducing expensive attorney review time) or to winnow collections to the key documents for depositions, pretrial pleadings, and trial.

The document databases that can assist lawyers in developing advanced ontologies and mathematical models are not limited to "discovery" documents. Search tools can be used in overall case management to search across pleadings, legal research, discovery responses, expert reports, and attorney work product. For example, in addition to searching discovery documents, a legal team in a labor dispute might want to search the interrogatory responses, pleadings, and depositions for all references to the concept of "strike." This is a potential growth area for vendors specializing in case management software.

Apart from the authorities listed in this section, there is still little by way of published reports or cases discussing or challenging the use of these various tools. It is only a matter of time, however, before more widespread deployment will lead to the development of a fuller body of case law.

Resistance by the Legal Profession

Some litigators continue to primarily rely upon manual review of information as part of their review process.[28] Principal rationales are: (1) concerns that computers cannot be programmed to replace the human intelligence required to make complex determinations on relevance and privilege; (2) the perception that there is a lack of scientific validity of search technologies necessary to defend against a court challenge; and (3) widespread lack of knowledge (and confusion) about the capabilities of automated search tools.

Other parties and litigators may accept simple keyword searching, yet be reluctant to use alternative search techniques. They may not be convinced that the chosen method would withstand a court challenge. They may perceive a risk that problem documents will not be found despite the additional effort; and an opposite risk that documents might be missed which would otherwise be picked up in a straight keyword search. Moreover, acknowledging that there is no one solution for all situations, they may opt for a tried-and-true lowest common denominator. Finally, litigators lack the time and resources to sort out these highly complex technical issues on a case-by-case basis.[29]

[28] *But see In re Instinet Group, Inc*., 2005 WL 3501708 (Del. Ch. Dec. 14, 2005). The court reduced plaintiffs' attorneys' fee claim by $1 million (75% of the total claim) for "obvious" inefficiencies in plaintiffs' counsel's review of paper printouts ("blowbacks") from digital files. The court stated that plaintiffs' counsel's decision to "blow back" the digital documents to paper "both added unnecessary expense and greatly increased the number of hours required to search and review the document production."

[29] *See, e.g*., R. Friedmann, http://prismlegal.com/wordpress/?cat=9 (Feb. 4, 2005)(suggesting that not one solution fits all cases); *see also id*. (July 30, 2003) (questioning the incremental value of sophisticated searching over simple searching because of the costs of implementation and need to build taxonomies and to test methodologies).

Challenging the Choice of Search Method

The challenge to a choice of search methodology used in a review prior to production can arise in one of two contexts: (1) a requesting party's objection to the unilateral use of a search method by a responding party; or (2) a court's *sua sponte* review of the use of a method or technology. Accordingly, the preferable method to reduce challenges – advocated by the proponents of the 2006 Federal Rules Amendments and experienced practitioners – is for a full and transparent discussion among counsel of the search terminology. Where the parties are in agreement on the method and a reasonable explanation can be provided, it is unlikely that a court will second-guess the process.

Absence agreement, a party has the presumption, under Sedona Principle 6, that it is in the best position to choose an appropriate method of searching and culling data. However, a unilateral choice of a search methodology may be challenged due to lack of a scientific showing that the results are accurate, complete and reliable. Since all automated search tools rely on some level of science, the challenging party may argue that the process used by the responding party is essentially an expert technology which has not been validated by subjecting it to peer review, and unbiased empirical testing or analysis.

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

As a practical matter, however, those who might object to a particular search and retrieval technology face several challenges. First, the legal system has, for decades blessed the use of keyword search tools and databases for discovery review. Second, even if such a challenge were permitted to proceed, the lack of a formally acknowledged baseline by which to measure the comparative accuracy and reliability of any search method precludes a comparison of the "new" method to traditional methods. And third, if human review or even keyword searching is the benchmark for accuracy and reliability, it arguably should not be difficult to compare the new technology favorably with either keyword searching or human review, especially when guided by a reasonable process. The discovery standard is, after all, reasonableness, not perfection.

Given the continued exponential growth in information, we would expect that a body of precedent will develop over time which references, if not critically analyzes, new and alternative search methods in use in particular legal contexts.

IV. SOME KEY TERMS, CONCEPTS AND HISTORY IN INFORMATION RETRIEVAL TECHNOLOGY

The evaluation of information retrieval ("IR") systems has, until now, largely been of greatest interest to computer scientists and graduate students in information and library science. Unlike performance benchmarking for computer hardware, there are no agreed-upon objective criteria for evaluating the performance of information retrieval systems. That is, for IR systems, the notion of effectiveness is subjective. Human judgment is ultimately the criteria for evaluating whether an IR system returns the relevant information in the correct manner. Two users may have differing needs when using an IR system. For example, one may want to find all potentially relevant documents. Another may want to correctly sort information by priority. Additionally, the subject matter and information type impact a user's information retrieval requirements.

Over the past 50 years, a large body of research has emerged concerning the evaluation of IR systems. The study of IR metrics helps quantify and compare the benefits of various search and information retrieval systems. In 1966, C.W. Cleverdon listed various "metrics" which have become the standard for evaluating IR systems within what has become known as the "Cranfield tradition."[30] Two of the metrics, *precision* and *recall*, are based on binary relationships. That is, either a document is relevant or it is not, and either a document is retrieved or it is not. Several modifications and additional metrics have been added in the IR literature since then, as the scientific field continues to add and refine techniques for measuring the efficiency of IR systems – both in terms of retrieval and also in user access to relevant information.

Measuring the effectiveness of information retrieval methods

Recall, by definition, is "an information retrieval performance measure that quantifies the fraction of known relevant documents which were effectively retrieved."[31] Another way to think about it is: out of the total number of relevant documents in the document collection, how many were retrieved correctly?

Precision is defined as "an information retrieval performance measure that quantifies the fraction of retrieved documents which are known to be relevant."[32] Put another way, how much of the returned result set is on target?

Recall and precision can be expressed by simple ratios:

$$\textbf{Recall} = \frac{\textbf{Number of responsive documents retrieved}}{\textbf{Number of responsive documents overall}}$$

$$\textbf{Precision} = \frac{\textbf{Number of responsive documents retrieved}}{\textbf{Number of documents retrieved}}$$

If a collection of documents contains, for example, 1000 documents, 100 of which are relevant to a particular topic and 900 of which are not, then a system that returned only these 100 documents in response to a query would have a precision of 1.0, and recall of 1.0.

If the system returned all 100 of these documents, but also returned 50 of the irrelevant documents, then it would have a precision 100/150 = .667 and still have a recall of 100/100 = 1.0.

If it returned only 90 of the relevant documents along with 50 irrelevant documents, then it would have a precision of 90/140 = 0.64 and a recall of 90/100 = 0.9.

Importantly for the practitioner, there is usually a trade off between precision and recall. One can often adjust a system to retrieve more documents, thereby increasing recall, but at the expense of retrieving more irrelevant documents, and thus decreasing precision.

One can cast either a narrow net and retrieve fewer relevant documents along with fewer irrelevant documents, or cast a broader net and retrieve more relevant documents, but at the expense of retrieving more irrelevant documents.[33]

[30] *See* Cyril W. Cleverdon et al., ASLIB Cranfield Research Project: Factors Determining the Performance of Indexing Systems (1966) (Vol. I , Design), *available at* http://www-nlpir.nist.gov/projects/irlib/pubs/cranv1p1/cranv1p1_index/cranv1p1_toc.html; Cyril W. Cleverdon et al., ASLIB Cranfield Research Project: Report of Cranfield II (1966) (Vol. II, Test Results), *available at* http://www-nlpir.nist.gov/projects/irlib/pubs/cranv2/cranv2_index/cranv2_toc.html; *see generally*, C.J. Van Rjiisbergen, Information Retrieval (2d ed. 1979), *available at* http://www.dcs.gla.ac.uk/Keith/Preface.html.

[31] *See* Ricardo Baeza-Yates & Berthier Ribeiro-Neto, Modern Information Retrieval 437-455 (1999) (glossary), *available at* http://www.sims.berkeley.edu/~hearst/irbook/glossary.html.

[32] *Ibid*.

[33] There are many other common metrics that are considered in IR literature, including f-measure, mean average precision and average search length. F-measure is an approximation of the cross-over point between precision and recall, which allows one to see where the compromise is between the two. Mean average precision determines the existing precision level for each retrieved relevant item. Average search length is the average position of a relevant retrieved item. Still other terms include "fallout," the ratio of the number of non-relevant items retrieved to the total number of items retrieved," and "elusion," the proportion of responsive documents that have been missed.

Measuring the Efficiency of Information Retrieval Methods

Efficiency is important to the usability of an IR system, but it does not affect the quality of the results. Efficiency is measured in two ways. The first measurement is the mean time for returning search results. This can be measured by average time to return the results or the computational complexity of the search. The second measurement is the mean time it takes a user to complete a search. This measurement is more subjective and is a function of the usability of the IR system.

The Blair and Maron Study

The leading study testing recall and precision in a legal setting was conducted by David Blair and M.E. Maron in 1985.[34] It is a classic in showing the problem caused by the rich use of human language among the many people that can be involved in a dispute, and how difficult it is to take such richness into account in a search for informational records.

Indeed, Blair and Maron found that attorneys were only about 20% effective at thinking up all of the different ways that document authors could refer to words, ideas, or issues in their case.

The case involved a San Francisco Bay Area Rapid Transit (BART) accident in which a computerized BART train failed to stop at the end of the line. There were about 40,000 documents totaling about 350,000 pages in the discovery database. The attorneys worked with experienced paralegal search specialists to find all of the documents that were relevant to the issues. The attorneys estimated that they had found more than 75% of the relevant documents, but more detailed analysis found that the number was actually only about 20%. The authors found that the different parties in the case used different words, depending on their role. The parties on the BART side of the case referred to "the unfortunate incident," but parties on the victim's side called it a "disaster." Other documents referred to the "event," "incident," "situation," "problem," or "difficulty." Proper names were often not mentioned.

As Roitblat notes, *supra*, n.34, Blair and Maron even found "that the terms used to discuss one of the potentially faulty parts varied greatly depending on where in the country the document was written. Some people called it an 'air truck,' a 'trap correction,' 'wire warp,' or 'Roman circle method.' After 40 hours of following a 'trail of linguistic creativity' and finding many more examples, Blair and Maron gave up trying to identify all of the different ways in which the document authors had identified this particular item. They did not run out of alternatives, they only ran out of time."

The Impact of Ambiguity and Variation on Precision and Recall

Since the Blair and Maron study, some further efforts have been made to study the precision/recall issues in a legal discovery context, some of which have been performed by members of The Sedona Conference®.[35] This field requires further study.

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

Ambiguity refers to the tendency of words and expressions to have different meanings when used in different contexts. These contexts are "referential variants" or *variation*. If one and only one word or expression is found in only one and only one context, it would present no ambiguity and no

[34] David C. Blair & M.E. Maron, "An evaluation of retrieval effectiveness for a full-text document-retrieval system," *Communications of the ACM* 289 (1985). The discussion that follows of the Blair and Maron study is drawn directly from Herbert L. Roitblat, "Search and Information Retrieval Science," 8 *Sedona Conf. J. at* 225 (2007).

[35] *See, e.g.*, Anne Kershaw, "Automated Document Review Proves Its Reliability," DIGITAL DISCOVERY & E-EVIDENCE, Nov. 2005, at 10, 10-12 (client-sponsored private study); Howard Turtle, "Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance," 1994 PROCEEDINGS OF THE 17TH ANNUAL INTERNATIONAL ACM SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL 212-220 (using structured caselaw in Westlaw databases); *see also* Text REtrieval Conference, http://trec.nist.gov/, discussed *infra* Part VII.C.

variation. A search for that term would retrieve all of the documents in which the term appeared, and all of the documents would be relevant. While there may not be an exact mathematical comparison, generally speaking, the lower the variation in the contexts, the lower the likely overall recall, and the lower the ambiguity of the search term, the better the precision of the result.

But as the Blair and Maron study demonstrates, human language is highly ambiguous and full of variation. In the years since Blair and Maron, the IR community has been engaged in research and development of methods, tools, and techniques that compensate for endemic ambiguity and variation in human language, and thus maximize the recall and precision of searches.

V. BOOLEAN AND BEYOND: A WORLD OF SEARCH METHODS, TOOLS AND TECHNIQUES

In the twenty years since the Blair and Maron study, a variety of new search tools and techniques have been introduced to help find relevant information and to help weed out irrelevant information. Understanding these various tools and methods is critical. All automated methods are not created equal, and do not perform the same function and task. It is important to know what each methodology does when it is used alone or in conjunction with other methodologies.

Clearly, different search methods have different functions and values in different circumstances. There is no one best system for all situations, a key fact for practitioners learning the technique of search and retrieval technology.

A more detailed description of search methods and techniques is set out in the Appendix. These methods can be grouped into three broad categories, but there are hybrid and cross-cutting approaches that defy easy placement in any particular "box."

Keywords and Boolean Operators

First, there are *keyword based methods*, ranging from the simple use of keywords alone, to the use of strings of keywords with what are known as "Boolean operators" (including AND, OR, "AND NOT" or "BUT NOT").

Statistical Techniques

Second, there are a variety of *statistical techniques*, which analyze word counts (how many times the same keyword will appear in a document, or will appear near other keywords). One such approach is called "Bayesian," derived from a famous mathematical theorem. Querying the data set using combinations of one or more of these types of Bayesian methods may well result in returning a broader slice of the data than merely using a simple keyword search, or a keyword search with Boolean operators.

Categorizations of Data Sets

Third, there are other techniques depending on *categorizations of the entire data* set with various methodologies heavily reliant on setting up (*i.e.*, coming to a consensus on) a *thesaurus, taxonomy* or "*ontology*" of related words or terms. These techniques can be used to categorize the entire data set into specified categories all at once – or continually, as more data is added to the data set.

However, data sets generally need to be indexed to use any of the latter alternative methodologies – where the indexing will take more time depending on what one indexes (*e.g.*, indexing all of the data will take substantially longer than indexing selected coded fields).

There are a variety of indexing tools, some of which are available as open source tools. Indexing structured data may take less time than indexing data in an unstructured form. Indexing a set number of structured fields (*i.e.* coded data) will be much faster because only those designated fields are indexed. Indexing an unstructured data set is time consuming because of the need to index all the *words* (except for and, a, the, or other common words). Knowing what is being indexed will be important to set expectations in terms of timing and making the data useful for querying or review.

Alternative search methods to keywords can, in some instances, free the user from having to guess, for every document, what word the author might have used. For example, there are more than 120 words that could be used in place of the word "think" (*e.g.*, guess, surmise, anticipate). As the Blair and Maron study shows, people coming in after the fact are actually very poor at guessing the right words to use in a search – words that find the documents a person is looking for without overwhelming the retrieval with irrelevant documents. In light of this fact, alternative search methods may serve to help to organize large collections of documents in ways that people have trouble doing.

Using a thesaurus, taxonomy, or ontology generally gives the results one would expect, because these systems explicitly incorporate one's expectations about what is related to what. They are most useful when one has (or can buy) a good idea of the conceptual relations to be found in one's documents – or one has the time and resources needed to develop them. Clustering, Bayesian classifiers, and other types of systems have the power to discover relationships in the text that might not have been anticipated. This means that one gets unexpected results from time to time, which can be of great value, but can also be somewhat disconcerting (or even wrong). An example: after training on a collection of medical documents, one of these systems learned that Elavil and Klonopin were related (they are both anti-anxiety drugs). A search for Elavil turned up all the documents that contained that word, along with documents containing the word "Klonopin" even without the word "Elavil."

Such systems can discover the meaning of at least some acronyms, jargon, and code words appropriate to the context of the specific document collection. No one has to anticipate their usage in all possible relational contexts; the systems, however, can go help to derive them directly from the documents processed.

Finally, none of these systems is magical. Language is sometimes shared just between two people, who have invented a shorthand or code. All tools require common sense, based on a thought-out approach. Some techniques may be difficult to understand to those without technical backgrounds, but they need not be mysterious. If a vendor will not explain how a system works, it is most likely because of ignorance. Ask for someone who can provide an explanation.

There is no magic to the science of search and retrieval: only mathematics, linguistics, and hard work. If lawyers do not become conversant in this area, they risk surrendering the intellectual jurisdiction to other fields.

VI. PRACTICAL GUIDANCE IN EVALUATING THE USE OF AUTOMATED SEARCH AND RETRIEVAL METHODS

Practice Point 1. 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.

For the reasons articulated in prior sections, the demands placed on practitioners and parties in litigation and elsewhere increasingly dictate that serious consideration be given to the use of automated search and retrieval methods in a wide variety of cases and contexts. Particularly, but not exclusively, in large and complex litigation, where discovery is expected to encompass hundreds of thousands to hundreds of millions of potentially responsive electronic records, there is no reasonable possibility of marshalling the human labor involved in undertaking a document by document, manual review of the potential universe of discoverable materials. This is increasingly true both for parties responding to a discovery request, and for parties who propound discovery only to receive a massive amount of material in response. Where the infeasibility of undertaking manual review is acknowledged, utilizing automated search methods may not only be reasonable and valuable, but necessary.

Even in less complex settings, sole reliance on manual review may nevertheless be an inefficient use of scarce resources. This is especially the case where automated search tools used on the front end of discovery may prove to be useful in a variety of ways, including for sampling, categorizing or grouping documents in order to facilitate later manual review.

Of course, the use of automated search methods is not intended to be mutually exclusive with manual review; indeed, in many cases, both automated and manual searches will be conducted: with initial searches by automated means to cull down a large universe of material to more manageable size, followed by a secondary manual review process. So too, while automated search methods may be used to find privileged documents out of a larger set, it remains the case that the majority of practitioners still will rely on manual review processes to identify the bases for privilege to be asserted for each document.

Practice Point 2. Success 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.

As discussed above, the decision to employ an automated search method or technology cannot be made in a vacuum, on the assumption that the latest "tool" will solve a discovery obligation. Rather, to maximize the chances of success in terms of finding responsive documents, a well-thought out strategy capitalizing on "human knowledge" available to a party should be put into action at the earliest opportunity. This knowledge can take many forms.

First, an evaluation of the legal setting a party finds itself in is of paramount importance, since the nature of the lawsuit or investigation, the field of law involved, and the specific causes of action under which a discovery obligation arises must all be taken into account. For example, keyword searches alone in highly technical patent cases may prove highly efficacious. In other types of cases, including those with broad causes of action and involving subjective states of intent, a practitioner should consider alternative search methods.

Second, in any legal setting involving consideration of automated methods for conducting searches, counsel and client should perform a "relevance needs analysis," to first define the target universe of documents that is central to the relevant causes of action. This would include not only assessing relevant subject areas, and "drilling down" with as much specificity as possible, but also analyzing the parties who would be the "owners" of relevant data. Time and cost considerations must also be factored in, including budgeting for human review time. These practice points apply whether your client is a defendant and holds a universe of potentially discoverable data, or your client is a plaintiff party who is expecting to receive similarly massive data in response to requests for documents.

Practice Point 3. The choice of a specific search and retrieval method will be highly dependent on the specific legal context in which it is to be employed.

The choice of a search and retrieval method for a given situation depends upon a number of factors.

For example, a search method that eliminates false positive "noise" (achieving high levels of precision) may not yield the highest number of relevant documents. In other cases, such as sampling, a search method will be graded on its ability to measure statistical significance of the occurrence of a particular word or concept. There are a number of overarching factors that lawyers should consider in evaluating the use of particular search and retrieval methods in particular settings.

First, the "heterogeneity" of the overall relevant universe of electronically stored information is a significant factor. Electronically stored information that is potentially relevant may be found in multiple locations and in a variety of forms, including structured and unstructured active computer environments, removable media, backup tapes, and the variety of email applications and file formats. In some cases, information that provides historical, contextual, tracking or managerial insight (such as metadata) may be relevant to a specific matter and demand specialized data mining search tools. Yet in other cases, it will be irrelevant.

Next, the volume and condition of the electronically stored information, and the extent to which electronically stored information is contained within static or dynamic electronic applications is relevant to the decisions made by the advocate or investigator.

Third, the time it will take to use a particular search and information retrieval method and its cost, as compared to other automated methods or human review, must be considered.

Fourth, the goals of the search are a factor (*e.g.* capturing or finding as many responsive documents as possible regardless of time and cost vs. finding responsive documents as efficiently as possible, *i.e.*, with the least number of nonresponsive documents). In other words, one must consider the desired trade off between recall and precision. Given the particular setting, the party seeking to employ one or more search methods should assess the relative importance in that setting of finding responsive electronically stored information versus the importance of eliminating non-responsive data. Depending on this assessment, one or more alternative search methodologies may prove to be a better match in the context of a particular task.

Fifth, one must consider the skills, experience, financial and practical logistical constraints of the representatives of the party making the selection (attorneys, litigation support staff, vendors).

Sixth, there is the status of electronic discovery in the matter, including the extent to which activities including preservation and collection are occurring in addition to processing and/or attorney review.

Seventh, one must investigate published papers supporting the reliability of the search and information retrieval method for particular types of data, or in particular settings.

Practice Point 4. Parties should perform due diligence in choosing a particular information retrieval product or service from a vendor.

The prudent practitioner should ask questions regarding search and retrieval features and the specific processing and searching rules that are applied to such features. Some tools are fully integrated into a vendor's search and review system, whereas others are "stand alone" tools that may be used separately from the review platform. It is essential not only to understand how the various tools function, but also to understand how the tools fit within the overall workflow planned for discovery. A practitioner should inquire as to what category or categories the specific tool fits into, how it functions, and what third party technology lies behind the tool.

It is also essential that specific methods or tools be made understandable to the court, opposing parties, and your own client. How data is captured and indexed (and how long it takes to build an index) also may affect a decision on use: it is therefore important to understand how a particular system deals with rolling input and output over time, in terms of its flexibility. The ability to perform searches across metadata, to search across multiple indices or stores of data, to search embedded data, to refine search results (nested searches), to save queries, to capture duplicates and perform de-duplication, to trace email threads, and to provide listings of related terms or synonyms, are all examples of the kind of specific functional requirements that should be inquired about.

Other types of due diligence inquiries may involve administrative matters (*e.g.*, understanding maintenance and upkeep, additional charges, system upgrades, availability of technicians, system performance), quality control issues (*e.g.*, prior testing of the method or tool in question; how databases and dictionaries supporting concept searching were populated; how strong is the application development group of the provider), and, finally, any relevant licensing issues, involving proprietary software or escrow agreements with third parties.

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.

Just as with past practice involving manual searches through traditional paper document collections, there is no requirement that "perfect" searches will occur – only that lawyers and parties act reasonably in the good faith performance of their discovery and legal obligations. From decades of information retrieval research, we know that a 100% rate of recall, *i.e.*, the ability to retrieve *all* responsive documents from a given universe of electronic data, is an unachievable goal. As discussed in prior sections, the richness of human language, with its attendant elasticity, results in all present day automated search methods falling short.

It is also important to recognize that there will be a measure of statistical variation associated with alternative search methods, *i.e.*, some responsive documents will be found by one search method while being missed by others. Even the same search method (such as one based on statistical properties of how words appear in the data set), may return different results if new documents are added to the searched universe.

Particularly in the context of a large data set, a search method should be judged by its overall results (such as using average measures of recall and precision), rather than being judged by whether it produces the identical document set as compared with a different technique. One possible benchmark to employ when considering use of an alternative search method is to compare the results of such a search against a similar search utilizing keywords and Boolean operators alone.

However, it is important not to compare "apples with oranges." Given the present state of information science, it would be a mistake to assume that one search method will work optimally across all types of possible inquiries or data sets (*e.g.*, what works well in finding word processing documents in a given proprietary format may not be as optimal for finding information in structured databases, or in a collection of scanned images). This is another area where, consistent with the above principles, a good deal of thought should be given at the outset to the precise problem, in terms of its scope and relevancy considerations, before committing to a particular search method.

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)

The *Treppel* decision and other recent case law indicates that courts are becoming more comfortable with addressing search and retrieval issues, particularly in the context of blessing or ordering parties to share information that would lead to the development of more refined search protocols. The fact that some courts have waded into these issues demonstrates how rapidly the law has been evolving even in advance of the 2006 amendments to the Federal Rules of Civil Procedure.[36]

Under newly modified Rule 26(f), the parties' initial planning is expected to address "[a]ny issues relating to disclosure or discovery of electronically stored information," as well as "[a]ny issues relating to preserving discoverable information." These initial discussions on preservation and production easily should encompass a specific discussion on search methods and protocols to be employed by one or both parties. While disclosure of these methods and protocols is not mandated or legally required under this rule, the advantages of collaborating should strongly be considered.

[36] *See* Kenneth J. Withers, "The December 2006 Amendments to the Federal Rules of Civil Procedure," 4 NW. J. OF TECH. & INTELL. PROP. 171 (2006), *available at* http://www.law.northwestern.edu/journals/njtip/v4/n2/3 (what "probably strikes the reader [of *Treppel*] as matter-of-fact, sensible, and routine, would have been extraordinary a scant six years ago, when the last major revision of the discovery rules went into effect [in 2000])."

Reaching an early consensus on the scope of searches has the potential to minimize the overall time, cost, and resources spent on such efforts, as well as minimizing the risk of collateral litigation challenging the reasonableness of the search method employed.[37]

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).

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

Practice Point 8. Parties and the courts should be alert to new and evolving search and information retrieval methods.

What constitutes a reasonable search and information retrieval method is subject to change, given the rapid evolution of technology. The legal community needs to be vigilant in examining new and emerging techniques and methods which claim to yield better search results. In particular settings, lawyers should endeavor to incorporate evolving technological progress at the earliest opportunity in the planning stages of discovery or other legal setting involving search and retrieval issues.

VII. FUTURE DIRECTIONS IN SEARCH AND RETRIEVAL SCIENCE

What prospects exist for improving present day search and retrieval methodologies? And how can lawyers play a greater role in working with the information retrieval research community based on a shared interest in how to improve the accuracy and efficiency of information retrieval?

A. Harnessing the Power of Artificial Intelligence (AI)

A statement from page 36 of The Sedona Conference®, Navigating The Vendor Proposal Process (2007 ed.), under the general heading "Advanced Search and Retrieval Technology," bears repetition here: "Technology is developing that will allow for electronic relevancy assessments and subject matter, or issue coding. These technologies have the potential to dramatically change the way electronic discovery is handled in litigation, and could save litigants millions of dollars in document review costs. Hand-in-hand with electronic relevancy assessment and issue coding, it is anticipated that advanced searching and retrieval technologies may allow for targeted collections and productions, thus reducing the volume of information involved in the discovery process."

The growing enormity of data stores, the inherent elasticity of human language, and the unfulfilled goal of computational thinking to approximate the ability and subtlety of human language behavior all present steep challenges to the AI community in developing optimal search and retrieval techniques.

But the future continues to hold promise. Not only is there the possibility of applying sophisticated artificial intelligence means to data mining of traditional texts, but looming immediately on the horizon are new and better approaches to image and voice pattern recognition. Clearly, all forms of data stored in corporations and institutions will be fair game in terms of being within the scope of future information demands in legal settings.

Finding information on the Web sometimes is easier than finding documents on one's own hard drive. The post-Google burst of interest in building better search engines for the Web can only

[37] *See* G. Paul and J. Baron, *Information Inflation, supra* n.10, at Paragraphs 50-55 (discussing an iterative collaboration process that includes adoption of multiple "meet and confers" to discuss and refine preliminary search results).

help lead to new and better search techniques applied in more well-defined contexts, such as corporate and institutional intranets and data stores.

A recent "2020 Science" report issued by Microsoft anticipates the near-term development of "novel data mining technologies and novel analysis techniques," including "active learning" in the form of "autonomous experimentation" and "artificial scientists," in replacement of "'traditional' machine learning techniques [that] have failed to bring back the knowledge out of the data."[38] Beyond the short-term horizon, scientists are expected to embrace emergent technologies including the use of genetic algorithms, nanotechnology, quantum computing, and a host of other advanced means of information processing. The field of future AI research in the specific domain of search and retrieval is wide open.

B. The Role of Process in the Search and Retrieval Challenge

Every search and retrieval technology has its own methodology to ensure the technology works properly – a set of instructions outlining the workflow for the tool. How well these methods are applied significantly impacts the performance, and therefore, the results generated by the technology. This is where process comes in. Process functions to provide order and structure by setting guidelines and procedures designed to ensure that a technology performs as intended. Effectively applied, process will then drive the consistent and predictable application of the search and retrieval technology. The results derived from the consistent and predictable application of search and retrieval tools will then establish the technology's credibility and value.

The Important Nature of Process

A process is a considered series of events, acts or operations leading to a result or an effect. A process, like a technology, is a "tool" that can be used to assist in completing a task. The use of a well-defined and controlled process promotes consistency, reliability and predictability of the results and ensures the efficient use of the resources required to produce them. As such, a process does not find the answer to, or attain the objective of a task on its own. Process, no matter how well designed and executed can not replace the exercise of judgment, however, process promotes the exercise of judgment by ensuring that the most accurate and reliable information is available when making decisions. In the search and retrieval context, this means the availability of consistent and reliable information to assist parties in making informed decisions.

The use of process promotes consistency by establishing a defined approach to a task. The resulting consistency promotes reliability and predictability. Reliability and predictability allow for better planning, performance and cost management. All together, risk is reduced and confidence is promoted.

Search and retrieval should be visualized as a process which enables a party to distinguish potentially discoverable information from among a broader set of electronic data for purposes of production. It consists of several process steps that take place in the context of a particular search and retrieval technology. Because the application of process is flexible, it can be used to address unique conditions that might be associated with a technology, such as where the use of a search and retrieval technology itself creates issues. For example, the use of search and retrieval technologies to address significant volumes of information may not address all problems: as review volumes increase, even with carefully crafted and tested search criteria, the likelihood of being swamped by false positives increases greatly. Additionally, greater volume increases the likelihood of the omission of some relevant documents. By developing and implementing process steps that consistently address these issues, their impact can be diminished and the reasonableness and good faith of the technology can be established.

[38] *See* http://research.microsoft.com/towards2020science/downloads.htm, p 15.

"Process" as a Measure of Reasonableness and Good Faith

Search and retrieval in this new era requires the establishment and recognition of a new standard. A standard of absolute perfection is and always has been unrealistic, but now, with quantitative data available, we know perfection is not only unrealistic, but also quite simply unachievable.

Rather than perfection, which expects that every relevant, non-privileged document will be found and produced, the standard against which we measure these new technologies and processes must be based upon the same principles that have traditionally governed discovery – reasonableness and good faith. Although these terms conjure thoughts of ambiguity and uncertainty, they can actually represent a well-defined set of expectations when placed within the context of process.

A process that emphasizes reasonableness and good faith is fully consistent with what is required under the discovery process. Discovery of information relevant to a dispute gathered by an opponent is often central to a fair and efficient resolution.[39] A party need only identify and produce that which is relevant, as defined by the rules, with the degree of diligence expected and available by experienced practitioners acting reasonably.[40] As noted in Sedona Principles 6 and 11, a party may choose to implement this approach in a reasonable manner, which is left to the good judgment of the party.

Sound process applied to the use of search and retrieval technology can readily establish a measurable means for conducting discovery that satisfies the rules. Reasonableness and good faith can be defined and measured by identifying performance criteria based on their attributes. Accordingly, the unreasonable and unattainable goal of "perfection" should not be allowed to be an enemy of the attainable and measurable goal of reasonableness.

As search and retrieval technologies and associated processes are developed, parties will no doubt want to use them in order to achieve defensible and credible results. If a party fails to adhere to appropriate performance guidelines it will be subject to scrutiny and criticism. Therefore, established process in conjunction with sound technology can serve as a benchmark for conducting discovery in the future. Furthermore, defensibility in court will very likely depend on the implementation of, and adherence to, processes developed for use with a search and retrieval technology.

Implementing Process

Using a search and retrieval technology in conjunction with an implementing process in the complex context of electronic discovery will involve multiple phases of activity, with iterative feedback opportunities at appropriate decision points to allow integration of what a case team learns after each exercise of the process in order to calibrate and maximize the technology's capability to identify relevant information. It is through this feedback that case teams will acquire sound information to use in making both strategic and tactical decisions.

The initial search and retrieval process should be designed with the intent that it serve as a pilot process that can be evaluated and modified as the team learns more about the corpus of information to be reviewed. One useful approach is to initiate the process by focusing on the information collected from a few of the custodians who were at the center of the facts at issue in the litigation or investigation. Focusing on information collected from the core custodians, which has a higher likelihood of being relevant, will help the team efficiently develop its understanding of the issues and language used by the custodians, thus allowing them to more efficiently develop and implement an appropriate search and retrieval process.

[39] *Hickman v. Taylor*, 329 U.S. at 507; *see supra*, n.1.

[40] Under Rule 26(g)(1), an attorney of record is expected to certify that to the best of his or her "knowledge, information, and belief, formed after a reasonable inquiry," that disclosures are "complete and correct" as of the time they were made. Similarly, under Rule 26(g)(2), an attorney must certify that to the best of his or her "knowledge, information, and belief, formed after a reasonable inquiry," that discovery requests, responses, and objections" are made "consistent with these rules."

The initial selection and refinement of search terms can also benefit from the application of sampling techniques that can help the review team to rank the precision and recall of various terms or concepts. Reviewing samples of information that include selected search terms or concepts and ranking their relative value based on their efficacy in retrieving relevant information (recall) and their efficiency in excluding non-relevant information (precision) can help the review team to focus the selection of terms.[41]

The development of process control logs and second-level review techniques can also help the review team to ensure that the designed process is consistently applied to all of the information to be reviewed. Additionally, a second-level review process based on statistical sampling techniques can ensure the achievement of acceptable levels of quality. While these techniques are relatively unknown in the typical review processes in use today, their widespread adoption in businesses of all types should drive their implementation in large document review projects in the near future.

C. How The Legal Community Can Contribute to The Growth of Knowledge

A consensus is forming in the legal community that human review of documents in discovery is expensive, time consuming, and error-prone. There is growing consensus that the application of linguistic and mathematic-based content analysis, search and retrieval technologies, and tools, techniques and process in support of the review function can effectively reduce the cost, time, and error rates.

Recommendations

1. ***The legal community should support collaborative research with the scientific and academic sectors aimed at establishing the efficacy of a range of automated search and information retrieval methods.***

2. ***The legal community should encourage the establishment of objective benchmarking criteria, for use in assisting lawyers in evaluating the competitive legal and regulatory search and retrieval services market.***

As stated, in the 20 years since the Blair and Maron study, there has been little in the way of peer-reviewable research establishing the efficacy of various methods of automated content analysis, search, and retrieval as applied to a legal discovery context. A program of research into the relative efficacy of search and retrieval methods would acknowledge that each alternative should be viewed in the context of its suitability to specific document review tasks. Different technologies, tools and techniques obviously have different strengths. Moreover, the outcome of the application of advanced content analysis, search and retrieval methods can have significant differences based on expertise of the operator. Ideally, a research program would advance the goals of setting minimum or baseline standards for what constitutes adequate information retrieval, as well as reaching agreement on how to benchmark competing methods against agreed-upon objective evaluation measures.

In this regard, The Sedona Conference® supported the introduction of a new "Legal Track" in 2006 for the TREC research program run by the National Institute of Standards and Technology. NIST is a federal technology agency that works with industry to develop and apply technology, measurements and standards. TREC is designed "to encourage research in information retrieval from large text collections."[42] The TREC legal track involves an evaluation of a set of search methodologies

[41] *See* text at Part IV, *supra*.

[42] The Text Retrieval Conference (TREC) was started in 1992. *See* http://trec.nist.gov. Its purpose is to support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. TREC is overseen by a program committee consisting of representatives from government, industry, and academia. Each TREC track involves a test database of documents and topics Participants run their own retrieval systems on the data, and return to NIST a list of the retrieved top-ranked documents. NIST generally pools the individual results, judges the retrieved documents for correctness, and evaluates the results. The TREC cycle ends with a workshop that is a forum for participants to share their experiences. The TREC test collections and evaluation software are available to the retrieval research community at large, so organizations can evaluate their own retrieval systems at any time. TREC has successfully met its dual goals of improving the state-of-the-art in information retrieval and of facilitating technology transfer, and many of today's commercial search engines include technology first developed in TREC.

based on lawyer relevancy assessments on topics drawn from a large public database of OCR-ed documents. The results coming out of the 2006 legal track represent the type of objective research study into the relative efficacy of Boolean and other search methods that the legal community should further encourage.[43]

However, a need exists to scale up the TREC research to accommodate the potential retrieval of millions or tens or hundreds of millions of arguably relevant documents among a greater universe of terabytes, petabytes, exabytes, and beyond.

Members of The Sedona Conference® community have and will continue to participate in collaborative workshops and other fora focused on issues involving information retrieval.[44] How best to leverage the work of the IR community to date is an enterprise beyond the scope of this paper. The Sedona Conference® intends to remain in the forefront of the efforts of the legal community in seeking out centers of excellence in this area, including the possibility of fostering private-public partnerships aimed at focused research.

[43] *See* Jason R. Baron, David D. Lewis & Douglas W. Oard, "TREC 2006 Legal Track Overview," 2006 FIFTEENTH TEXT RETRIEVAL CONFERENCE (TREC 2006) PROCEEDINGS, *available at* http://trec.nist.gov/pubs/trec15/papers/LEGAL06.OVERVIEW.pdf; *see also* TREC 2007 Legal Track, http://trec-legal.umiacs.umd.edu/ (additional documentation relating to TREC 2006 Legal Track).

[44] *See, e.g., Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings* ("DESI Workshop"), held at the Eleventh International Conference on Artificial Intelligence and Law (ICAIL 2007), June 4, 2007, Palo Alto, CA, papers *available at* http://www.umiacs.umd.edu/~oard/desi-ws/.

APPENDIX: Types of Search Methods

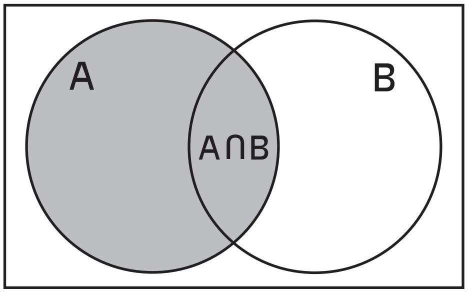
This appendix is a "survey" of different forms of search methods found in the information science literature, and which form the basis of offerings by vendors in the legal marketplace. The list is not definitive. Indeed, as the main body of the Commentary makes clear, rapid technological progress will inevitably affect how methods are described, perfected, and then replaced with new ways of performing search and retrieval.

A second caveat: the following search methods are not intended to be mutually exclusive. Indeed, many products tout the benefits of hybrid, combined, or cumulative approaches to performing searches.

A. Boolean Search Models

A "Boolean search" utilizes the principles of Boolean logic named for George Boole, a British born mathematician. Boolean logic is a method for describing a "set" of objects or ideas. Boolean logic was applied to information retrieval as computers became more widely accepted. Boolean search statements can easily be applied to large sets of unstructured data and the results exactly match the search terms and logical constraints applied by the operators.

As used in set theory, a Boolean notation demonstrates the relationship between the sets or groups, indicating what is in each set alone (set union), what is jointly contained in both (set intersection), and what is contained in neither (set differences). The operators of AND (intersection or ∩), OR (union or ∪) and AND NOT or BUT NOT (difference) are the primary operations of Boolean logic. These relationships can easily be seen within a Venn diagram (see below).



OR is a Boolean operator that states the set may contain any, some or all of the keywords searched. The purpose of this command is to encompass alternative vocabulary terms. OR is represented by the union of the sets A ∪ B (the entire shaded areas above). The use of OR expands the resulting Boolean set.

AND is a Boolean operator used to identify the intersection of two sets or two keywords. The purpose of this command is to help construct more complex concepts from more simple vocabulary words. AND is represented by the middle intersecting area above (A ∩ B). The use of AND restricts the resulting Boolean set.

NOT is a Boolean operator used to eliminate unwanted terms. The purpose of this command (preceded by either AND or BUT) is to help suppress multiple meanings of the same term; in other words, eliminating ambiguity.

Different search engines or search tools may provide additional Boolean-type operators or connectors to create more complex search statements. These may include:

- **Parenthesis:** A Boolean search may include the use of parentheses to force a logical order to the execution of the search, as well as to create more refined and flexible criteria. Any number of logical ANDs (or any number of logical ORs) may be chained together without ambiguity; however, the combination of ANDs and ORs and AND NOTs or BUT NOTs sometimes can lead to ambiguous cases. In such cases, parentheses may be used to clarify the order of operations. The operations within the innermost pair of parentheses are performed first, followed by the next pair out, etc., until all operations are completed.

- **Proximity or NEAR/WITHIN operator:** Another extension to Boolean searching, this technique checks the position of terms and only matches those within the specified distance. This is a useful method for establishing relevancy between search criteria, as well as for paring down irrelevant matches and getting better results (improving precision). Some search engines let you define the order, in addition to the distance. For example: *budget w/10 deficit* might mean "deficit within the 10 words following the word budget".

- **Phrase Searching:** Some search engines provide an option to search a set of words as a phrase, either by typing in quote marks (" ") or by using a command. When they receive this kind of search, the engines will generally locate all words that match the search terms, and then discard those which are not next to each other in the correct order. To perform this task efficiently, the index typically will store the position of the word in the document, so the search engine can tell where the words are located.

- **Wildcard operators** (also sometimes referred to as truncation and stemming). This search capability allows the user to widen the search by searching a word stem or incomplete term. It is typically a symbol such as a question mark (?), asterisk (*), or exclamation point (!). The search system may also allow the user to restrict the truncation to a certain number of letters by adding additional truncation symbols. For example: Teach?? would find teaches and teacher but would not find teaching. In addition, some systems will allow for internal truncation such as wom?n would find women or woman. The * and ! terms have broader application: for example, hous* would find house, housemate, Houston, household or other similar words with the stem "hous."

B. Probabilistic Search Models: Bayesian Classifiers

Probability theories are used in information retrieval to make decisions regarding relevant documents. The most prominent of these are so-called "Bayesian" systems or methods, based on Bayes' Theorem. The theorem was developed by Thomas Bayes, an eighteenth century British mathematician. A Bayesian system sets up a formula that places a value on words, their interrelationships, proximity and frequency. By computing these values, a relevancy ranking can be determined for each document in a search result. This weighting may be based on a variety of factors:

- Frequency of terms within a document- the more times it appears, the more weight it carries.

- Closer to the top – documents with the term in the title are more heavily weighted

- Adjacency or proximity – the closer the terms are to each other, the higher the weighting

- Explicit or implicit feedback on relevance

(Note: other types of search models apply these types of concepts or ideas as well.)

Bayesian systems frequently utilize a "training set" of highly relevant documents to increase understanding, and therefore the probability measures of the system. During training, the system examines each word in the training documents and computes the probability with which that word occurs in each category.

For example, the word "potato" may occur in 5 documents in the category "kitchen tools" (*e.g.*, "potato peeler"), in 7 documents in the category "farm products," and in one document in the category "garden tools." When a new document is then found to contain the word "potato," the Bayesian classifier will interpret this new document as most likely to be a member of the category "farm products" than either of the other two. The same process is repeated for all of the words in the document. Each word in the document provides evidence for which of the categories the document belongs to. The Bayesian classifier combines all of this evidence, using Bayes' rule, and determines the most likely category.

Bayesian classifiers provide powerful tools for comparing documents and organizing documents into useful categories with a moderate amount of effort.

C. Fuzzy Search Models

Boolean and probabilistic search models rely on exact word matches to form the results to a query. Exact matching is very strict: either a word matches or it doesn't. Fuzzy search is an attempt to improve search recall by matching more than the exact word: fuzzy matching techniques try to reduce words to their core and then match all forms of the word. The method is related to stemming in Boolean classifiers, discussed above.

Some algorithms for fuzzy matching use the understanding that the beginning and end of English words are more likely to change than the center, so they count matching letters and give more weight to words with the matching letters in the center than at the edges. Unfortunately, this can sometimes bring up results that make little sense (a search for tivoli might bring up ravioli).

Many systems allow one to assign a degree of "fuzziness" based on the percentage of characters that are different. Fuzzy searching, or matching, has at least two different variations: finding one or more matching strings of a text, and finding similar strings within a fixed string set often referred to as a dictionary. Fuzzy searching has many applications in legal information retrieval including: spellchecking, email addresses and OCR clean-up.

D. Statistical Methods: Clustering

Systems may use statistics or other machine-learning tools to recognize what category certain information belongs to. The simplest of these is the use of "statistical clustering." Clustering is the process of grouping together documents with similar content. There are a variety of ways to define similarity, but one way is to count the number of words that overlap between each pair of documents. The more words they have in common, the more likely they are to be about the same thing.

Many clustering tools build hierarchical clusters of documents. Some organize the documents into a fixed number of clusters. The quality or "purity" of clustering (*i.e.*, the degree to which the cluster contains only what it should) is rarely as high as that obtained using custom built taxonomies or ontologies, but since they require no human intervention to construct, clustering is often an economical and effective first pass at organizing the documents in a collection.

Some systems improve the quality of clusters that are produced by starting with a selected number of clusters, each containing selected related documents. These selected documents then function as "seeds" for the clusters. Other related documents are then joined to them to form clusters that correspond to their designer's interests. Then, additional documents are added to these clusters if they are sufficiently similar.

E. Machine Learning Approaches to Semantic Representation

Bayesian classifiers are often considered "naïve" because they assume that every word in a document is independent of every other word in the document. In contrast, there is a class of concept learning technologies that embrace the notion that words are often correlated with one another, and that there is value in that correlation. These methods are also referred to as "dimensionality reduction techniques" or "dimension reduction systems."

These systems recognize there is redundancy among word usage and take advantage of that redundancy to find "simpler" representations of the text. For example, a document that mentions "lawsuits" is also likely to mention "lawyers," "judges," "attorneys," etc. These words are not synonyms, but they do share certain meaning characteristics. The presence of any one of these words would be suggestive of their common theme. Documents that mentioned any of these terms would likely be about law. Conversely, in searching for one of these words, one might be almost as satisfied to find a document that did not contain that exact word, but did contain one of these related words.

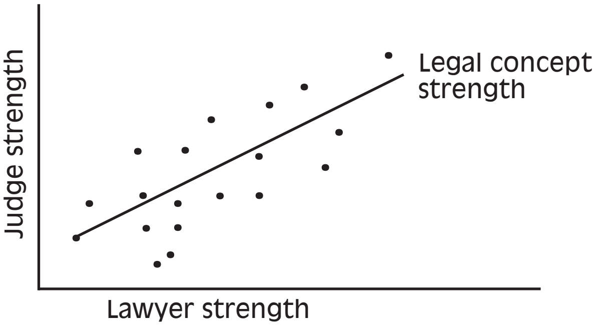


Figure 1. Dimension reduction – the original dimensions of "lawyer" and "judge" are combined into a single dimension. Each point in the graph represents a document. Its location in the graph shows how much the document is related to each dimension.

The figure above illustrates the kind of relationships such systems find. The word "lawyer" tends to occur in the same context as the word "judge." Each document has a certain strength along the "lawyer" dimension, related, for example, to how many times the word "lawyer" appears. Similarly, documents have strength along the "judge" dimension, related, for example, to how many times the word "judge" appears. These systems find a new dimension that summarizes the relationship between "lawyer" and "judge." In this example, we are reducing the dimensions from two to one.

Mathematically, we can then describe documents by how much strength they have along this dimension and not concern ourselves with its strength along the original "lawyer" or "judge" dimensions. The new dimension is a summary of the original dimensions, and the same thing can be done for all words in all the documents. We can locate documents along these new, reduced, dimensions or we can represent words along these dimensions in a similar way.

Similarly, multiple words can be represented along dimensions. And, instead of having just one summary dimension, we can have many of them. Instead of describing a document by how it relates to each of the words it contains, as is done with Vector Space Models,[45] we can describe the document by how it relates to each of these reduced dimensions. Latent Semantic Indexing (LSI, also called Latent Semantic Analysis) is probably the best known of these dimension-reducing techniques, but there are others, including neural networks and other kinds of statistical language modeling.

Such techniques are similar to one another in that they learn the representations of the words in the documents from the documents themselves. Their power comes from reducing the dimensionality of the documents. They simplify representation, and make recognizing meaning easier.

For example, a collection of a million documents might contain 70,000 or more unique words. Each document in this collection can be represented as a list of 70,000 numbers, where each number stands for each word (say the frequency with which that word occurs in that document). Using these techniques, one can represent each document by its strength along each of the reduced dimensions.

[45] *See* H. Roitblat, *supra*, n.34.

One can think of these strengths as a *meaning signature*, where similar words will have similar meaning signatures. Documents with similar meanings will have similar meaning signatures. As a result, the system can recognize documents that are related, even if they have different words, because they have similar meaning signatures.

F. Concept and Categorization Tools: Thesauri, Taxonomies and Ontologies

To deal with the problem of synonymy, some systems rely on a thesaurus, which lists alternative ways of expressing the same or similar ideas. When a term is used in a query, the system uses a thesaurus to automatically search for all similar terms. The combination of query term and the additional terms identified by the thesaurus can be said to constitute a "concept."

The quality of the results obtained with a thesaurus depends on the quality of the thesaurus, which, in turn, depends on the effort expended to match the vocabulary and usage of the organization using it. Generic thesauri, which may attempt to represent the English language or are specialized for particular industries, are sometimes available to provide a starting point, but each group or organization has its own jargon and own way of talking that require adjustment for effective categorization. In America, for example, the noun "jumper" is a child's one-piece garment. In Australia, the noun "jumper" is a sweater. In America, a 3.5 inch removable disk device was called a "floppy" during its heyday. But in Australia, it was called a "stiffy."

Taxonomies and ontologies are also used to provide conceptual categorization. Taxonomy is a hierarchical scheme for representing classes and subclasses of concepts. The figure below shows a part of a taxonomy for legal personnel. Attorneys, lawyers, etc. are all kinds of law personnel. The only relations typically included in a taxonomy are inclusion relations. Items lower in the taxonomy are subclasses of items higher in the taxonomy. For example, the NAICS (North American Industry Classification System) is one generally available taxonomy that is used to categorize businesses. In this taxonomy, the category "Information" has subclasses of "Publishing" and "Motion Picture and Sound Recording Industries" and "Broadcasting."

One can use this kind of taxonomy to recognize the conceptual relationship among these different types of personnel. If your category includes law personnel, then any document that mentions attorney, lawyer, paralegal, etc. should be included in that category. Like thesauri, there are a number of commercially available taxonomies for various industries.

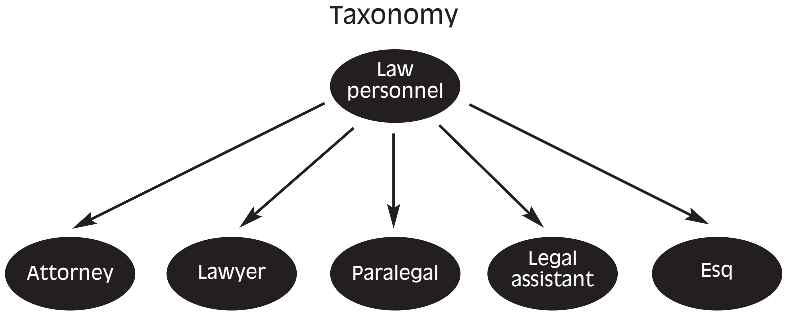


Figure 2. A simple taxonomy for law personnel.

Predefined taxonomies exist for major business functions and specific industries. It may be necessary to adapt these taxonomies to one's particular organization or matter.

An ontology is a more generic species of taxonomy, often including a wider variety of relationship types than are found in the typical taxonomy. An ontology specifies the relevant set of conceptual categories and how they are related to one another. The figure below shows part of an ontology covering subject matter similar to that described in the preceding taxonomy. For clarity, only a subset of the connections between categories is shown. According to this ontology, if your category includes attorneys, you may also be interested in documents that use words such as "lawyer," "paralegal," or "Esq." Like taxonomies, ontologies are most useful when they are adapted to the specific information characteristics of the organization.

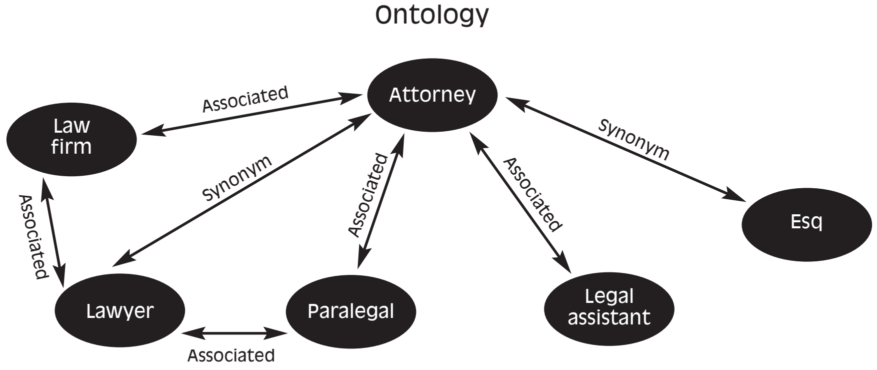


Figure 3. A section of an ontology of legal personnel.

Taxonomies, ontologies, and thesauri are all knowledge structures. They represent explicit knowledge about some subject. An expert writes down the specific relations she knows about. Although there are tools that help the expert create these structures, they still tend to represent only the information the expert can explicitly describe as important.

The structure of the thesaurus, taxonomy, or ontology can be used as the organizing principle for a collection of documents. Rules are derived that specify how documents with specific words in them are related to each of these categories, and the computer can then be used to organize the documents into the corresponding categories.

These rules can be created explicitly, or they can be created using machine-learning techniques. Explicit rules are created by knowledge engineers. For example, one rule might include a Boolean statement like this: (acquir* or acquisition or divest* or joint venture or alliance or merg*) and (compet* or content or program*) that specifies the critical words that must appear for a document to be assigned to the "merger" category. The effectiveness of rules like these depends critically on the ability of the knowledge engineers to guess the specific words that document authors actually used. Syntactic rules may also be employed by some systems. For example, a system may only look for specific words when they are part of the noun phrase of a sentence.

G. Presentation/Visualization Tools

Presentation and visualization software technologies may incorporate search and retrieval functionality that may be found to have useful applications. These technologies can organize information (*e.g.*, emails) so that a researcher can more efficiently study the research topic (including finding relevant emails). They also are good at highlighting patterns of "social networks" within an organization that would not necessarily be apparent by more traditional searches. Subject to some exceptions, the results of any search and retrieval query can be presented in a variety of forms, including as a:

1. List – items in sequence, for example messages ordered by sent date
2. Table – items aggregated into rows by columns, for example messages by sender
3. Group – items categorized or totaled, for example count of messages by sender

4. Cluster – items in groups organized by spatial proximity, for example relevant groups spiraling out to less relevant groups
5. Tree – items in parent/child hierarchy, for example, folder and subfolder(s)
6. Timeline – items arrayed by a time element, for example a list/group of items arrayed by sent date
7. Thread – items grouped by conversation
8. Network – items arrayed by person, for example a diagram of message traffic between sender(s) and recipient(s)
9. Map – items plotted by geography, for example items plotted by city and state of origin
10. Cube – items in a multi-dimensional pivot table; includes, table, group, timeline and tree functionality

In practice, a researcher can load search results into a presentation technology for an organized view, and then drill-down to access discrete items of significant interest or concern. This often iterative process may help a researcher to learn more about, act on, and manage search results.

TSC

Advanced Learning in a Panoramic Setting℠

Visit www.thesedonaconference.org

Exhibit C

JONES DAY

51 LOUISIANA AVENUE, N.W. • WASHINGTON, D.C. 20001.2113
TELEPHONE: +1.202.879.3939 • FACSIMILE: +1.202.626.1700

Direct Number: (202) 879-3669
jberman@JonesDay.com

March 1, 2012

VIA E-MAIL AND FIRST CLASS MAIL

Thomas C. Gricks, III
Fifth Avenue Place
120 Fifth Avenue
Suite 2700
Pittsburgh, PA 15222-3001
412-577-5205
tgricks@schnader.com

Re: Document Production in the Dulles Jet Center Litigation

Tom,

It was nice speaking with you yesterday. As we discussed, I would like to come to an agreement regarding the format in which our clients are to produce electronic documents. Further, after discussions with counsel for some of the other plaintiffs, I have several follow-up questions regarding your proposal to use predictive coding as a method of determining which documents are relevant to the discovery requests that have been served on Landow Aviation Limited Partnership, Landow Aviation I, Inc. and Landow & Company Builders, Inc.

Format of Electronic Documents

With regard to the format of your production, below is a general list of my production preferences. Please let me know if you have different preferences, and of course other issues may arise.

a. Family-units should be produced as kept in the normal course of business. For instance, all email should be produced with all of its attachments regardless of the relevancy of any individual attachment.

b. Unless otherwise specified below, each document should be produced as a uniquely endorsed (Bates stamped) single-page Group IV TIF image provided with a page level image load file that designates document level boundaries, an associated document level extracted text file, and document level metadata fields provided in a delimited load file.

The metadata fields to be provided are: Beginning Production Number, Ending Production Number, Beginning Attachment Range, Ending Attachment Range, Custodian, Original Location Path, Email Folder Path, Document Type, Author, File Name, File Size, MD5 Hash, Date Last Modified, Date Created, Date Last Accessed, Date Sent, Date Received, Recipients, Copyees, Blind Copyees, Email Subject, Path to Native File

c. Some file types should be produced natively, such as:

   i. CAD drawings and other engineering drawings,

   ii. Microsoft Excel or other spreadsheet or database files,

   iii. Microsoft Powerpoint, and

   iv. Audio, video, and graphics files.

In the larger production, each relevant native files should be accounted for with a unique record in the load file identified with a unique Bates number, associated metadata, and a single-page Group IV .TIFF image slip sheet branded with the unique Bates number.

d. We will provide you with specific load file specifications as we near the production date.

Predictive Coding Issues

1. Which custodians did you collect potentially relevant data from? For each custodian, please provide us with a folder directory indicating how the custodian stored his or her data in the normal course of business.

2. In terms of data size, can you confirm that the initial data collection was 8 terabytes in size, which you have narrowed to 250-300 gigabytes of potentially relevant data? What methods did you use to cull the data from the initial 8 terabytes to this estimated 250-300 gigabytes?

3. How many discrete documents/files does this 250-300 gigabytes of data represent? While there are methods for estimating the number of documents based on data volume, I would like to know the precise count.

4. Which file extensions account for more than 10% of the 250-300 gigabytes volume?

5. To help the parties to understand the data you have collected, the plaintiffs previously provided a list of proposed search terms, which is attached as Exhibit A. I have slightly modified the search instructions to ensure that email addresses of the named individuals are captured by the search. I did so by adding the following addition instructions:

   > More generally, it is assumed that search terms that form a part of a longer term would yield a "hit" for the longer term. For example, it is assumed that a search for "condemn" would also yield "hits" for "condemned" and "condemnation." Likewise, a search for “Kim” would return a hit for tkim@domainname.com.

   With this additional instruction, can you provide us with a count of how many unique hits and unique documents are returned for each search term in the attached list of search terms?

6. Can you confirm that rather than using traditional human review or using a review based on search terms, you are proposing to use OrcaTec’s predictive coding model to identify the relevant documents within the 250-300 gigabytes of data?

7. Can you provide the statistical sampling methodology that is going to be deployed against this 250-300 gigabytes of data? Specifically, what is the sample size that is used for the initial “seed set”? How is this “seed set” selected? Once the “seed set” is reviewed, how is it applied to the remaining data? What is the confidence level threshold at which a document is deemed to be responsive? For documents falling below that threshold confidence level, what is your proposed workflow for re-review?

8. Are you processing all file types? If there are certain file extensions that will not be included in the predictive coding work-flow, how will these be reviewed for relevancy. For instance, it is my understanding that CAD drawings cannot be processed and loaded to most review tools.

9. What is the method you are proposing to use for the identification and withholding of privileged documents within the 250-300 gigabyte set.

I look forward to your response. After you have had a chance to assemble the requested information, perhaps we should schedule a conference call to work through the issues.

Very truly yours,

/s

Jonathan Berman

cc: Jonathan M. Stern, Esq.
Morgan W. Campbell, Esq.
Brandon H. Elledge, Esq.
Richard D. Gable, Jr., Esq.

SEARCH TERMS FOR LANDOW E-PRODUCTION

Note: each line represents one search term. So, for example, the search term "Thomas Kim" would not yield a "hit" if the searched data only contains the first name "Thomas" or the last name "Kim." Instead, it would require the words "Thomas" and "Kim" to appear together as the single term "Thomas Kim." On the other hand, "Sergio" is one search term, and would yield a hit even if the last name "Plaza" is not found in the searched data.

Note: The below list assumes that the search would not discriminate between upper case and lower case letters. For example, although one of the search terms listed is "Hangar," it is intended that this would also encompass the terms "hangar" and "HANGAR."

Note: It is assumed that a search for a singular will also yield a plural-so For example, the search term "weld" would yield a "hit" for "welds." More generally, it is assumed that search terms that form a part of a longer term would yield a "hit" for the longer term. For example, it is assumed that a search for "condemn" would also yield "hits" for "condemned" and "condemnation." Likewise, a search for “Kim” would return a hit for tkim@domainname.com.

Terms:

Dulles Jet
Jet Center
DJC
Hangar
Hanger
Aviation tenant
Eaglespan
Eagle Span
Loveland
EagleBeam
PageMark
Sergio
Plaza
Jerry
Curtis
Ehi
Lambert
Monte Osborn
Kerri White
Thomas Kim
Thomas M. Kim
Kevin Stearns
Marty Babb
Bascon
Beke
Mason

Proudfoot
George W. Lucas
Wade
Lucas
Hart
Tamanko
Eugene Owen
Gene Owen
Whitacre
Frommer
Copley
Remley
Cagley
Alison Copley
Mark Holhubner
Shelly Nichols
William Boothe
Bill Boothe
Scott Harper
Amy Ruhe
Chris Verlander
Jeff Weinheimer
Rich Tanner
David Davis
Chander
Nangia
Roy Daniel
Micky
Mickey
Curro
Pinnacle
Kip
Kipton
Ping
Brian Wagner
Michael Walkley
Robert Edge
Christopher B. Smith
MWAA
Metropolitan Washington Airports
Kimmel
Seedlock
Mlinarcik
Christopher U. Browne
David A. Jones
Frank D. Holly, Jr.

Halterman
Independent Testing
ITIS
Sturgeon
John Young
National Door
P.C. Cummings
DGS
Ted Brennan
Daniel Schuster
Don Wickesser
Tom Cason
J.P. Dwyer
Schuster
Schnabel
Cepull
Rabe
Huprich
Faber
Tawfik
Hafid
Sungkar
Aden
Mohamed
Abuzied
Lucciano
Campos
Ruben
Taruselli
Joanne Sloane
Yemi
Bamigbade
Odorisio
Justin Jubie
Michael Baker
Christopher DiChiaro
Weld
Structural steel
Structural calculation
Rafter
Column
Corrugated web
Corrugated metal
Haunch
Splice
Magnetic particle

Mag particle
Ultrasound
Ultrasonic
AWS
American Welding Society
PEMB
Pre-Engineered Metal Building
Metal Building System
Inspect
Deflect
AISC
American Institute of Steel Construction
IBC
International Building Code
ASCE
Virginia Uniforn Statewide Building Code
VUSBC
USBC
MBMA
Metal Building Manufacturer
Kodak
CSC
Computer Sciences Corp.
Computer Sciences Corporation
Jet Aviation
Arcadia
Global Aerospace
Global Express *9052*
Global Express N620K
N59AP
BAE
N800LA
Armstrong
Phoenix Steel
Dark Horse
Ultimate Building, Inc.
Gallier
JRF
Walker Iron
MBCI
Metal Building Components
NCI
Metallic Building Company
R&M Steel Co.
Door Engineering
Dominion Caisson

Snow load
Failure load
Blizzard
Snowmageddon
Condemn
Collapse
Bowden
Ryerson
Norco
Cargill
Hercules
Homan Welding
Alexandria Surveys
Allyn
Kilsheimer
Emilio
KCE
KTA Group
Smith Group
Guiliani
Bums and McDonnell
HLW
Soliman
Triphase
deconstruct
Vika
Miller Long
Melallurgical Technologies
General Dynamics
Jeff Kudlac
Gary Rogerson
Scott Hoffman
Quality Control
Quality Assurance
M.I.C.
MIC
N767DT

Exhibit D

From: Gricks III, Thomas C.
To: "Jonathan Berman"
Cc: brandon.elledge@hklaw.com; Stern, Jonathan M.; mcampbell@dglitigators.com; rgable@gibbonslaw.com; William G Laxton Jr
Subject: RE: DJC Litigation / e-discovery issues
Date: Thursday, March 15, 2012 6:13:00 PM
Attachments: image001.png

Jonathan:

I am writing in response to your letter of March 1, 2012 to further our discussions on the production of electronically stored information.

Format of Production

Since you are suggesting a TIFF production, document families cannot be produced in the ordinary course, as such. Rather, each document is produced independently, and the parent-child relationship is maintained through the fields in the dat file. I do agree with the notion that we should treat document families as a whole, both in the production and the withholding of families. Therefore, families with one or more responsive documents will be generally be produced, but families with one or more privileged documents will be withheld in their entirety. We can address fine tuning privileges of individual family member documents if necessary upon the preparation of a privilege log that details the privilege and the family relationship.

I am generally in agreement with your proposed production format. We can produce single page TIFFs with a load file and dat file, with all associated text and metadata. However, I do not believe we need to include all of the metadata fields that you are suggesting. For example, the Original Location Path, Email Folder Path, Date Last Accessed and Path to Native File are not likely to provide any pertinent information so as to make them necessary for every record, and including them in the database may unnecessarily complicate production. Should those data points be absolutely necessary for specific documents or records, available data will be maintained and can likely be provided if and when needed.

I am happy to produce appropriate files natively, and would suggest that we do so only for those files that will not TIFF adequately. Generally, that would include graphics files (images, audio, video) and spreadsheets/databases. However, I am not sure we really need or want to produce PowerPoint decks natively, unless there is some relevant animation that exists and must be viewed. For consistency purposes, I think it is generally better to produce every document that will TIFF in an image format with a Bates number. That said, I am not sure there are any ppt files in the ESI, so it may not be an issue in any event. I am sure you realize that we cannot produce native files unless we have them in native form, so the production of ESI that we have received from other parties in non-native form, or that does not exist in our ESI in native form, cannot be converted back.

Predictive Coding Issues

1. We conducted two comprehensive collections, including Ghost and Forensic images of the

hard drives from the following personal computers and laptops at the Landow offices and the Dulles Jet Center: (DJC) Room 54-WS-DJC05, Room 55-WS-DJC, Construction Trailer, Room 60-DJC, White, Ross, Front Desk 1/Line Service, Front Desk 2; (Landow) Landow, Nathan; Curro, Mickey; Line Room Dulles/Line Service 3; McNeely, John; McCann, Mary; Chen, Fay; Herlson, Diane; Frisque, Michele; Callaghan, Meagan; Harris, Steve; Beach-Uhlman, Judy; Landow, David; Landow, Michael; C Byrd 1; C Byrd 2. We also collected email and data files that were resident on the following: (DJC) Server; (Landow) Timberline Server, Server 2003. I do not plan to provide any directory trees for these images, as they are simply too burdensome to prepare and irrelevant to the production of ESI in any manner.

2. The collection is indeed 8 terabytes, as it consists primarily of numerous large, imaged hard drives. The data from the initial images of those drives was de-NISTed and de-duped on a custodian level, and graphics files were removed from the dataset, to isolate true data file types, such as doc(x), xls(x), pdf, msg, etc. (We will have to review the graphics files separately to determine whether they represent any pertinent ESI.) The resultant collection set contains 130GB of email containers and 70GB of loose native files. It is estimated that the second collection will add an additional twenty-five percent (25%) to the volume of ESI after de-duping against the first set, since they were conducted at different times.

3. The data from the initial collection consists of 200 email containers and roughly 210,000 loose files. Beyond this estimate of data volume, I do not see any benefit to expanding all of the email containers to determine how many documents they may contain.

4. Obviously, from the above, the email containers represent the largest volume of data. As for the loose files, below is a summary of the most common file types for the native files:

File Type	# of Files	% of Total
htm	50,700	24.1
txt	50,100	23.9
html	39,400	18.8
wpd	17,800	8.5
pdf	16,100	7.7
xls	12,000	5.7
doc	8,400	4
msg / eml	4,400	2.1
zip	2,300	1.1
rtf	2,000	1
docx	800	0.4
xlsx	500	0.2

5. There is no practical way of generating the keyword data that you are seeking, nor does it

serve any useful purpose in the analysis of the proposal to use predictive coding to generate a useful review set of documents from the mass of ESI with which we are dealing. To evaluate the data in the manner in which you are proposing would require the loading and special indexing of all of the ESI, and likely several weeks of analysis. You are seeking acronym searches, string searches, stemming, wildcards and the like. This is a massive undertaking, and it will not give us much useful information. We have done some very limited analysis to understand the scope of a keyword analysis, which provides at least some guidance. Just to give you an idea of the difficulties we will see with keywords from this limited review, we found the following relationships (relevant/non-relevant percentages): Dulles Jet (0/100); Jet Center (36/64); hangar (67/33); sergio (0/100); plaza (0/100); mickey (29/71); curro (0/100); column (47/53); inspection (15/85).

6. The most cost effective means of culling a collection set into a good review set is presently predictive coding, and that is what we are proposing. Whether we use OrcaTec or another vendor (and that decision has not been made), all of the predictive coding tools generate a review set with greater precision and recall, and less cost, than a traditional linear review.

7. The predictive coding will not necessarily commence with a seed set, as such, or a sample. The procedure will depend on the vendor. For example, OrcaTec and Equivio operate by random or modified-random generation of a small set (usually about 100 documents) for coding as relevant or non-relevant. Recommind uses a seed set, and then returns additional sets of documents anticipated to be relevant for further coding as relevant or non-relevant. Each predictive coding tool uses some form of statistical linguistic-mathematical algorithm, whether latent semantic analysis, probabilistic latent semantic analysis, latent Dirichlet analysis or some variation, to model the coding results for relevant and perhaps non-relevant documents. Once the model stabilizes, and the tool is consistent in categorization with the reviewer, the tool uses the model to locate all of the similar documents in the collection set, categorizing them as relevant or non-relevant. It is at this point that statistical sampling comes in, as the resultant review set should be checked statistically to ensure that recall equals or exceeds that of linear review, within an agreed confidence level and interval. If not, additional documents would be coded through the tool to further refine the model, and the categorization and sampling would be redone.

8. I would suggest that we do not need to include non-text files in the predictive coding process – things like drawings, images, audio, video, etc. I would propose to review those documents separately, just as in any other production.

9. I have not reviewed the documents sufficiently to know what filters will definitely be applied for privilege review. However, I generally like to apply bulk filtering techniques using lawyers names; firm names; sender, recipient and domain searches; and other general searches to eliminate potentially privileged or confidential material. That enables us to make the bulk of the production more quickly. Then I typically conduct a linear review on the documents withheld to identify those that are truly privileged. Privileged documents are logged, and non-privileged documents are produced in a subsequent production.

I trusts this adequately responds to your questions. Please let me know when you would like to discuss the above issues in greater detail, and I would also like to discuss reasonable financial arrangements in connection with this production. We would like to move this component of the litigation forward as promptly as these issues can be resolved.

Tom

From: Jonathan Berman [mailto:jberman@JonesDay.com]
Sent: Thursday, March 15, 2012 10:50 AM
To: Gricks III, Thomas C.
Cc: brandon.elledge@hklaw.com; Stern, Jonathan M.; mcampbell@dglitigators.com; rgable@gibbonslaw.com; William G Laxton Jr
Subject: Re: DJC Litigation / e-discovery issues

Dear Tom,

I am writing again regarding plaintiffs' attempts to resolve the electronic discovery issues.

Although one week ago you indicated that you were preparing a response to my letter of March 1, I have yet to receive any substantive response to that letter. (My March 1 letter is attached here again for your convenience.)

Since the lack of a resolution is delaying both your document production and mine, I would like to move forward. Please let me know when we can meet-and-confer on the issues set out in the letter.

Best regards,

Jonathan Berman

JONES DAY

Jonathan Berman

51 Louisiana Ave., NW • Washington, DC 20001-2113
DIRECT 202.879.3669 • **FAX** 202.626.1700 • **EMAIL** JBerman@JonesDay.com

From: "Gricks III, Thomas C." <TGricksIII@Schnader.com>
To: Jonathan Berman <jberman@JonesDay.com>
Cc: "Stern, Jonathan M." <JStern@Schnader.com>, "mcampbell@dglitigators.com" <mcampbell@dglitigators.com>, "brandon.elledge@hklaw.com" <brandon.elledge@hklaw.com>, "rgable@gibbonslaw.com" <rgable@gibbonslaw.com>, William G Laxton Jr <wglaxton@JonesDay.com>

Date: 03/08/2012 10:25 AM
Subject: Re: DJC Litigation / e-discovery issues

I am currently out of the office in meetings, but will prepare a response to frame our discussion.

Sent from my iPad

On Mar 8, 2012, at 10:20 AM, "Jonathan Berman" <jberman@JonesDay.com> wrote:

Dear Tom,

Please let me know when we can meet and confer on the issues raised in my letter of March 1, attached below.

Best regards,

Jonathan Berman

<ATT00001.gif> <ATT00002.gif>

Jonathan Berman

51 Louisiana Ave., NW • Washington, DC 20001-2113
DIRECT 202.879.3669 • **FAX** 202.626.1700 • **EMAIL** JBerman@JonesDay.com

----- Forwarded by Jonathan Berman/JonesDay on 03/08/2012 10:18 AM -----

From: Jonathan Berman/JonesDay
To: tgricks@schnader.com
Cc: "Stern, Jonathan M." <JStern@Schnader.com>, mcampbell@dglitigators.com, brandon.elledge@hklaw.com, rgable@gibbonslaw.com, William G Laxton Jr/JonesDay@JonesDay
Date: 03/01/2012 05:10 PM
Subject: DJC Litigation / e-discovery issues

Dear Tom,

Attached please find a letter regarding e-discovery issues.

Best regards,

Jonathan Berman

<ATT00003.gif> <ATT00004.gif>

Jonathan Berman

51 Louisiana Ave., NW • Washington, DC 20001-2113
DIRECT 202.879.3669 • **FAX** 202.626.1700 • **EMAIL** JBerman@JonesDay.com

==========

This e-mail (including any attachments) may contain information that is private, confidential, or protected by attorney-client or other privilege. If you received this e-mail in error, please delete it from your system without copying it and notify sender by reply e-mail, so that our records can be corrected.

==========

==========

This e-mail (including any attachments) may contain information that is private, confidential, or protected by attorney-client or other privilege. If you received this e-mail in error, please delete it from your system without copying it and notify sender by reply e-mail, so that our records can be corrected.

==========

<JD_3060047_1.pdf>
<Search Terms for Landow E-Production.pdf>

==========

This e-mail (including any attachments) may contain information that is private, confidential, or protected by attorney-client or other privilege. If you received this e-mail in error, please delete it from your system without copying it and notify sender by reply e-mail, so that our records can be corrected.

==========

Exhibit E

LexisNexis®

MONIQUE DA SILVA MOORE, et al., Plaintiffs, -against- PUBLICIS GROUPE & MSL GROUP, Defendants.

11 Civ. 1279 (ALC) (AJP)

UNITED STATES DISTRICT COURT FOR THE SOUTHERN DISTRICT OF NEW YORK

2012 U.S. Dist. LEXIS 23350*; *18 Wage & Hour Cas. 2d (BNA) 1479

February 24, 2012, Decided
February 24, 2012, Filed

PRIOR HISTORY: *Moore v. Publicis Groupe SA, 2012 U.S. Dist. LEXIS 19857 (S.D.N.Y., Feb. 14, 2012)*

COUNSEL: **[*1]** For Monique Da Silva Moore, on behalf of herself and all others similarly situated, Plaintiff: Jeremy Heisler, LEAD ATTORNEY, Sanford, Wittels & Heisler, LLP, New York, NY; David W. Sanford, PRO HAC VICE, Sanford, Wittels & Heisler, LLP (DC), Washington, DC; Deepika Bains, Steven Lance Wittels, Sanford Wittels & Heisler, LLP, New York, NY; Janette Lynn Wipper, PRO HAC VICE, SANFORD WITTELS & HEISLER, LLP (San Francisco, San Francisco, CA; Siham Nurhussein, Clifford Chance US, LLP (NYC), New York, NY.

For Maryellen O'Donohue, on behalf of herself and all others similarly situated, Laurie Mayers, on behalf of herself and all others similarly situated, Heather Pierce on behalf of herself and all others similarly situated, Katherine Wilkinson, on behalf of herself and all others similarly situated, Plaintiffs: Jeremy Heisler, LEAD ATTORNEY, Sanford, Wittels & Heisler, LLP, New York, NY; David W. Sanford, PRO HAC VICE, Sanford, Wittels & Heisler, LLP (DC), Washington, DC; Deepika Bains, Steven Lance Wittels, Sanford Wittels & Heisler, LLP, New York, NY; Janette Lynn Wipper, PRO HAC VICE, SANFORD WITTELS & HEISLER, LLP (San Francisco, San Francisco, CA.

For Publicis Groupe, Defendant: Melissa **[*2]** Ruth Kelly, LEAD ATTORNEY, Morgan, Lewis & Bockius LLP (New York), New York, NY; Paul Clayton Evans, PRO HAC VICE, Morgan Lewis & Bockius, LLP (PA), Philadelphia, PA; Paul Clayton Evans, Morgan, Lewis & Bockius LLP, Philadelphia, PA.

For MSL Group, Defendant: Noel P. Tripp, Paul J. Siegel, LEAD ATTORNEYS, Jeffrey W. Brecher, Jackson Lewis LLP(Melville N.Y.), Melville, NY; Brett Michael Anders, Jackson Lewis LLP(NJ), Morristown, NJ; Victoria Woodin Chavey, Jackson Lewis LLP, Hartford, CT.

JUDGES: Andrew J. Peck, United States Magistrate Judge.

OPINION BY: Andrew J. Peck

OPINION

<u>OPINION AND ORDER</u>

ANDREW J. PECK, United States Magistrate Judge:

In my article Search, Forward: Will manual document review and keyword searches be replaced by computer-assisted coding?, I wrote:

> To my knowledge, no reported case (federal or state) has ruled on the use of computer-assisted coding. While anecdotally it appears that some lawyers are using predictive coding technology, it also appears that many lawyers (and their clients) are waiting for a judicial decision approving of computer-assisted review.
>
> Perhaps they are looking for an opinion concluding that: "It is the opinion of this court that the use of predictive coding is a proper **[*3]** and acceptable means of conducting searches under the Federal Rules of Civil Procedure, and furthermore that the software provided for this purpose by [insert name of your favorite vendor] is the software of choice in this court." If so, it will be a long wait.
>
> . . . .
>
> Until there is a judicial opinion approving (or even critiquing) the use of predictive coding, counsel will just have to rely on this article as a sign of judicial approval. In my opinion, computer-assisted coding should be used in those cases where it will help "secure the just, speedy, and inexpensive" (*Fed. R. Civ. P. 1*) determination of cases in our e-discovery world.

Andrew Peck, Search, Forward, L. Tech. News, Oct. 2011, at 25, 29. This judicial opinion now recognizes that computer-assisted review is an acceptable way to search for relevant ESI in appropriate cases.[1]

[1] To correct the many blogs about this case, initiated by a press release from plaintiffs' vendor -- **[*4]** the Court did not order the parties to use predictive coding. The parties had agreed to defendants' use of it, but had disputes over the scope and implementation, which the Court ruled on, thus accepting the use of computer-assisted review in this lawsuit.

CASE BACKGROUND

In this action, five female named plaintiffs are suing defendant Publicis Groupe, "one of the world's 'big four' advertising conglomerates," and its United States public relations subsidiary, defendant MSL Group. (See Dkt. No. 4: Am. Compl. ¶¶ 1, 5, 26-32.) Plaintiffs allege that defendants have a "glass ceiling" that limits women to entry level positions, and that there is "systemic, company-wide gender discrimination against female PR employees like Plaintiffs." (Am. Compl. ¶¶ 4-6, 8.) Plaintiffs allege that the gender discrimination includes

> (a) paying Plaintiffs and other female PR employees less than similarly-situated male employees; (b) failing to promote or advance Plaintiffs and other female PR employees at the same rate as similarly-situated male employees; and (c) carrying out discriminatory terminations, demotions and/or job reassignments of female PR employees when the company reorganized its PR practice **[*5]** beginning in 2008 . . . .

(Am. Compl. ¶ 8.)

Plaintiffs assert claims for gender discrimination under Title VII (and under similar New York State and New York City laws) (Am. Compl. ¶¶ 204-25), pregnancy discrimination under Title VII and related violations of the Family and Medical Leave Act (Am. Compl. ¶¶ 239-71), as well as violations of the Equal Pay Act and Fair Labor Standards Act (and the similar New York Labor Law) (Am. Compl. ¶¶ 226-38).

The complaint seeks to bring the Equal Pay Act/FLSA claims as a "collective action" (i.e., opt-in) on behalf of all "current, former, and future female PR employees" employed by defendants in the United States "at any time during the applicable liability period" (Am. Compl. ¶¶ 179-80, 190-203), and as a class action on the gender and pregnancy discrimination claims and on the New York Labor

Law pay claim (Am. Compl. ¶¶ 171-98). Plaintiffs, however, have not yet moved for collective action or class certification at this time.

Defendant MSL denies the allegations in the complaint and has asserted various affirmative defenses. (See generally Dkt. No. 19: MSL Answer.) Defendant Publicis is challenging the Court's jurisdiction over it, and the parties **[*6]** have until March 12, 2012 to conduct jurisdictional discovery. (See Dkt. No. 44: 10/12/11 Order.)

COMPUTER-ASSISTED REVIEW EXPLAINED

My Search, Forward article explained my understanding of computer-assisted review, as follows:

> By computer-assisted coding, I mean tools (different vendors use different names) that use sophisticated algorithms to enable the computer to determine relevance, based on interaction with (i.e., training by) a human reviewer.
>
> Unlike manual review, where the review is done by the most junior staff, computer-assisted coding involves a senior partner (or [small] team) who review and code a "seed set" of documents. The computer identifies properties of those documents that it uses to code other documents. As the senior reviewer continues to code more sample documents, the computer predicts the reviewer's coding. (Or, the computer codes some documents and asks the senior reviewer for feedback.)
>
> When the system's predictions and the reviewer's coding sufficiently coincide, the system has learned enough to make confident predictions for the remaining documents. Typically, the senior lawyer (or team) needs to review only a few thousand documents to train the computer.
>
> Some **[*7]** systems produce a simple yes/no as to relevance, while others give a relevance score (say, on a 0 to 100 basis) that counsel can use to prioritize review. For example, a score above 50 may produce 97% of the relevant documents, but constitutes only 20% of the entire document set.
>
> Counsel may decide, after sampling and quality control tests, that documents with a score of below 15 are so highly likely to be irrelevant that no further human review is necessary. Counsel can also decide the cost-benefit of manual review of the documents with scores of 15-50.

Andrew Peck, Search, Forward, L. Tech. News, Oct. 2011, at 25, 29.[2]

> [2] From a different perspective, every person who uses email uses predictive coding, even if they do not realize it. The "spam filter" is an example of predictive coding.

My article further explained my belief that Daubert would not apply to the results of using predictive coding, but that in any challenge to its use, this Judge would be interested in both the process used and the results:

> [I]f the use of predictive coding is challenged in a case before me, I will want to know what was done and why that produced defensible results. I may be less interested in the science **[*8]** behind the "black box" of the vendor's software than in whether it produced responsive documents with reasonably high recall and high precision.
>
> That may mean allowing the requesting party to see the documents that were used to train the computer-assisted coding system. (Counsel would not be required to explain why they coded documents as responsive or non-responsive, just what the coding was.) Proof of a valid "process,"

> including quality control testing, also will be important.
>
> . . . .
>
> Of course, the best approach to the use of computer-assisted coding is to follow the Sedona Cooperation Proclamation model. Advise opposing counsel that you plan to use computer-assisted coding and seek agreement; if you cannot, consider whether to abandon predictive coding for that case or go to the court for advance approval.

Id.

THE ESI DISPUTES IN THIS CASE AND THEIR RESOLUTION

After several discovery conferences and rulings by Judge Sullivan (the then-assigned District Judge), he referred the case to me for general pretrial supervision. (Dkt. No. 48: 11/28/11 Referral Order.) At my first discovery conference with the parties, both parties' counsel mentioned that they had been discussing an "electronic **[*9]** discovery protocol," and MSL's counsel stated that an open issue was "plaintiff's reluctance to utilize predictive coding to try to cull down the" approximately three million electronic documents from the agreed-upon custodians. (Dkt. No. 51: 12/2/11 Conf. Tr. at 7-8.)[3] Plaintiffs' counsel clarified that MSL had "over simplified [plaintiffs'] stance on predictive coding," i.e., that it was not opposed but had "multiple concerns . . . on the way in which [MSL] plan to employ predictive coding" and plaintiffs wanted "clarification." (12/2/11 Conf. Tr. at 21.)

The Court did not rule but offered the parties the following advice:

> Now, if you want any more advice, for better or for worse on the ESI plan and whether predictive coding should be used, . . . I will say right now, what should not be a surprise, I wrote an article in the October Law **[*10]** Technology News called Search Forward, which says predictive coding should be used in the appropriate case.
>
> Is this the appropriate case for it? You all talk about it some more. And if you can't figure it out, you are going to get back in front of me. Key words, certainly unless they are well done and tested, are not overly useful. Key words along with predictive coding and other methodology, can be very instructive.
>
> I'm also saying to the defendants who may, from the comment before, have read my article. If you do predictive coding, you are going to have to give your seed set, including the seed documents marked as nonresponsive to the plaintiff's counsel so they can say, well, of course you are not getting any [relevant] documents, you're not appropriately training the computer.

(12/2/11 Conf. Tr. at 20-21.) The December 2, 2011 conference adjourned with the parties agreeing to further discuss the ESI protocol. (12/2/11 Conf. Tr. at 34-35.)

The ESI issue was next discussed at a conference on January 4, 2012. (Dkt. No. 71: 1/4/12 Conf. Tr.) Plaintiffs' ESI consultant conceded that plaintiffs "have not taken issue with the use of predictive coding or, frankly, with the confidence levels **[*11]** that they [MSL] have proposed . . . ." (1/4/12 Conf. Tr. at 51.) Rather, plaintiffs took issue with MSL's proposal that after the computer was fully trained and the results generated, MSL wanted to only review and produce the top 40,000 documents, which it estimated would cost $200,000 (at

[3] When defense counsel mentioned the disagreement about predictive coding, I stated that: "You must have thought you died and went to Heaven when this was referred to me," to which MSL's counsel responded: "Yes, your Honor. Well, I'm just thankful that, you know, we have a person familiar with the predictive coding concept." (12/2/11 Conf. Tr. at 8-9.)

$5 per document). (1/4/12 Conf. Tr. at 47-48, 51.) The Court rejected MSL's 40,000 documents proposal as a "pig in a poke." (1/4/12 Conf. Tr. at 51-52.) The Court explained that "where [the] line will be drawn [as to review and production] is going to depend on what the statistics show for the results," since "[p]roportionality requires consideration of results as well as costs. And if stopping at 40,000 is going to leave a tremendous number of likely highly responsive documents unproduced, [MSL's proposed cutoff] doesn't work." (1/4/12 Conf. Tr. at 51-52; see also id. at 57-58; Dkt. No. 88: 2/8/12 Conf. Tr. at 84.) The parties agreed to further discuss and finalize the ESI protocol by late January 2012, with a conference held on February 8, 2012. (1/4/12 Conf. Tr. at 60-66; see 2/8/12 Conf. Tr.)

Custodians

The first issue regarding the ESI protocol involved the selection of which custodians' emails **[*12]** would be searched. MSL agreed to thirty custodians for a "first phase." (Dkt. No. 88: 2/8/12 Conf. Tr. at 23-24.) MSL's custodian list included the president and other members of MSL's "executive team," most of its HR staff and a number of managing directors. (2/8/12 Conf. Tr. at 24.)

Plaintiffs sought to include as additional custodians seven male "comparators," explaining that the comparators' emails were needed in order to find information about their job duties and how their duties compared to plaintiffs' job duties. (2/8/12 Conf. Tr. at 25-27.) Plaintiffs gave an example of the men being given greater "client contact" or having better job assignments. (2/8/12 Conf. Tr. at 28-30.) The Court held that the search of the comparators' emails would be so different from that of the other custodians that the comparators should not be included in the emails subjected to predictive coding review. (2/8/12 Conf. Tr. at 28, 30.) As a fallback position, plaintiffs proposed to "treat the comparators as a separate search," but the Court found that plaintiffs could not describe in any meaningful way how they would search the comparators' emails, even as a separate search. (2/8/12 Conf. Tr. at 30-31.) **[*13]** Since the plaintiffs likely could develop the information needed through depositions of the comparators, the Court ruled that the comparators' emails would not be included in phase one. (2/8/12 Conf. Tr. at 31.)

Plaintiffs also sought to include MSL's CEO, Olivier Fleuriot, located in France and whose emails were mostly written in French. (2/8/12 Conf. Tr. at 32-34.) The Court concluded that because his emails with the New York based executive staff would be gathered from those custodians, and Fleuriot's emails stored in France likely would be covered by the French privacy and blocking laws,[4] Fleuriot should not be included as a first-phase custodian. (2/8/12 Conf. Tr. at 35.)

[4] See, e.g., *Societe Nationale Industrielle Aerospatiale v. U.S. Dist. Ct. for the S.D. of Iowa, 482 U.S. 522, 107 S. Ct. 2542, 96 L. Ed. 2d 461 (1987)*; see also The Sedona Conference, International Principles on Discovery, Disclosure & Data Protection (2011), available at http://www.thesedonaconference.org/dltForm?did=IntlPrinciples2011.pdf.

Plaintiffs sought to include certain managing directors from MSL offices at which no named plaintiff worked. (2/8/12 Conf. Tr. at 36-37.) The Court ruled that since plaintiffs had not yet moved **[*14]** for collective action status or class certification, until the motions were made and granted, discovery would be limited to offices (and managing directors) where the named plaintiffs had worked. (2/8/12 Conf. Tr. at 37-39.)

The final issue raised by plaintiffs related to the phasing of custodians and the discovery cutoff dates. MSL proposed finishing phase-one discovery completely before considering what to do about a second phase. (See 2/8/12 Conf. Tr. at 36.) Plaintiffs expressed concern that there would not be time for two separate phases, essentially seeking to move the phase-two custodians back into phase one. (2/8/12 Conf. Tr. at 35-36.) The Court found MSL's separate phase approach to be more sensible and noted that if necessary, the Court would extend the discovery cutoff to allow the parties to pursue discovery in phases. (2/8/12 Conf. Tr. at 36, 50.)

Sources of ESI

The parties agreed on certain ESI sources, including the "EMC SourceOne [Email] Archive," the

"PeopleSoft" human resources information management system and certain other sources including certain HR "shared" folders. (See Dkt. No. 88: 2/8/12 Conf. Tr. at 44-45, 50-51.) As to other "shared" folders, neither side **[*15]** was able to explain whether the folders merely contained forms and templates or collaborative working documents; the Court therefore left those shared folders for phase two unless the parties promptly provided information about likely contents. (2/8/12 Conf. Tr. at 47-48.)

The Court noted that because the named plaintiffs worked for MSL, plaintiffs should have some idea what additional ESI sources, if any, likely had relevant information; since the Court needed to consider proportionality pursuant to *Rule 26(b)(2)(C)*, plaintiffs needed to provide more information to the Court than they were doing if they wanted to add additional data sources into phase one. (2/8/12 Conf. Tr. at 49-50.) The Court also noted that where plaintiffs were getting factual information from one source (e.g., pay information, promotions, etc.), "there has to be a limit to redundancy" to comply with *Rule 26(b)(2)(C)*. (2/8/12 Conf. Tr. at 54.)[5]

[5] The Court also suggested that the best way to resolve issues about what information might be found in a certain source is for MSL to show plaintiffs a sample printout from that source. (2/8/12 Conf. Tr. at 55-56.)

The Predictive Coding Protocol

The parties agreed to use a **[*16]** 95% confidence level (plus or minus two percent) to create a random sample of the entire email collection; that sample of 2,399 documents will be reviewed to determine relevant (and not relevant) documents for a "seed set" to use to train the predictive coding software. (Dkt. No. 88: 2/8/12 Conf. Tr. at 59-61.) An area of disagreement was that MSL reviewed the 2,399 documents before the parties agreed to add two additional concept groups (i.e., issue tags). (2/8/12 Conf. Tr. at 62.) MSL suggested that since it had agreed to provide all 2,399 documents (and MSL's coding of them) to plaintiffs for their review, plaintiffs can code them for the new issue tags, and MSL will incorporate that coding into the system. (2/8/12 Conf. Tr. at 64.) Plaintiffs' vendor agreed to that approach. (2/8/12 Conf. Tr. at 64.)

To further create the seed set to train the predictive coding software, MSL coded certain documents through "judgmental sampling." (2/8/12 Conf. Tr. at 64.) The remainder of the seed set was created by MSL reviewing "keyword" searches with Boolean connectors (such as "training and Da Silva Moore," or "promotion and Da Silva Moore") and coding the top fifty hits from those searches. **[*17]** (2/8/12 Conf. Tr. at 64-66, 72.) MSL agreed to provide all those documents (except privileged ones) to plaintiffs for plaintiffs to review MSL's relevance coding. (2/8/12 Conf. Tr. at 66.) In addition, plaintiffs provided MSL with certain other keywords, and MSL used the same process with plaintiffs' keywords as with the MSL keywords, reviewing and coding an additional 4,000 documents. (2/8/12 Conf. Tr. at 68-69, 71.) All of this review to create the seed set was done by senior attorneys (not paralegals, staff attorneys or junior associates). (2/8/12 Conf. Tr. at 92-93.) MSL reconfirmed that "[a]ll of the documents that are reviewed as a function of the seed set, whether [they] are ultimately coded relevant or irrelevant, aside from privilege, will be turned over to" plaintiffs. (2/8/12 Conf. Tr. at 73.)

The next area of discussion was the iterative rounds to stabilize the training of the software. MSL's vendor's predictive coding software ranks documents on a score of 100 to zero, i.e., from most likely relevant to least likely relevant. (2/8/12 Conf. Tr. at 70.) MSL proposed using seven iterative rounds; in each round they would review at least 500 documents from different concept **[*18]** clusters to see if the computer is returning new relevant documents. (2/8/12 Conf. Tr. at 73-74.) After the seventh round, to determine if the computer is well trained and stable, MSL would review a random sample (of 2,399 documents) from the discards (i.e., documents coded as non-relevant) to make sure the documents determined by the software to not be relevant do not, in fact, contain highly-relevant documents. (2/8/12 Conf. Tr. at 74-75.) For each of the seven rounds and the final quality-check random sample, MSL agreed that it would show plaintiffs all the documents it looked at including those deemed not relevant (except for privileged documents). (2/8/12 Conf. Tr. at 76.)

Plaintiffs' vendor noted that "we don't at this point agree that this is going to work. This is new technology and it has to be proven out." (2/8/12 Conf. Tr. at 75.) Plaintiffs' vendor agreed, in general, that computer-assisted review works, and works better than most alternatives. (2/8/12 Conf. Tr. at 76.) Indeed, plaintiffs' vendor noted that "it is fair to say [that] we are big proponents of it." (2/8/12 Conf. Tr. at 76.) The Court reminded the parties that computer-assisted review "works better than most **[*19]** of the alternatives, if not all of the [present] alternatives. So the idea is not to make this perfect, it's not going to be perfect. The idea is to make it significantly better than the alternatives without nearly as much cost." (2/8/12 Conf. Tr. at 76.)

The Court accepted MSL's proposal for the seven iterative reviews, but with the following caveat:

> But if you get to the seventh round and [plaintiffs] are saying that the computer is still doing weird things, it's not stabilized, etc., we need to do another round or two, either you will agree to that or you will both come in with the appropriate QC information and everything else and [may be ordered to] do another round or two or five or 500 or whatever it takes to stabilize the system.

(2/8/12 Conf. Tr. at 76-77; see also id. at 83-84, 88.)

On February 17, 2012, the parties submitted their "final" ESI Protocol which the Court "so ordered." (Dkt. No. 92: 2/17/12 ESI Protocol & Order.)[6] Because this is the first Opinion dealing with predictive coding, the Court annexes hereto as an Exhibit the provisions of the ESI Protocol dealing with the predictive coding search methodology.

OBSERVATIONS ON PLAINTIFF'S OBJECTIONS TO THE COURT'S RULINGS

On February 22, 2012, plaintiffs filed objections to the Court's February 8, 2012 rulings. (Dkt. No. 93: Pls. *Rule 72(a)* Objections; see also Dkt. No. 94: Nurhussein Aff.; Dkt. No. 95: Neale Aff.) While those objections are before District Judge Carter, a few comments are in order.

Plaintiffs' Reliance on Rule 26(g)(1)(A) is Erroneous

Plaintiffs' objections to my February 8, 2012 rulings assert that my acceptance of MSL's predictive coding approach "provides unlawful 'cover' for MSL's counsel, who has a duty under *FRCP 26(g)* to 'certify' that their client's document production is 'complete' and 'correct' as of the time it was made. *FRCP 26(g)(1)(A)*." (Dkt. No. 93: Pls. *Rule 72(a)* Objections at 8 n.7; accord, **[*21]** id. at 2.) In large-data cases like this, involving over three million emails, no lawyer using any search method could honestly certify that its production is "complete" -- but more importantly, *Rule 26(g)(1)* does not require that. Plaintiffs simply misread *Rule 26(g)(1)*. The certification required by *Rule 26(g)(1)* applies "with respect to a disclosure." *Fed. R. Civ. P. 26(g)(1)(A)* (emphasis added). That is a term of art, referring to the mandatory initial disclosures required by *Rule 26(a)(1)*. Since the *Rule 26(a)(1)* disclosure is information (witnesses, exhibits) that "the disclosing party may use to support its claims or defenses," and failure to provide such information leads to virtually automatic preclusion, see *Fed. R Civ. P. 37(c)(1)*, it is appropriate for the

[6] Plaintiffs included a paragraph noting its objection to the **[*20]** ESI Protocol, as follows:

> Plaintiffs object to this ESI Protocol in its entirety. Plaintiffs submitted their own proposed ESI Protocol to the Court, but it was largely rejected. The Court then ordered the parties to submit a joint ESI Protocol reflecting the Court's rulings. Accordingly, Plaintiffs jointly submit this ESI Protocol with MSL, but reserve the right to object to its use in this case.

(ESI Protocol ¶ J.1 at p. 22.)

Rule 26(g)(1)(A) certification to require disclosures be "complete and correct."

Rule 26(g)(1)(B) is the provision that applies to discovery responses. It does not call for certification that the discovery response is "complete," but rather incorporates the *Rule 26(b)(2)(C)* proportionality principle. Thus, *Rule 26(g)(1)(A)* has absolutely nothing to do with MSL's obligations to respond to plaintiffs' discovery requests. **[*22]** Plaintiffs' argument is based on a misunderstanding of *Rule 26(g)(1)*.[7]

[7] *Rule 26(g)(1)* provides:

> (g) Signing Disclosures and Discovery Requests, Responses, and Objections.
>
> > (1) *Signature Required; Effect of Signature*. Every disclosure under *Rule 26(a)(1)* or *(a)(3)* and every discovery request, response, or objection must be signed by at least one attorney of record in the attorney's own name . . . . By signing, an attorney or party certifies that to the best of the person's knowledge, information, and belief formed after a reasonable inquiry:
> >
> > > (A) with respect to a disclosure, it is complete and correct as of the time it is made; and
> > >
> > > (B) with respect to a discovery request, response, or objection, it is:
> > >
> > > > (i) consistent with these rules and warranted by existing law or by a nonfrivolous argument for extending, modifying, or reversing existing law, or for establishing new law;
> > > >
> > > > (ii) not interposed

> for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation; and
>
> (iii) neither unreasonable nor unduly burdensome or expensive, considering the needs of the case, prior discovery in the case, the amount in controversy, and the importance of the issues **[*23]** at stake in the action.

Fed. R. Civ. P. 26(g)(1) (emphasis added).

Rule 702 and Daubert Are Not Applicable to Discovery Search Methods

Plaintiffs' objections also argue that my acceptance of MSL's predictive coding protocol "is contrary to *Federal Rule of Evidence 702*" and "violates the gatekeeping function underlying *Rule 702*." (Dkt. No. 93: Pls. *Rule 72(a)* Objections at 2-3; accord, id. at 10-12.)[8]

> [8] As part of this argument, plaintiffs complain that although both parties' experts (i.e., vendors) spoke at the discovery conferences, they were not sworn in. (Pls. *Rule 72(a)* Objections at 12: "To his credit, the Magistrate [Judge] did ask the parties to bring [to the conference] the ESI experts they had hired to advise them regarding the creation of an ESI protocol. These experts, however, were never sworn in, and thus the statements they made in court at the hearings were not sworn testimony made under penalty of perjury.") Plaintiffs never asked the Court to have the experts testify to their qualifications or be sworn in.

Federal Rule of Evidence 702 and the Supreme Court's Daubert decision[9] deal with the trial court's role as gatekeeper to exclude unreliable expert testimony from being **[*24]** submitted to the jury at trial. See also Advisory Comm. Notes to *Fed. R. Evid. 702*. It is a rule for admissibility of evidence at trial.

> [9] *Daubert v. Merrell Dow Pharms., Inc., 509 U.S. 579, 113 S. Ct. 2786, 125 L. Ed. 2d 469 (1993).*

If MSL sought to have its expert testify at trial and introduce the results of its ESI protocol into evidence, Daubert and *Rule 702* would apply. Here, in contrast, the tens of thousands of emails that will be produced in discovery are not being offered into evidence at trial as the result of a scientific process or otherwise. The admissibility of specific emails at trial will depend upon each email itself (for example, whether it is hearsay, or a business record or party admission), not how it was found during discovery.

Rule 702 and Daubert simply are not applicable to how documents are searched for and found in discovery.

Plaintiffs' Reliability Concerns Are, At Best, Premature

Finally, plaintiffs' objections assert that "MSL's method lacks the necessary standards for assessing whether its results are accurate; in other words, there is no way to be certain if MSL's method is reliable." (Dkt. No. 93: Pls. *Rule 72(a)* Objections at 13-18.) Plaintiffs' concerns may be appropriate **[*25]** for resolution during or after the process (which the Court will be closely supervising), but are premature now. For example, plaintiffs complain that "MSL's method fails to include an agreed-upon standard of relevance that is transparent and accessible to all parties. . . . Without this standard, there is a high-likelihood of delay as the parties resolve disputes with regard to individual documents on a case-by-case basis." (Id. at 14.) Relevance is determined by plaintiffs' document demands. As statistics show, perhaps only 5% of the disagreement among reviewers comes from close questions of relevance, as opposed to reviewer error. (See page 18 n.11 below.) The issue regarding relevance standards might be significant if MSL's proposal was not totally transparent. Here, however, plaintiffs will see how MSL has coded every email used in the seed set (both relevant and not relevant), and the Court is available to quickly resolve any issues.

Plaintiffs complain they cannot determine if "MSL's method actually works" because MSL does not describe how many relevant documents are permitted to be located in the final random sample of documents the software deemed irrelevant. (Pls. *Rule 72(a)* **[*26]** Objections at 15-16.) Plaintiffs argue that "without any decision about this made in advance, the Court is simply kicking the can down the road." (Id. at 16.) In order to determine proportionality, it is necessary to have more information than the parties (or the Court) now has, including how many relevant documents will be produced and at what cost to MSL. Will the case remain limited to the named plaintiffs, or will plaintiffs seek and obtain collective action and/or class action certification? In the final sample of documents deemed irrelevant, are any relevant documents found that are "hot," "smoking gun" documents (i.e., highly relevant)? Or are the only relevant documents more of the same thing? One hot document may require the software to be re-trained (or some other search method employed), while several documents that really do not add anything to the case might not matter. These types of questions are better decided "down the road," when real information is available to the parties and the Court.

FURTHER ANALYSIS AND LESSONS FOR THE FUTURE

The decision to allow computer-assisted review in this case was relatively easy -- the parties agreed to its use (although disagreed about **[*27]** how best to implement such review). The Court recognizes that computer-assisted review is not a magic, Staples-Easy-Button, solution appropriate for all cases. The technology exists and should be used where appropriate, but it is not a case of machine replacing humans: it is the process used and the interaction of man and machine that the courts needs to examine.

The objective of review in ediscovery is to identify as many relevant documents as possible, while reviewing as few non-relevant documents as possible. Recall is the fraction of relevant documents identified during a review; precision is the fraction of identified documents that are relevant. Thus, recall is a measure of completeness, while precision is a measure of accuracy or correctness. The goal is for the review method to result in higher recall and higher precision than another review method, at a cost proportionate to the "value" of the case. See, e.g., Maura R. Grossman & Gordon V. Cormack, Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, Rich. J.L.& Tech., Spring 2011, at 8-9, available at http://jolt.richmond.edu/v17i3/article11.pdf.

The slightly more **[*28]** difficult case would be where the producing party wants to use computer-assisted review and the requesting party objects.[10] The question to ask in that situation is what methodology would the requesting party suggest instead? Linear manual review is simply too expensive where, as here, there are over three million emails to review. Moreover, while some lawyers still consider manual review to be the "gold standard," that is a myth, as statistics clearly show that computerized searches are at least as accurate, if not more so, than manual review. Herb Roitblatt, Anne Kershaw, and Patrick Oot of the Electronic Discovery Institute conducted an empirical assessment to "answer the question of whether there was a benefit to engaging in a traditional human review or whether computer systems could be relied on to produce comparable results," and concluded that "[o]n every measure, the performance of the two computer systems was at least as accurate (measured against the original review) as that of human re-review." Herbert L. Roitblatt, Anne Kershaw & Patrick Oot, Document Categorization in Legal Electronic Discovery: Computer Classification vs. Manual Review, 61 J. Am. Soc'y for Info. Sci. **[*29]** & Tech. 70, 79 (2010).[11]

[10] The tougher question, raised in Klein Prods. LLC v. Packaging Corp. of Am. before Magistrate Judge Nan Nolan in Chicago, is whether the Court, at plaintiffs' request, should order the defendant to use computer-assisted review to respond to plaintiffs' document requests.

[11] The Roitblatt, Kershaw, Oot article noted that "[t]he level of agreement among human reviewers is not strikingly high," around 70-75%. They identify two sources for this variability: fatigue ("A document that they [the reviewers] might have categorized as responsive when they were more attentive might then be categorized [when the reviewer is distracted or fatigued] as non-responsive or vice versa."), and differences in "strategic judgment." Id. at 77-78. Another study found that responsiveness "is fairly well defined, and that disagreements among assessors are largely attributable to human error," with only 5% of reviewer disagreement attributable to borderline or questionable issues as to relevance. Maura R. Grossman & Gordon V. Cormack, Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error? 9 (DESI IV: 2011 ICAIL Workshop on Setting Standards for **[*30]** Searching Elec. Stored Info. in Discovery, Research Paper), *available at* http://www.umiacs.umd.edu/~oard/desi4/papers/grossman3.pdf.

Likewise, Wachtell, Lipton, Rosen & Katz litigation counsel Maura Grossman and University of Waterloo professor Gordon Cormack, studied data from the Text Retrieval Conference Legal Track (TREC) and concluded that: "[T]he myth that exhaustive manual review is the most effective -- and therefore the most defensible -- approach to document review is strongly refuted. Technology-assisted review can (and does) yield more accurate results than exhaustive manual review, with much lower effort." Maura R. Grossman & Gordon V. Cormack, Technology-Assisted Review in E-Discovery Can Be More Effective and More Efficient Than Exhaustive Manual Review, Rich. J.L.& Tech., Spring 2011, at 48.[12] The technology-assisted reviews in the Grossman-Cormack article also demonstrated significant cost savings over manual review: "The technology-assisted reviews require, on average, human review of only 1.9% of the documents, a fifty-fold savings over exhaustive manual review." Id. at 43.

Because of the volume of ESI, lawyers frequently have turned to keyword searches to cull email (or other ESI) down to a more manageable volume for further manual review. Keywords have a place in production of ESI -- indeed, the parties here used keyword searches (with Boolean connectors) to find documents for the expanded seed set to train the predictive coding software. In too many cases, however, the way lawyers choose keywords is the equivalent of the child's game of "Go Fish."[13] The requesting party guesses which keywords might produce evidence to support its case without having much, if any, knowledge of the responding party's "cards" (i.e., the terminology used by the responding party's custodians). Indeed, the responding party's counsel often does not know what is in its own client's "cards."

Another problem with keywords is that they often are over-inclusive, **[*32]** that is, they find responsive documents but also large numbers of irrelevant documents. In this case, for example, a keyword search for "training" resulted in 165,208 hits; Da Silva Moore's name resulted in 201,179 hits; "bonus" resulted in 40,756 hits; "compensation" resulted in 55,602 hits; and "diversity" resulted in 38,315 hits. (Dkt. No. 92: 2/17/12 ESI Protocol Ex. A.) If MSL had to manually review all of the keyword hits, many of which would not be relevant (i.e., would be false positives), it would be quite costly.

Moreover, keyword searches usually are not very effective. In 1985, scholars David Blair and M. Maron collected 40,000 documents from a Bay Area Rapid Transit accident, and instructed experienced attorney and paralegal searchers to use keywords and other review techniques to retrieve at least 75% of the documents relevant to 51 document requests. David L. Blair & M. E. Maron, An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System, 28 Comm. ACM 289 (1985). Searchers believed they met the goals, but their average recall was just 20%. Id. This result has been replicated in the TREC Legal Track studies over the past few years.

Judicial decisions **[*33]** have criticized specific keyword searches. Important early decisions in this area came from two of the leading judicial scholars in ediscovery, Magistrate Judges John Facciola (District of Columbia) and Paul Grimm (Maryland). See *United States v. O'Keefe, 537 F. Supp. 2d 14, 24 (D.D.C. 2008)* (Facciola, M.J.); *Equity Analytics, LLC v. Lundin, 248 F.R.D. 331, 333 (D.D.C. 2008)* (Facciola, M.J.); *Victor Stanley, Inc. v. Creative Pipe, Inc., 250 F.R.D. 251, 260, 262 (D. Md. 2008)* (Grimm, M.J.). I followed their lead with Willaim A. Gross Construction Associates, Inc., when I wrote:

> 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").

[12] Grossman and Cormack also note that "not **[*31]** all technology-assisted reviews . . . are created equal" and that future studies will be needed to "address which technology-assisted review process(es) will improve most on manual review." Id.

[13] See Ralph C. Losey, "Child's Game of 'Go Fish' is a Poor Model for e-Discovery Search," in Adventures in Electronic Discovery 209-10 (2011).

> . . . .
>
> Electronic discovery requires cooperation between opposing counsel and transparency in all aspects of preservation and production of ESI. Moreover, where counsel are using keyword searches for retrieval of ESI, they at a minimum must carefully craft the appropriate keywords, with input from the ESI's custodians as to **[*34]** the words and abbreviations they use, and the proposed methodology must be quality control tested to assure accuracy in retrieval and elimination of "false positives." It is time that the Bar -- even those lawyers who did not come of age in the computer era -- understand this.

William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. 134, 134, 136 (S.D.N.Y. 2009) (Peck, M.J.).

Computer-assisted review appears to be better than the available alternatives, and thus should be used in appropriate cases. While this Court recognizes that computer-assisted review is not perfect, the Federal Rules of Civil Procedure do not require perfection. See, e.g., *Pension Comm. of Univ. of Montreal Pension Plan v. Banc of Am. Sec., 685 F. Supp. 2d 456, 461 (S.D.N.Y. 2010)*. Courts and litigants must be cognizant of the aim of *Rule 1*, to "secure the just, speedy, and inexpensive determination" of lawsuits. *Fed. R. Civ. P. 1*. That goal is further reinforced by the proportionality doctrine set forth in *Rule 26(b)(2)(C)*, which provides that:

> On **[*35]** motion or on its own, the court must limit the frequency or extent of discovery otherwise allowed by these rules or by local rule if it determines that:
>
>> (i) the discovery sought is unreasonably cumulative or duplicative, or can be obtained from some other source that is more convenient, less burdensome, or less expensive;
>>
>> (ii) the party seeking discovery has had ample opportunity to obtain the information by discovery in the action; or
>>
>> (iii) the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues.

Fed. R. Civ. P. 26(b)(2)(C).

In this case, the Court determined that the use of predictive coding was appropriate considering: (1) the parties' agreement, (2) the vast amount of ESI to be reviewed (over three million documents), (3) the superiority of computer-assisted review to the available alternatives (i.e., linear manual review or keyword searches), (4) the need for cost effectiveness and proportionality under *Rule 26(b)(2)(C)*, and (5) the transparent process proposed by **[*36]** MSL.

This Court was one of the early signatories to The Sedona Conference Cooperation Proclamation, and has stated that "the best solution in the entire area of electronic discovery is cooperation among counsel. This Court strongly endorses The Sedona Conference Proclamation (available at www.TheSedonaConference.org)." *William A. Gross Constr. Assocs., Inc. v. Am. Mfrs. Mut. Ins. Co., 256 F.R.D. at 136*. An important aspect of cooperation is transparency in the discovery process. MSL's transparency in its proposed ESI search protocol made it easier for the Court to approve the use of predictive coding. As discussed above on page

10, MSL confirmed that "[a]ll of the documents that are reviewed as a function of the seed set, whether [they] are ultimately coded relevant or irrelevant, aside from privilege, will be turned over to" plaintiffs. (Dkt. No. 88: 2/8/12 Conf. Tr. at 73; see also 2/17/12 ESI Protocol at 14: "MSL will provide Plaintiffs' counsel with all of the non-privileged documents and will provide, to the extent applicable, the issue tag(s) coded for each document . . . . If necessary, counsel will meet and confer to attempt to resolve any disagreements regarding the coding **[*37]** applied to the documents in the seed set.") While not all experienced ESI counsel believe it necessary to be as transparent as MSL was willing to be, such transparency allows the opposing counsel (and the Court) to be more comfortable with computer-assisted review, reducing fears about the so-called "black box" of the technology.[14] This Court highly recommends that counsel in future cases be willing to at least discuss, if not agree to, such transparency in the computer-assisted review process.

[14] It also avoids the GIGO problem, i.e., garbage in, garbage out.

Several other lessons for the future can be derived from the Court's resolution of the ESI discovery disputes in this case.

First, it is unlikely that courts will be able to determine or approve a party's proposal as to when review and production can stop until the computer-assisted review software has been trained and the results are quality control verified. Only at that point can the parties and the Court see where there is a clear drop off from highly relevant to marginally relevant to not likely to be relevant documents. While cost is a factor under *Rule 26(b)(2)(C)*, it cannot be considered in isolation from the results of the **[*38]** predictive coding process and the amount at issue in the litigation.

Second, staging of discovery by starting with the most likely to be relevant sources (including custodians), without prejudice to the requesting party seeking more after conclusion of that first stage review, is a way to control discovery costs. If staging requires a longer discovery period, most judges should be willing to grant such an extension. (This Judge runs a self-proclaimed "rocket docket," but informed the parties here of the Court's willingness to extend the discovery cutoff if necessary to allow the staging of custodians and other ESI sources.)

Third, in many cases requesting counsel's client has knowledge of the producing party's records, either because of an employment relationship as here or because of other dealings between the parties (e.g., contractual or other business relationships). It is surprising that in many cases counsel do not appear to have sought and utilized their client's knowledge about the opposing party's custodians and document sources. Similarly, counsel for the producing party often is not sufficiently knowledgeable about their own client's custodians and business terminology. Another **[*39]** way to phrase cooperation is "strategic proactive disclosure of information," i.e., if you are knowledgeable about and tell the other side who your key custodians are and how you propose to search for the requested documents, opposing counsel and the Court are more apt to agree to your approach (at least as phase one without prejudice).

Fourth, the Court found it very helpful that the parties' ediscovery vendors were present and spoke at the court hearings where the ESI Protocol was discussed. (At ediscovery programs, this is sometimes jokingly referred to as "bring your geek to court day.") Even where as here counsel is very familiar with ESI issues, it is very helpful to have the parties' ediscovery vendors (or in-house IT personnel or in-house ediscovery counsel) present at court conferences where ESI issues are being discussed. It also is important for the vendors and/or knowledgeable counsel to be able to explain complicated ediscovery concepts in ways that make it easily understandable to judges who may not be tech-savvy.

CONCLUSION

This Opinion appears to be the first in which a Court has approved of the use of computer-assisted review. That does not mean computer-assisted review **[*40]** must be used in all cases, or that the exact ESI protocol approved here will be appropriate in all future cases that utilize computer-assisted review. Nor does this Opinion endorse any vendor (the Court was very careful not to mention the names of the parties' vendors in the body of this Opinion, although it is revealed in the attached ESI Protocol), nor any particular computer-assisted review tool. What the Bar should take away from this Opinion is that computer-assisted review is an available tool and should be seriously considered for use in large-data-volume cases where it may save the producing party (or both parties) significant amounts of legal fees in document review. Counsel no longer have to worry about being the "first" or "guinea pig" for judicial acceptance of computer-assisted review. As with keywords or any other technological solution to ediscovery, counsel must design an appropriate process, including use of available technology, with appropriate quality control testing, to review and produce relevant ESI while adhering to *Rule 1* and *Rule 26(b)(2)(C)* proportionality. Computer-assisted review now can be considered judicially-approved for use in appropriate cases.

SO **[*41]** ORDERED.

Dated: New York, New York

February 24, 2012

/s/ Andrew J. Peck

Andrew J. Peck

United States Magistrate Judge

EXHIBIT

PARTIES' PROPOSED PROTOCOL RELATING TO THE PRODUCTION OF ELECTRONICALLY STORED INFORMATION ("ESI")

A. Scope

1. General. The procedures and protocols outlined herein govern the production of electronically stored information ("ESI") by MSLGROUP Americas, Inc. ("MSL") during the pendency of this litigation. The parties to this protocol will take reasonable steps to comply with this agreed-upon protocol for the production of documents and information existing in electronic format. Nothing in this protocol will be interpreted to require disclosure of documents or information protected from disclosure by the attorney-client privilege, work-product doctrine or any other applicable privilege or immunity. It is Plaintiffs' position that nothing in this protocol will be interpreted to waive Plaintiffs' right to object to this protocol as portions of it were mandated by the Court over Plaintiffs' objections, including Plaintiffs' objections to the predictive coding methodology proposed by MSL.

2. Limitations and No-Waiver. This protocol provides a general framework for **[*42]** the production of ESI on a going forward basis. The Parties and their attorneys do not intend by this protocol to waive their rights to the attorney work-product privilege, except as specifically required herein, and any such waiver shall be strictly and narrowly construed and shall not extend to other matters or information not specifically described herein. All Parties preserve their attorney client privileges and other privileges and there is no intent by the protocol, or the production of documents pursuant to the protocol, to in any way waive or weaken these privileges. All documents produced hereunder are fully protected and covered by the Parties' confidentiality and clawback agreements and orders of the Court effectuating same.

3. Relevant Time Period. January 1, 2008 through February 24, 2011 for all non-email ESI relating to topics besides pay discrimination and for all e-mails. January 1, 2005 through February 24, 2011 for all non-e-mail ESI relating to pay discrimination for New York Plaintiffs.

B. ESI Preservation

1. MSL has issued litigation notices to designated employees on February 10, 2010, March 14, 2011 and June 9, 2011.

C. Sources

1. The Parties have identified the **[*43]** following sources of potentially discoverable ESI at MSL. Phase I sources will be addressed first, and Phase II sources will be addressed after Phase I source searches are complete. Sources marked as "N/A" will not be searched by the Parties.

Data Source	Description	Phase

	Data Source	Description	Phase
a	EMC SourceOne Archive	Archiving System used to capture and store all incoming and outbound e-mails and selected instant message conversations saved through IBM Sametime (see below).	I
b	Lotus Notes E-mail	Active corporate system that provides e-mail communication and calendaring functions.	N/A
c	GroupWise E-mail	Legacy corporate system that provided e-mail communication and calendarin, functions.	N/A
d	IBM Sametime	Lotus Notes Instant Messaging and collaboration application.	N/A
e	Home Directories	Personal network storage locations on the file server(s) dedicated to individual users. (With the exception of 2 home directories for which MSL will collect and analyze the data to determine the level of duplication as compared to the EMC SourceOne Archive. The parties will meet and confer regarding the selection of the two custodians.)	II
f	Shared Folders	Shared network storage locations on the file server(s) that are accessible by individual users, groups of users or entire departments. (With the exception of the following Human Resources shared folders which will be in Phase I: Corporate HR, North America HR and New York HR.)	II
g	Database Servers	Backend databases (e.g. Oracle, SQL, MySQL) used to store information for front end applications or other purposes.	N/A
h	Halogen Software	Performance management program provided by Halogen to conduct performance evaluations.	I
i	Noovoo	Corporate Intranet site.	II
j	Corporate Feedback	E-mail addresses that employees may utilize to provide the company with comments, suggestions and overall feedback.	I
k	Hyperion Financial Management ("HFM")	Oracle application that offers global financial consolidation, reporting and analysis.	N/A
l	Vurv/Taleo	Talent recruitment software.	II
m	ServiceNow	Help Desk application used to track employee computer related requests.	N/A

	Data Source	Description	Phase
n	PeopleSoft	Human resources information management system.	I
o	PRISM	PeopleSoft component used for time and billing management.	I
p	Portal	A project based portal provided through Oracle/BEA Systems.	II
q	Desktops/Laptops	Fixed and portable computers provided to employees to perform work related activities. (With the exception of 2 desktop/laptop hard drives for which MSL will collect and analyze the data to determine the level of duplication as compared to the EMC SourceOne Archive. The parties will meet and confer regarding the selection of the two custodians.)	II
r	Publicis Benefits Connection	Web based site that maintains information about employee benefits and related information.	II
s	GEARS	Employee expense reporting system.	II
t	MS&L City	Former corporate Intranet.	N/A
u	Adium	Application which aggregates instant messages.	N/A
y	Pidgin	Application which aggregates instant message.	N/A
w	IBM Lotus Traveler and MobileIron	Mobile device synchronization and security system.	N/A
y	Mobile Communication Devices	Portable PDAs, smart phones, tablets used for communication.	N/A
z	Yammer	Social media and collaboration portal.	N/A
aa	SalesForce.com	Web-based customer relationship management application.	N/A
bb	Removable Storage Devices	Portable storage media, external hard drives, thumb drives, etc. used to store copies of work related ESI.	N/A

a. [*44] EMC SourceOne - MSL uses SourceOne, an EMC e-mail archiving system that captures and stores all e-mail messages that pass through the corporate e-mail system. In addition, if a user chooses to save an instant messaging chat conversation from IBM Sametime (referenced below), that too would be archived in SourceOne. Defendant MSL also acknowledges that calendar items are regularly ingested into the SourceOne system. The parties have agreed that this data source will be handled as outlined in section E below.

b. Lotus Notes E-mail - MSL currently maintains multiple Lotus Notes Domino servers in various data centers around the world. All e-mail communication and calendar items are journaled in real time to the EMC SourceOne archive. The parties have agreed to not collect any information from this data source at this time.

c. GroupWise E-mail -- Prior to the implementation of the Lotus Notes environment, GroupWise was used for all e-mail and calendar functionality. Before the decommissioning of the GroupWise servers, MSL created backup tapes of all servers that housed the GroupWise e-mail databases. The parties have agreed to not collect any information from this data source at this time.

d. **[*45]** IBM Sametime -- MSL provides custodians with the ability to have real time chat conversations via the IBM Sametime application that is part of the Lotus Notes suite of products.

e. Home Directories -- Custodians with corporate network access at MSL also have a dedicated and secured network storage location where they are able to save files. MSL will collect the home directory data for 2 custodians and analyze the data to determine the level of duplication of documents in this data source against the data contained in the EMC SourceOne archive for the same custodians. (The parties will meet and confer regarding the selection of the two custodians.) The results of the analysis will be provided to Plaintiffs so that a determination can be made by the parties as to whether MSL will include this data source in its production of ESI to Plaintiffs. If so, the parties will attempt to reach an agreement as to the approach used to collect, review and produce responsive and non-privileged documents.

f. Shared Folders -- Individual employees, groups of employees and entire departments at MSL are given access to shared network storage locations to save and share files. As it relates to the Human Resources **[*46]** related shared folders (i.e., North America HR Drive (10.2 GB), Corporate HR Drive (440 MB), NY HR Drive (1.9 GB), Chicago HR Drive (1.16 GB), Boston HR Drive (43.3 MB), and Atlanta HR Drive (6.64 GB)), MSL will judgmentally review and produce responsive and non-privileged documents from the North America HR Drive, Corporate HR Drive, and NY HR Drive. MSL will produce to Plaintiffs general information regarding the content of other Shared Folders. The parties will meet and confer regarding the information gathered concerning the other Shared Folders and discuss whether any additional Shared Folders should be moved to Phase I.

g. Database Servers -- MSL has indicated that it does not utilize any database servers, other than those that pertain to the sources outlined above in C, which are likely to contain information relevant to Plaintiffs' claims.

h. Halogen Software -- MSL **[*47]** utilizes a third party product, Halogen, for performance management and employee evaluations. The parties will meet and confer in order to exchange additional information and attempt to reach an agreement as to the scope of data and the approach used to collect, review and produce responsive and non-privileged documents.

i. Noovoo -- MSL maintains a corporate Intranet site called "Noovoo" where employees are able to access Company-related information. MSL will provide Plaintiffs with any employment-related policies maintained within Noovoo.

j. Corporate Feedback -- MSL has maintained various e-mail addresses that employees may utilize to provide the company with comments, suggestions and overall feedback. These e-mail addresses include "powerofone@mslworldwide.com", "poweroftheindividual@mslworldwide.com", "townhall@mslworldwide.com" and "whatsonyourmind@mslworldwide.com". The parties have agreed that all responsive and non-privileged ESI will be produced from these e-mail accounts and any other e-mail accounts that fall under this category of information. At present, MSL intends to manually review the contents of each of these e-mail accounts. However, if after collecting the contents **[*48]** of each of the e-mail accounts MSL determines that a manual review would be impractical, the parties will meet and confer as to the approach used to collect, review and produce responsive and non-privileged documents.

k. Hyperion Financial Management ("HFM") -- MSL uses an Oracle application called HFM that offers global financial consolidation, reporting and analysis capabilities.

l. Vurv/Taleo -- **[*49]** Since approximately 2006, MSL used an application known as Vury as its talent recruitment software. As of August 31, 2011, as a result of Vury being purchased by Taleo, MSL has been using a similar application by Taleo as its talent recruitment software. The application, which is accessed through MSL's public website, allows users to search for open positions as well as input information about themselves. To the extent Plaintiffs contend they were denied any specific positions, they will identify same and the Parties will meet and confer to discuss what, if any, information exists within Vurv/Taleo regarding the identified position. If information exists in Vurv/Taleo or another source regarding these positions, MSL will produce this information, to the extent such information is discoverable.

m. ServiceNow -- MSL utilizes ServiceNow as its Help Desk application. This system covers a wide variety of requests by employees for computer-related assistance (e.g., troubleshoot incidents, install software, etc.).

n. PeopleSoft -- MSL utilizes PeopleSoft, an Oracle-based software product, to record employee data such as date of hire, date of termination, promotions, salary increases, transfers, **[*50]** etc. MSL has produced data from this source and will consider producing additional data in response to a specific inquiry from Plaintiffs.

o. PRISM -- MSL utilizes PRISM for tracking time and billing. It is used primarily to track an employee's billable time. MSL will consider producing additional data in response to a specific inquiry from Plaintiffs.

p. Portal -- MSL maintains a portal provided through Oracle/BEA Systems. The portal is web-based and is used for light workflow activities (such as reviewing draft documents).

q. Desktops/Laptops -- MSL provided employees with desktop and/or laptop computers to assist in work related activities. MSL will collect the desktop/laptop hard drive data for 2 custodians and analyze the data to determine the level of duplication of documents in this data source against the data contained in the EMC SourceOne archive for the same custodians. (The parties will meet and confer regarding the selection of the two custodians.) The results of the analysis will be provided to Plaintiffs so that a determination can be made by the parties as to whether MSL will include this data source in its production of ESI to Plaintiffs. If so, the Parties will attempt **[*51]** to reach an agreement as to the approach used to collect, review and produce responsive and non-privileged documents.

r. Publicis Benefits Connection -- Plaintiffs understand that MSL provides employees with access to a centralized web based site that provides access to corporate benefits information and other related content.

s. GEARS -- MSL maintains a centralized web-based expense tracking and reporting system called "GEARS" where users are able to enter expenses and generate reports.

t. MS&L City -- MSL maintained a corporate web-based Intranet prior to migrating to Noovoo.

u. Adium -- This is a free and open source instant messaging client for Mac OS X users.

v. Pidgin -- Pidgin is a chat program which lets users log into accounts on multiple chat networks simultaneously. However, the data resides with a third party messaging provider (e.g. AIM, Yahoo!, Google Talk, MSN Messenger, etc.).

w. IBM Lotus Traveler and MobileIron -- MSL maintains these systems for e-mail device sync and security features for employees' mobile devices, including Blackberry devices, iPhones, iPads, Android phones, and Android tablets.

x. Mobile Communication Devices - MSL provides mobile devices and/or connectivity **[*52]** including Blackberry devices, iPhones, iPads, Android phones, and Android tablets to designated employees.

y. Yammer -- This is an instant messaging application hosted externally, used for approximately one year in or around 2008 through 2009.

z. SalesForce.com -- This is a web-based customer relationship management application but it was not widely used.

aa. Removable Storage Devices -- MSL does not restrict authorized employees from using removable storage devices.

D. Custodians

1. The Parties agree that MSL will search the e-mail accounts of the following individuals as they exist on MSL's EMC SourceOne archive. (Except where a date range is noted, the custodian's entire e-mail account was collected from the archive.)

	Custodian Name	Title
1.	Lund, Wendy	Executive VP of Global Client and Business Development
2.	Fite, Vicki	Managing Director, MSL Los Angeles
3.	Wilson, Renee	President, NE Region, Managing Director NY
4.	Brennan, Nancy (1/1/08 to 5/31/08)	SVP/Director Corporate Branding
5.	Lilien (Lillien, Kashanian), Tara	SVP, North America Human Resources
6.	Miller, Peter	Executive Vice President, CFO
7.	Masini, Rita	Chief Talent Officer
8.	Tsokanos, Jim	President of the Americas
9.	Da Silva Moore, Monique	Director Healthcare Practice, Global
10.	O'Kane, Jeanine (2/8/10 to 2/24/11)	Director of Healthcare North America
11.	Perlman, Carol	Senior VP
12.	Mayers, Laurie	SVP MS&L Digital
13.	Wilkinson, Kate	Account Executive
14.	Curran, Joel (5/1/08 to 5/31/10)	Managing Director MSL Chicago
15.	Shapiro, Maury	North American CFO
16.	Baskin, Rob (1/1/08 to 12/31/08)	Managing Director
17.	Pierce, Heather	VP
18.	Branam, Jud (1/1/08 to 1/31/10)	Managing Director, MS&L Digital
19.	McDonough, Jenni (1/1/08 to 12/31/08)	VP, Director of Human Resources
20.	Hannaford, Donald (1/1/08 to 3/1/08)	Managing Director
21.	On, Bill (1/1/08 to 2/24/11)	Managing Director
22.	Dhillon, Neil (9/8/08 to 5/31/10)	Managing Director MSL Washington DC
23.	Hubbard, Zaneta	Account Supervisor
24.	Morgan, Valerie (1/1/08 to 2/24/11)	HR Director
25.	Daversa, Kristin (1/1/08 to 2/24/11)	HR Director
26.	Vosk, Lindsey (1/1/08 to 2/24/11)	HR Manager
27.	Carberry, Joe (1/1/08 to 2/24/11)	President, Western Region
28.	Sheffield, Julie (1/1/08 to 2/24/11)	HR/Recruiting Associate
29.	MaryEllen O'Donohue	SVP (2010)
30.	Hass, Mark	CEO (former)
31.	Morsman, Michael	Managing Director, Ann Arbor (former)

E. **[*53]** Search Methodology[1]

1. General. The Parties have discussed the methodologies or protocols for the search and review of ESI collected from the EMC SourceOne archive and the following is a summary of the Parties' agreement on the use of Predictive Coding. This section relates solely to the EMC SourceOne data source (hereinafter referred to as the "e-mail collection").

2. General Overview of Predictive Coding Process. MSL will utilize the Axcelerate software by Recommind to search and review the e-mail collection for production in this case.

The process begins with Jackson Lewis attorneys developing an understanding of the entire e-mail collection while identifying a small number of documents, the initial seed set, that is representative of the categories to be reviewed and coded (re-

[1] As noted in Paragraphs A(1) and J of this Protocol, Plaintiffs object to the predictive coding methodology proposed by MSL.

levance, privilege, issue-relation). It is the step when the first seed sets are generated which is done by use of search and analytical tools, including keyword, Boolean and concept search, concept grouping, and, as needed, up to 40 other automatically populated filters available within **[*54]** the Axcelerate system. This assists in the attorneys' identification of probative documents for each category to be reviewed and coded.

Plaintiffs' counsel will be provided with preliminary results of MSL's hit counts using keyword searches to create a high priority relevant seed set, and will be invited to contribute their own proposed keywords. Thereafter, Plaintiffs' counsel will be provided with the non-privileged keyword hits -- both from MSL's keyword list and Plaintiffs' keyword list -- which were reviewed and coded by MSL. Plaintiffs' counsel will review the documents produced and promptly provide defense counsel with their own evaluation of the initial coding applied to the documents, including identification of any documents it believes were incorrectly coded. To the extent the parties disagree regarding the coding of a particular document, they will meet and confer in an effort to resolve the dispute prior to contacting the Court for resolution. The irrelevant documents so produced shall be promptly returned after review and analysis by Plaintiffs' counsel and/or resolution of any disputes by the Court.

The seed sets are then used to begin the Predictive Coding process. Each **[*55]** seed set of documents is applied to its relevant category and starts the software "training" process. The software uses each seed set to identify and prioritize all substantively similar documents over the complete corpus of the e-mail collection. The attorneys then review and code a judgmental sample of at least 500 of the "computer suggested" documents to ensure their proper categorization and to further calibrate the system by recoding documents into their proper categories. Axcelerate learns from the new corrected coding and the Predictive Coding process is repeated.

Attorneys representing MSL will have access to the entire e-mail collection to be searched and will lead the computer training, but they will obtain input from Plaintiffs' counsel during the iterative seed selection and quality control processes and will share the information used to craft the search protocol as further described herein. All non-privileged documents reviewed by MSL during each round of the iterative process (i.e., both documents coded as relevant and irrelevant) will be produced to Plaintiffs' counsel during the iterative seed set selection process. Plaintiffs' counsel will review the documents produced **[*56]** and promptly provide defense counsel with its own evaluation of the initial coding applied to the documents, including identification of any documents it believes were incorrectly coded. To the extent the Parties disagree regarding the coding of a particular document, they will meet and confer in an effort to resolve the dispute prior to contacting the Court for resolution. Again, the irrelevant documents so produced shall be promptly returned after review and analysis by Plaintiffs' counsel and/or resolution of any disputes by the Court.

At the conclusion of the iterative review process, all document predicted by Axcelerate to be relevant will be manually reviewed for production. However, depending on the number of documents returned, the relevancy rating of those documents, and the costs incurred during the development of the seed set and iterative reviews, MSL reserves the right to seek appropriate relief from the Court prior to commencing the final manual review.

The accuracy of the search processes, both the systems' functions and the attorney judgments to train the computer, will be tested and quality controlled by both judgmental and statistical sampling. In statistical sampling, **[*57]** a small set of documents is randomly selected from the total corpus of the documents to be tested. The small set is then reviewed and an error rate calculated therefrom. The error rates can then be reliably projected on the total corpus, having a margin of error directly related to the sample size.

3. Issue Tags. The parties agree that, to the extent applicable, as part of the seed set training described above, as well as during the iterative review process, all documents categorized as relevant and not privileged, to the extent applicable, also shall be coded with one or more of the following agreed-upon issue tags:

a. Reorganization.

b. Promotion/Assignments.

c. Work/Life Balance.

d. Termination.

e. Compensation.

f. Maternity/Pregnancy.

g. Complaints/HR.

h. Publicis Groupe/Jurisdiction.

This issue coding will take place during the initial random sample, creation of the seed set and initial and iterative training (see paragraphs 4, 5 and 6 below). This input shall be provided to Plaintiffs' counsel along with the initial document productions. Plaintiffs' counsel shall promptly report any disagreements on classification, and the parties shall discuss these issues in good faith, so that the seed **[*58]** set training may be improved accordingly. This issue-tagging and disclosure shall take place during the described collaborative seed set training process. The disclosures here made by MSL on its issue coding are not required in the final production set.

4. Initial Random Sample. Using the Axcelerate software to generate a random sample of the entire corpus of documents uploaded to the Axcelerate search and review platform, MSL's attorneys will conduct a review of the random sample for relevance and to develop a baseline for calculating recall and precision. To the extent applicable, any relevant documents also will be coded with one or more of the issue tags referenced in paragraph E.3 above. The random sample consists of 2,399 documents, which represents a 95% confidence level with a confidence estimation of plus or minus 2%. The Parties agree to utilize the random sample generated prior to the finalization of this protocol. However, during Plaintiffs' counsel's review of the random sample, they may advise as to whether they believe any of the documents should be coded with one or more of the subsequently added issue codes (i.e., Complaints/HR and Publicis Groupe/Jurisdiction) and **[*59]** will, as discussed above, indicate any disagreement with MSL's classifications.

5. Seed Set.

a. Defendant MSL. To create the initial seed set of documents that will be used to "train" the Axcelerate software as described generally above, MSL primarily utilized keywords listed on Exhibits A and B to this protocol, but also utilized other judgmental analysis and search techniques designed to locate highly relevant documents, including the Boolean, concept search and other features of Axcelerate. Given the volume of hits for each keyword (Exhibit A), MSL reviewed a sampling of the hits and coded them for relevance as well as for the following eight preliminary issues: (i) Reorganization; (ii) Promotion; (iii) Work/Life Balance; (iv) Termination; (v) Compensation; and (vi) Maternity. Specifically, except for key words that were proper names, MSL performed several searches within each set of key word hits and reviewed a sample of the hits. The Axcelerate software ranked the hits in order of relevance based on the software's analytical capabilities and the documents were reviewed in decreasing order of relevance (i.e., each review of the sample of supplemental searches started with the highest **[*60]** ranked documents). Exhibit B identifies the supplemental searches conducted, the number of hits, the number of documents reviewed, the number of documents coded as potentially responsive and general comments regarding the results. In addition, to the extent applicable, documents coded as responsive also were coded with one or more issue tags. MSL will repeat the process outlined above and will include the newly defined issues and newly added custodians. MSL will provide Plaintiffs' counsel with all of the non-privileged documents and will provide, to the extent applicable, the issue tag(s) coded for each document, as described above. Plaintiffs' counsel shall promptly review and provide notice as to any documents with which they disagree where they do not understand the coding. If necessary, counsel will meet and confer to attempt to resolve any disagreements regarding the coding applied to the documents in this seed set.

b. Plaintiffs. To help create the initial seed set of documents that will be used to "train" the Axcelerate software, Plaintiffs provided a list of potential key words to MSL. MSL provided Plaintiffs with a hit list for their proposed key words. This process was repeated **[*61]** twice with the hit list for Plaintiffs' most recent set of keywords attached as Exhibit C. MSL will review 4,000 randomly sampled documents from Plaintiffs' supplemental list of key words to be coded for relevance and issue tags.

MSL will provide Plaintiffs' counsel with all non-privileged documents and will provide, to the extent applicable, the issue tag(s) coded for each document. Plaintiffs' counsel shall promptly review and provide notice as to any documents with which they disagree with or where they do not understand the coding. If necessary, the Parties' counsel will meet and confer to attempt to resolve any disagreements regarding the coding applied to the documents in this seed set.

c. Judgmental Sampling. In addition to the above, a number of targeted searches were conducted by MSL in an effort to locate documents responsive to several of Plaintiffs' specific discovery requests. Approximately 578 documents have already been coded as responsive and produced to Plaintiffs. In addition, several judgmental searches were conducted which resulted in approximately 300 documents initially being coded as responsive and several thousand additional documents coded as irrelevant. The **[*62]** documents coded as relevant and non-privileged also will be reviewed by Plaintiffs' counsel and, subject to their feedback, included in the seed set. An explanation shall be provided by MSL's attorneys for the basis of the bulk tagging of irrelevant documents (primarily electronic periodicals and newsletters that were excluded in the same manner as spam junk mail is excluded). The explanation shall include the types of documents bulk tagged as irrelevant as well as the process used to identify those types of documents and other similar documents that were bulk tagged as irrelevant.

6. Initial And Iterative Training. Following the creation of the first seed set, the Axcelerate software will review the entire data set to identify other potentially relevant documents. MSL will then review and tag a judgmental based sample, consisting of a minimum of 500 documents, including all documents ranked as *highly relevant* or *hot*, of the new "Computer Suggested" documents, which were suggested by the Axcelerate software. MSL's attorneys shall act in consultation with the Axcelerate software experts to make a reasonable, good faith effort to select documents in the judgmental sample that will serve **[*63]** to enhance and increase the accuracy of the predictive coding functions. The results of this first iteration, both the documents newly coded as relevant and not relevant for particular issue code or codes, will be provided to Plaintiffs' counsel for review and comment. (All documents produced by the parties herein to each other, including, without limitation, these small seed set development productions, shall be made under the Confidentiality Stipulation in this matter as well as any clawback agreement that shall be reduced to an order acceptable to the Court. Any documents marked as irrelevant shall be returned to counsel for MSL at the conclusion of the iterative training phase, unless the relevancy of any documents are disputed, in which case they may be submitted to the Court for review.)

Upon completion of the initial review, and any related meet and confer sessions and agreed upon coding corrections, the Axcelerate software will be run again over the entire data set for suggestions on other potentially relevant documents following the same procedures as the first iteration. The purpose of this second and any subsequent iterations of the Predictive Coding process will be to further **[*64]** refine and improve the accuracy of the predictions on relevance and various other codes. The results of the second iteration shall be reviewed and new coding shared with Plaintiffs' counsel as described for the first iteration. This process shall be repeated five more times, for a total of seven iterations, unless the change in the total number of relevant documents predicted by the system as a result of a new iteration, as compared to the last iteration, is less than five percent (5%), and no new documents are found that are predicted to be *hot* (aka *highly relevant*), at which point MSL shall have the discretion to stop the iterative process and begin the final review as next described. If more than 40,000 documents are returned in the final iteration, then MSL reserves the right to apply to the Court for relief and limitations in its review obligations hereunder. Plaintiffs reserve the right, at all times, to challenge the accuracy and reliability of the predictive coding process and the right to apply to the Court for a review of the process.

7. Final Search and Production. All of the documents predicted to be relevant in the final iteration described in paragraph six above will be **[*65]** reviewed by MSL, unless it applies to the court for relief hereunder. All documents found by MSL's review to be relevant and non-privileged documents will be promptly produced to Plaintiffs.

If more than 40,000 documents are included in the final iteration, then MSL reserves its right to seek payment from Plaintiffs for all reasonable costs and fees MSL incurred related to the attorney review and production of more 40,000 documents.

8. Quality Control by Random Sample of Irrelevant Documents. In addition, at the conclusion of this search protocol development process described above, and before the final search and production described in Paragraph 7 above, MSL will review a random sample of 2,399 documents contained in the remainder of the database that were excluded as irrelevant. The results of this review, both the documents coded as relevant and not relevant, but not privileged, will be provided to Plaintiffs' counsel for review. (Any documents initially coded as "not relevant" will be provided subject to the Confidentiality Stipulation and any clawback agreements entered in this matter will be returned to counsel for MSL within 60 days of their production.) The purpose for this **[*66]** review is to allow calculation of the approximate degree of recall and precision of the search and review process used. If Plaintiffs object to the proposed review based on the random sample quality control results, or any other valid objection, they shall provide MSL with written notice thereof within five days of the receipt of the random sample. The parties shall then meet and confer in good faith to resolve any difficulties, and failing that shall apply to the Court for relief. MSL shall not be required to proceed with the final search and review described in Paragraph 7 above unless and until objections raised by Plaintiffs have been adjudicated by the Court or resolved by written agreement of the Parties.

F. Costs

1. MSL proposes to limit the costs of its final review and production of responsive ESI from the MSL email collection to an additional $200,000, above and beyond the approximately $350,000 it has already paid or is anticipated to pay in e-discovery related activities as previously described and disclosed to Plaintiffs.

2. Plaintiffs agree to bear all of the costs associated with their compliance with the terms of this protocol and with the receipt and review of ESI produced **[*67]** hereunder including the costs associated with its ESI experts at DOAR Litigation Consulting who will be involved with Plaintiffs in all aspects of this ESI protocol. Plaintiffs propose that MSL bear all of the costs associated with its obligations under the terms of this protocol and do not agree to limit the amount of information subject to the review and production of ESI by MSL.

G. Format of Production For Documents Produced From Axcelerate

1. TIFF/Native File Format Production. Documents will be produced as single-page TIFF images with corresponding multi-page text and necessary load files. The load files will include an image load file as well as a metadata (.DAT) file with the metadata fields identified on Exhibit D. Defendant MSL will produce spreadsheets (.xls files) and PowerPoint presentations (.ppt files) in native form as well as any documents that cannot be converted to TIFF format (e.g., audio or video files, such as mp3s, ways, megs, etc.). In addition, for any redacted documents that are produced, the documents' metadata fields will be redacted where required. For the production of ESI from non-email sources, the parties will meet and confer to attempt to reach an agreement **[*68]** of the format of production.

2. Appearance. Subject to appropriate redaction, each document's electronic image will convey the same information and image as the original document. Documents that present imaging or formatting problems will be promptly identified and the parties will meet and confer in an attempt to resolve the problems.

3. Document Numbering. Each page of a produced document will have a legible, unique page identifier "Bates Number" electronically "burned" onto the image at a location that does not obliterate, conceal or interfere with any information from the source document. The Bates Number for each page of each document will be created so as to identify the producing party and the document number. In the case of materials redacted in accordance with applicable law or confidential materials contemplated in any Confidentiality Stipulation entered into by the parties, a designation may be "burned" onto the document's image at a location that does not obliterate or obscure any information from the source document.

4. Production Media. The producing party will produce documents on readily accessible, computer or electronic media as the parties may hereafter agree upon, **[*69]** including CD-ROM, DVD, external hard drive (with standard PC compatible interface), (the "Production Media"). Each piece of Production Media will be assigned a production number or other unique identifying label corresponding to the date of the production of documents on the Production Media (e.g., "Defendant MSL Production April 1, 2012") as well as the sequence of the material in that production (e.g. "-001", "-002"). For example, if the production comprises document images on three DVDs, the producing party may label each DVD in the following manner "Defendant MSL Production April 1, 2012", "Defendant MSL Production April 1, 2012-002", "Defendant MSL Production April 1, 2012-003." Additional information that will be identified on the physical Production Media includes: (1) text referencing that it was produced in *da Silva Moore v. Publicis Groupe SA, et al.*; and (2) the Bates Number range of the materials contained on the Production Media. Further, any replacement Production Media will cross-reference the original Production Media and clearly identify that it is a replacement and cross-reference the Bates Number range that is being replaced.

5. Write Protection and Preservation. **[*70]** All computer media that is capable of write-protection should be write-protected before production.

6. Inadvertent Disclosures. The terms of the Parties' Clawback Agreement and Court Order shall apply to this protocol.

7. Duplicate Production Not Required. A party producing data in electronic form need not produce the same document in paper format.

H. Timing.

I. To the extent a timeframe is not specifically outlined herein, the parties will use their reasonable efforts to produce ESI in a timely manner consistent with the Court's discovery schedule.

2. The parties will produce ESI on a rolling basis.

I. General Provisions.

1. Any practice or procedure set forth herein may be varied by agreement of the parties, and first will be confirmed in writing, where such variance is deemed appropriate to facilitate the timely and economical exchange of electronic data.

2. Should any party subsequently determine it cannot in good faith proceed as required by this protocol, the parties will meet and confer to resolve any dispute before seeking Court intervention.

3. The Parties agree that e-discovery will be conducted in phases and, at the conclusion of the search process described in Section E above, the **[*71]** Parties will meet and confer regarding whether further searches of additional custodians and/or the Phase II sources is warranted and/or reasonable. If agreement cannot be reached, either party may seek relief from the Court.

J. Plaintiffs' Objection

1. Plaintiffs object to this ESI Protocol in its entirety. Plaintiffs submitted their own proposed ES! Protocol to the Court, but it was largely rejected. The Court then ordered the parties to submit a joint ES! Protocol reflecting the Court's rulings. Accordingly, Plaintiffs jointly submit this ESI Protocol with MSL, but reserve the right to object to its use in this case.

This protocol may be executed in counterparts. Each counterpart, when so executed, will be deemed and original, and will constitute the same instrument.

By:

JANETTE WIPPER, ESQ.

DEEPIKA BAINS, ESQ.

SIHAM NURHUSSEIN, ESQ.

SANFORD WITTELS & HEISLER, LLP

Attorneys for Plaintiffs and Class

555 Montgomery Street, Ste. 1206

San Francisco, CA 94111

Telephone: (415) 391-6900

Date: , 2012

By:

BRETT M. ANDERS, ESQ.

VICTORIA WOODIN CHAVEY, ESQ.
JEFFREY W. BRECHER, ESQ.
JACKSON LEWIS LLP
Attorneys for Defendant MSLGROUP

58 South Service Road, Suite 410
Melville, NY 11747
Telephone: (631) **[*72]** 247-0404
Date: , 2012