The Sedona Conference Journal

Volume 6 2005

Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery

Jason R. Baron

Toward A Federal Benchmarking Standard For Evaluating Information Retrieval Products Used In E-Discovery

Jason R. Baron[1]
National Archives and Records Administration
College Park, MD

"The Half of Knowledge is Knowing Where to find it"
--- Samuel Johnson, 1775

I. Introduction

On September 21, 2004, lawyers began their opening arguments in *U.S. v. Philip Morris,* before presiding Judge Gladys Kessler in the U.S. District Court for the District of Columbia, in what turned out to be an eight month bench trial.[2] The government's RICO[3] case against the tobacco industry, in which at one time the U.S. claimed entitlement to some $280 billion in damages,[4] has to date generated over 40 million pages in discovery.[5] Untold many millions more pages undoubtedly have been subject to searches conducted by all parties in a quest for relevant records.

One "baby universe" of records searched for purposes of responding to discovery propounded by the defendant tobacco companies consisted of 18 million email records of the Clinton White House, maintained in an electronic repository at the headquarters of NARA in College Park, Maryland. For NARA, the manual review process employed in this review -- once an initial automated keyword search had been conducted -- entailed a huge commitment of time and staff resources, stretching NARA resources to the limit. The problems encountered in searching this large database revolved in substantial part around an automated keyword search that generated large numbers of nonresponsive records (along with responsive ones), all of which needed to be sorted out through a labor-intensive manual review process.

The failure to greatly weed out false positives at the automated state of the search necessarily should prompt re-examination of the basic assumptions regarding the utility of the information retrieval techniques that were employed. This is especially the case given the expected exponential growth of NARA's holdings of electronic records. Can and should the same information retrieval techniques be employed in the future against databases that are between ten and a thousand times larger than even the current White House email database of records? Are there alternative

[1] Director of Litigation, Office of the General Counsel, National Archives and Records Administration (NARA). A prior version of this paper was presented at the Sedona Conference's Third Annual Meeting of the Working Group on Best Practices for Electronic Retention and Production, held in Phoenix, Arizona, October 15-16, 2004. I am grateful to have received comments on the prior version and/or to have had the opportunity to discuss the ideas presented herein with a number of individuals, including Daisy Abbot, Paul Berkman, David Blair, Macyl Burke, Robert Chadduck, Bruce Croft, Lucja Iwanska, Reagan Moore, Doug Oard, Herb Roitblatt, Seamus Ross, Ian Soboroff, Bill Underwood, and Ellen Voorhees. The views expressed in this paper are, however, solely the author's, and do not necessarily purport to represent the official position of NARA or any other component of the Federal government. Elizabeth Kim, George Mason U. School of Law (J.D. candidate 2006), assisted in providing initial legal research.

[2] Michael Janofsky, "Last Defense Witness Testifies in Federal Tobacco Trial," *New York Times,* June 3, 2005.

[3] Racketeer Influenced and Corrupt Organizations Act, 18 U.S.C. 1961-68.

[4] See Memorandum Opinion in *U.S. v. Philip Morris USA, Inc., et al.,* Civ. Action No. 99-2946 (D.D.C.), dated July 7, 2004, at 2 n.3, available at http://www.dcd.uscourts.gov/99-2496ak.pdf. The U.S. Court of Appeals subsequently vacated the government's RICO "disgorgement" claim, see *U.S. v. Philip Morris USA, Inc.,* 396 F.3d 1190 (D.C. Cir. 2005); however, a further appeal and additional remedies have been sought by the government. See Carol D. Leonnig, "Tobacco Escapes Huge Penalty: U.S. Seeks $10 Billion Instead of $130 Billion", *The Washington Post*, June 8, 2005.

[5] "Drowning in Paper," *Legal Times,* March 18, 2004, available at http://www.law.com/jsp/article.jsp?id=1079144426836.

automated ways to search for information that make for a more efficient process without sacrificing either the completeness or the accuracy of the overall search? And are there objective measures for evaluating the efficacy of very different types of search methodologies?

The Sedona Conference's January 2004 Best Practices Paper includes a brief comment regarding "search methodologies," including the notion of employing keyword and concept searching, as well as utilizing sampling techniques to refine the accuracy of searches while reducing cost.[6] A variety of alternative methodologies to simple keyword searching, including advanced Boolean searching, concept searching, and natural language searches, do in fact exist, as discussed below. The Sedona Conference's very raising of the notion of multiple search methodologies being employed is itself forward looking, given a paucity of case law in this area. Notwithstanding, however, the substantial buzz in the vendor community regarding the latest "search engine"-type products available to consumers, it seems clear that there is no agreed-upon standard for measuring different information retrieval methodologies (let alone particular software products) in the context of how they actually do or would perform in civil discovery.

Based on a substantial body of academic work in the area of information and text retrieval, a "benchmarking" standard for comparing the results of differing search methodologies is conceivable and could be developed. As proposed here, use of particular methodologies and particular "certified" products that have been found to meet a set of baseline or functional requirements for the accurate retrieval of responsive records could be stipulated to at the onset of litigation, either in the context of Rule 26(a) initial discovery, or at any time thereafter as part of the overall discovery process.

This paper represents an initial exercise in fleshing out what a benchmarking standard would entail, drawing upon the concepts of precision and recall as developed in the academic literature of information retrieval. Reference will also be made to the *Daubert* formulation of "known or potential rate[s] of error" to be associated with competing search methodologies, including notions of Type I and Type II errors (*i.e.,* false positives and false negatives). The modest proposal made below is a call for the development by an accredited national or international standards body of a testbed for the testing of software products meeting certain benchmark standards, along the lines of the Department of Defense's efforts in evaluating whether software products meet the DoD 5015.2 standard for electronic recordkeeping products.

II. The Information Retrieval Problem in Litigation: An Example

Over the course of a two year period, NARA responded to discovery propounded by the defendant tobacco companies in *U.S. v Philip Morris*.[7] As part of this effort, NARA staff in College Park, Maryland conducted keyword searches against Clinton email repositories. The main database consisted of email records captured by ARMS, the Automated Records Management System operated by IT staff in the Office of Administration (OA) component of the Executive Office of the President (EOP) on behalf of most White House staff. ARMS was created in July 1994 in the aftermath of the long-running White House email case captioned *Armstrong v. Executive Office of the President*.[8] The ARMS system operated as a form of electronic recordkeeping application, sitting "on top of" the DEC All-in-1 email system (and later Lotus Notes system), and functioning to capture email designated as "records" by approximately 1500 EOP end users under the Federal Records Act and the

[6] See THE SEDONA PRINCIPLES: Best Practices Recommendations & Principles for Addressing Electronic Document Production (January 2004 Version), Comment 11a, p. 39 (SEDONA PRINCIPLES), available at http://www.thesedonaconference.org/content/miscFiles/publications_html?grp=wgs110.

[7] The last paragraph of a 1,723 paragraph "Request to Produce Documents," filed by defendants under Rule 34 of the Federal Rules of Civil Procedure (FRCP), and directed to U.S. government agencies generally, requested that NARA search all of its facilities (including Presidential libraries, as well as its regional and national headquarters archives), for responsive records to the prior 1,722 requests to produce all documents concerning tobacco.

[8] 1 F.3d 1274 (D.C. Cir. 1993). *Armstrong*, otherwise known as the "PROFS" case, involved a series of rulings regarding the White House's duty to manage and preserve e-mail records under the Federal Records Act, including Iran-Contra related email residing only on National Security Council backup tapes. For a discussion of the *Armstrong* litigation, see Jason R. Baron, "E-mail Metadata in a Post-*Armstrong* World," paper in the *Proceedings of the 3rd IEEE Computer Society Metadata Conference,* April 1999, available at http://www.computer.org/proceedings/meta/1999/papers/83/jbaron.html; see also Jason R. Baron, "The PROFS Decade: NARA, E-mail, and the Courts," Chapter 6 in Bruce Ambacher, ed., *Thirty Years of Electronic Records* (Lanham, MD: Scarecrow Press 2003).

Presidential Records Act.[9] For purposes of tobacco discovery, EOP federal records components undertook their own keyword searches and worked separately with DOJ in producing responsive federal record emails; NARA, as legal custodian of all Clinton presidential records after the Administration ended, had lead responsibility for producing presidential record emails. Although the internal structure of ARMS involved a series of "buckets" standing as proxies for various EOP components, thus necessitating multiple automated searches, for purposes here the presidential components of ARMS can collectively be treated as if they consisted of one undifferentiated "big bucket" of email records capable of being searched on an across-the-board basis in response to the inputting of keywords.[10]

For the tobacco search, the straightforward approach NARA employed with respect to keyword searching produced on the order of 200,000 "hits" against the 18 million presidential record database. These hits were generated as the result of basically two searches: a first search was conducted based on a dozen simple keywords entered, such as "tobacco," "smoking," "cigarette," "tar" and "nicotine," with a later, second search, conducted in slightly more sophisticated fashion using Boolean operators such as "AND NOT," to eliminate potential sources of false positive hits. Through repeated sampling and a fair measure of trial and error, NARA staff were able to thus refine automated search queries. For example, the string "Marlboro AND NOT Upper Marlboro" was utilized to eliminate emails concerning Upper Marlboro, Maryland; similarly, the resulting "PMI [Philip Morris Institute] AND NOT presidential management intern"; and "b/w [Brown & Williamson] AND NOT photo." In some cases, NARA was able to persuade the parties not to demand search terms that were known to produce unduly large numbers of false positives.[11]

The results of NARA's automated searches of the ARMS database (*i.e.,* all "hits" based on keyword searches) were burned onto CDs, which in turn were handed over in controlled fashion to a small army of in-house archivists, lawyers, and law clerks for the purpose of their conducting further manual review over what amounted to over a six month process. This manual review effort consisted of reviewing both emails and all attachments to emails (necessitating often a review of multiple attachments), first for printing out potentially responsive records, second, for a re-review for responsiveness, and third, for multiple reviews to determine whether one or more grounds for asserting privilege might be claimed. Except for highly sensitive privileged email records which were withheld from production but otherwise recorded on privilege logs, all responsive emails - privileged and nonprivileged alike - were then made available in hard copy form to defendants as part of an agreed-upon "open record" or "clawback" procedure.[12] Substantially over 100,000 responsive email records were made available to defendants in this fashion.

In speculating about the next wave of potential litigation, NARA simply is unlikely to be able to employ a similarly manual review-intensive process for presidential emails (and other forms of e-records) where the potential universe of hits numbers in the millions or tens of millions of records (based not only on a Clinton era database, but also future databases containing even substantially larger numbers of emails from the incumbent and future presidential Administrations). Although a manual review for privileged material may or may not continue to be viewed as a necessity by future Presidents - a separate matter beyond the scope of this paper - there is without question a need for a

[9] Uniquely in government, OA's email system operates under two distinct records laws. Because some EOP components legally constitute federal agencies (*e.g.,* OMB, the Office of Science and Technology Policy, etc.) with independent governmental functions, they are covered under the Federal Records Act, 44 U.S.C. Chapters 21, 29, 31 & 33. Yet other White House components consist of personnel whose sole duty it is to advise and assist the President (*e.g.,* the White House Office, the Domestic Policy Council, the Council of Economic Advisers, the National Security Council (the latter being the result of a later holding in *Armstrong*, see 90 F.3d 553 (D.C. Cir. 1996), etc.), and are thus covered by the Presidential Records Act, 44 U.S.C. Chapter 22. One chief distinction between the records laws is that while federal records are immediately subject to the Freedom of Information Act, presidential records are generally exempt from FOIA access until five years beyond the end of a President's time in office, or in the case of certain categories of records designated by the President, up to 12 years beyond the end. See 44 U.S.C. 2204.

[10] Also left unaddressed here is a further complication in searching posed by the fact that as the result of certain types of missing email in ARMS, EOP and NARA staff had entered into a Tape Restoration Project (TRP) aimed at restoring email from backup tapes and creating a separate TRP database for eventual merger with ARMS. The TRP database was also searched for the tobacco discovery. See generally Memorandum of Understanding entered into between EOP and NARA, dated January 11, 2001, available at http://www.archives.gov/presidential_libraries/presidential_records/clinton_gore_email_records_memo.html (describing restoration efforts undertaken).

[11] One notable example: although the search term "TI" was put forward as potentially of interest in generating potential hits for emails concerning the "Tobacco Institute," NARA sampling revealed a lack of responsive emails but a large number of emails referencing Julie Andrews in *The Sound of Music* singing "Do, Re, Me . . ." (leading to "Ti [tea] . . . that will bring us back to Do").

[12] See SEDONA PRINCIPLES at 37 (Comment 10.d), for use of this form of open discovery. A different, procedurally simpler procedure was utilized by EOP staff in responding to discovery involving federal record emails, consisting of wholesale turning over of emails in electronic form on CDs under open records discovery subject to clawback. The sensitive nature of presidential record emails precluded use of a similar procedure, thus leading to the more manual review intensive process described above.

more efficient automated search mechanism to weed out false positive "hits" so as to minimize the tremendous inefficiencies in reviewing emails and attachments by means of a manual process for the purpose of establishing a sub-universe of responsive or relevant records in discovery.

III. ALTERNATIVE SEARCH METHODOLOGIES: SOME CHOICES

In the academic world, "information retrieval" is defined as "[the] actions, methods and procedures for recovering stored data to provide information on a given subject."[13] With the substantial caveat that the following does not purport to represent a comprehensive taxonomy, there would appear to be several types of non-mutually exclusive search methodologies worth considering as "alternatives" to brute force information retrieval by simple keyword inputs.

A. Advanced Boolean Searches.

Boolean logic has been described as a "'syntactical calculus,' a mathematical algebra that allows the researcher to tailor the search query by using the Boolean operators 'and,' 'or,' 'not,' or 'near.'"[14] Search queries may be tailored by linking terms using grammatical connectors ("term a" within same paragraph as "term b"), or numerical connectors ("term a" within so many words of "term b").[15] As made reference to above, one systematic means of making keyword searches more efficient is to employ an iterative process: if "term a" generates large numbers of the same sort of false positive, refine the search through a "term a AND NOT term b" command (Marlboro AND NOT "Upper Marlboro," where "Marlboro" cigarettes is the target responsive concept). In doing so, an evaluation may be warranted, however, as to whether documents including "term b," while in the main turning out to be false positives, nevertheless retain the capability to still end up being responsive to the original target "term a."

B. Statistical Techniques.

"Statistical document retrieval techniques extend the capabilities of Boolean systems by using frequency of occurrence and relevance ranking."[16] A relevance score for particular documents is obtained using an inverse frequency algorithm, which "assumes that for a document to score high, the query term must occur frequently in the document, but infrequently in the entire document set."[17] This technique "weights" the use of particular terms in a Boolean expression according to their perceived relevance ranking.

C. Concept Searching.

Concept searching "involves using a massive dictionary of the English language to find synonyms and related words."[18] The example is given for the word "motorcycle" retrieving "moped," "bike" or perhaps even "automobile," "car" or "vehicle."[19]

D. Natural Language.

Natural language queries are expressed using normal conversational syntax, either as spoken or in writing, without syntactical rules or conventions. The legal community is familiar with natural language queries utilizing Lexis and Westlaw.

[13] ISO 2382/1 (1984).

[14] Carol M. Bast, Ransford C. Pyle, "Legal Research in the Computer Age: A Paradigm Shift?," 93 *Law Library Journal* 285 (Spring 2001), text at n.39 (available on Westlaw).

[15] *Id.*

[16] William E. Underwood, Matthew G. Underwood, "Evaluation of Document Retrieval Technologies to Support Access to Presidential Electronic Records," PERPOS Technical Report ITTL/CISTD 02-3 (December 2002), George Tech Research Institute (available from the author).

[17] *Id.* at 3 & n.3.

[18] Margaret Carol Fine, "Zen and the Art of Internet Searching," *Law Technology Product News,* Vol. 7, No. 12, at 147 (Col. 1) (December 2000) (available on Westlaw).

[19] *Id.*

E. Fuzzy Logic.

"Fuzzy logic was first invented as a representation scheme and calculus for uncertain or vague notions. It is basically a multi-valued logic that allows more human-like interpretation and reasoning in machines by resolving intermediate categories between notations such as true/false, hot/cold etc used in Boolean logic."[20] Apart from its other intricacies, using fuzzy logic one might hope to attain "hits" based on misspellings of versions of words or incomplete words, such as "motercyler" for "motorcycler."[21] This technique takes on importance due to the error rate associated with scanned in data utilizing optical character recognition.

Many other variations on the above techniques exist under different labels, including references on the Web and in the information retrieval literature to case based reasoning, query expansion, probabilistic logic, Bayesian networks, vector spaces, parallel computing, and visual analytics, to name just a few.[22] Moreover, there is nothing to prevent combining one or more of the above techniques in hybrid fashion to optimize search results.

IV. MEASURES FOR EVALUATING SEARCH METHODOLOGIES IN E-DISCOVERY

An enormous academic literature already exists on the general subject of information retrieval (IR) and "searching."[23] Beginning with the so-called "Cranfield tests" conducted in the late 1950s and early 1960s involving indexing techniques, a huge cottage industry of IR academic research has devoted itself to following a traditional model of: (i) using test collections of documents; (ii) developing query or topic sets as tasks to perform; and (iii) making relevance judgments.[24] A number of computer science-oriented organizations and conferences are devoted in whole or in part to the subject, including with respect to comparing or evaluating different alternative search methodologies.[25]

However, so far as this author is aware, there has been little reference in the academic literature to the special type of data mining problems faced by lawyers as they confront demands from opposing parties in civil discovery to find all responsive documents in large electronic databases under arbitrary, externally-imposed deadlines.[26] It must be conceded that "the legal system is remarkable in its reliance on both precise and imprecise concepts."[27] Nevertheless, the prime measures of accuracy analyzed in the information retrieval literature arguably map to the litigation support context.

In response to a request to produce documents filed under FRCP 34, the term "documents" will be broadly defined by careful lawyers to include just about any electronic "data file" extant on an opposing party's computers. Such data files (referred to for purposes of this article as "documents")

[20] Pragya Agarwal, "Lotfi Zadeh: Fuzzy logic-Incorporating Real-World Vagueness," Center for Spacially Integrated Social Science, Univ. of California, available at http://www.csiss.org/classics/content/68.

[21] Fine, "Zen and the Art of Internet Searching," *supra* n.18.

[22] See generally, Mark Lager, "Spinning a Web Search," Univ. of California, Santa Barbara (1996), available at http://www.library.ucsb.edu/untangle/lager.html; Michael W. Berry et al., "Matricies, Vector Spaces, and Information Retrieval," 41 *SIAM Review* (1999), available at http://epubs.siam.org/sam-bin/getfile/SIREV/articles/34703.pdf. For interesting recent work in applying visualization techniques to email archives, with implications for improving future content analysis, see Adam Perer, Ben Schneiderman, Douglas W. Oard, *Using Rhythms of Relationships to Understand Email Archives* (2005), available at http://www.cs.umd.edu/hcil/emailviz/workshop.

[23] A Google search for "information retrieval" returned 4,210,000 entries (in 0.30 seconds) (accessed June 5, 2005). For early scholarship, see G. Salton, *Automatic Information Organization and Retrieval* (New York: McGraw-Hill 1968); C.J. van Rijsbergen, *Information Retrieval* (London: Butterworth 1979). More recent scholarly works include Ricardo Baeza-Yates, Berthier Ribeiro-Neto, *Modern Information Retrieval* (N.Y.: Addison-Wesley 1999); David A. Grossman, Ophir Frieder, *Information Retrieval: Algorithms and Heuristics* (Boston: Kluwer Academic Publishers 1998); Karen Sparck Jones, Peter Willett, eds., *Readings in Information Retrieval,* chap. 4 ("Evaluation") (San Francisco: Morgan Kauffman Pubs. 1997); Ian H. Whitten, Alistair Moffat, Timothy C. Bell, *Managing Gigabytes: Compressing and Indexing Documents and Images,* chap. 4 ("Querying") (New York: Van Nostrand Reinhold 1994). Websites providing further multiple links include: http://web.syr.edu/~diekemar/ir.html; http://www.searchtools.com/info; http://filebox.vt.edu/users/wfan/text_mining.html.

[24] Edie Rasmussen, "Evaluation in Information Retrieval" (2002), available at www.music-ir.org/evaluation/wp2/wp2_rasmussen.pdf.

[25] See, *e.g.,* Annual Text REtrieval Conferences (TREC), http://trec.nist.gov; ACM (Association for Computer Machinery) Special Interest Group on Information Retrieval, http://www.sigir.org; see also U. Mass. Amherst Center for Intelligent Information Retrieval, at http://ciir.cs.umass.edu.

[26] David C. Blair recognizes the issue in his monograph, *Language and Representation in Information Retrieval* (Amsterdam: Elsevier 1990), at ii-iii:

> There is a growing undercurrent of urgency in the study of Information Retrieval, because the problems with which it is concerned are pervasive and spreading . . . . Adding to the urgency is the fact that extraordinarily high standards of retrieval are required with increasing frequency. Large document data bases to support corporate or government litigation . . . are . . . examples of the growing number of large-scale systems where the consequences of poor retrieval can be dramatic.

See also Robin Widdison, "New Perspectives in Legal Information Retrieval," 10 *International Journal of Law and Information Technology* 41, 66 (2002) (available on Westlaw).

[27] Daniel E. Rose, Richard K. Belew, "Legal Information Retrieval: A Hybrid Approach," in *Proceedings of the Second International Conference on AI and Law,* Vancouver, 138-146 (1989).

are either responsive to discovery, or they are not. They also will be retrieved in discovery, or they will not. The resulting 2 x 2 set of possible permutations is simple to contemplate:

	RESPONSIVE	**NON-RESPONSIVE**
Retrieved	**A**	**B**
Not retrieved	**C**	**D**

Cell A: Responsive documents to a discovery request successfully retrieved.
Cell B: Nonresponsive documents retrieved in error (*i.e.*, false positives).
Cell C: Responsive documents failed to be retrieved (*i.e.*, false negatives).
Cell D: Nonresponsive documents properly left unretrieved.[28]

The "Recall" rate is a measure of the ability of a given retrieval methodology to find all of the potential responsive documents in a given collection. The recall fraction is expressed as:

$$\text{RECALL} = \frac{\text{number of responsive documents actually retrieved}}{\text{number of potentially responsive documents}} = \frac{A}{A + C}$$

"Precision" is a measure of the ability of a given retrieval methodology to find truly responsive documents amongst all the documents in a given collection.

$$\text{PRECISION} = \frac{\text{number of responsive documents retrieved}}{\text{total number of documents retrieved}} = \frac{A}{A + B}$$

An example of the difference between the two measures: Assume a database of 1000 documents, of which 100 truly are responsive to a discovery request. The search methodology used retrieves 200 documents, but of these 200 "hits" only 50 turn out to be responsive.

Recall rate = 50/100 = 50% Precision = 50/200 = 25%

Ideally, a good retrieval methodology will serve to recall a very high percentage of truly responsive documents, without sacrificing "production efficiency" in also generating large numbers of "hits" containing nonresponsive documents. The relationship between recall and precision is an inverse one: increasing the Recall rate invariably leads to a corresponding loss of Precision, as more and more documents are retrieved to find the elusive remaining needle in the rest of the haystack. This idea is captured in the generic graph below[29]:

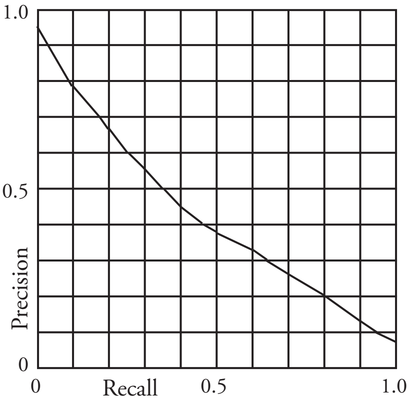

Fig. 1. A typical precision-recall graph

[28] Display format and legend modified from "Quantitative Issues in Information Retrieval," available at http://www.diagnosticstrategies.com/info_retrieval.htm.
[29] Diagram borrowed from Kavi Mahesh, Oracle Corporation, "Text Retrieval Quality: A Primer," available at http://www.oracle.com/technology/products/text/htdocs/imt_quality.htm.

The academic literature analyzes and evaluates precision levels of particular retrieval techniques when measured against a particular arbitrarily selected rate of recall.[30] There is, however, no "one" recall rate (other than the utopian 100%) that corresponds to a generally accepted benchmark level for use by the legal profession in responding to discovery. (Parties are expected to conduct "reasonable" searches of their holdings, not "perfect" ones.) One widely-cited early study by Blair and Manon suggested that lawyers overestimate the number of responsive documents uncovered by the particular search methodology they choose to use.[31] A more recent "pilot test," involving a comparison of how human review fared versus automated search techniques of a sample from a proprietary test collection of 48,000 documents, suggests that lawyers using only themselves as reviewers failed to do as well as automated techniques in finding relevant documents.[32]

Based on unscientific surveys and anecdotal evidence, it is generally perceived to be the case that neither Boolean nor natural language searches provide dramatically different rates of "recall."[33] It is also fairly widely held in the librarian and legal communities that natural language searching provides less overall precision in results, in the manner defined above.[34] On the other hand, one important study conducted by Howard Turtle comparing retrieval effectiveness indicated that natural language querying provides better retrieval performance (in terms of recall) than expert searchers using a Boolean retrieval system, when searching full text legal materials in the context of a controlled WESTLAW environment.[35] All of these propositions remain, however, largely untested with respect to the latest generation of search methodologies as applied to a potential universe of corporate or institutional data subject to civil discovery.

A different way of analyzing the "production inefficiency" built into any process of responding to civil discovery would be to consider the "rates of error" associated with particular methodologies, in a manner suggested by the Supreme Court's 1993 opinion in *Daubert v. Merrell Dow Pharmaceuticals, Inc.*[36] There, the Court set out a multi-factor test for introducing scientific evidence in the courtroom, which included "in the case of a particular scientific technique" the fact that a "court ordinarily should consider the known or potential rate of error" associated with the technique, and "the existence and maintenance of standards controlling the technique's operation."[37] The two well-recognized types of errors associated with *Daubert* are Type I errors in the nature of false positives, and Type II errors in the nature of false negatives (*i.e.*, the latter being the "missed document rate"[38]).[39] The retrieval of large numbers of false positive unresponsive documents is certainly burdensome and vexatious; however, the failure to find responsive documents can be critical in particular litigations (*i.e.*, the failure to find one or more "smoking guns"). This asymmetry as between the importance of Type I and Type II errors takes on special significance in matters of legal discovery, as compared with other contexts.[40] In other words, lawyers' "risk tolerance" for missing responsive records is low.

At least as a starting point for further discussion, Type II errors appear important enough to specifically control for. Accordingly, in considering alternative methodologies to brute force keyword

[30] It should also be noted that the academic literature cited *supra*, n.23, contains a number of other statistical measures, including combining Recall and Precision rates into a single number. One such combined measure is the harmonic mean of R and P, or M, where $M=2RP/(R+P)=2A/(2A+B+C)$. See http://www.diagnosticstrategies.com/info_retrieval.htm. A critique of search methodologies in a legal context using the harmonic mean of R and P is beyond the scope of this paper.

[31] The original study, reported in 1985, involved approximately 350,000 pages (in 40,000 documents) gathered in connection with litigation involving a San Francisco BART train accident. Although the attorneys in the litigation initially estimated that they had found 75% of the relevant documents, Blair and Manon estimated a true recall rate of only 20%. See David C. Blair and M. E. Manon, "An evaluation of retrieval effectiveness for a full-text document retrieval system," *Communications of the ACM,* 28(3): 290-299 (1985).

[32] See Anne Kershaw, Sherry Harris, et al., "Technology Trends," paper presented at The Sedona Conference® Third Annual Meeting of the Working Group on Best Practices for Electronic Document Retention and Production, Oct. 2004 (available from the author).

[33] See, *e.g.,* Sheilla E. Désert, "Westlaw is Natural v. Boolean Searching: A Performance Study," 85 *Law Library Journal,* 713, 715-16 (1993); Christopher G. Wren, Jill Robinson Wren, *Using Computers in Legal Research: A Guide to Lexis and Westlaw* 26 (Madison: Adams & Ambrose, 1994).

[34] See Underwood, *supra* n.16 ("A difficulty with natural language-based document retrieval systems based on query expansion is that while they increase the recall of relevant documents, they do this at the expense of retrieval of many more irrelevant documents than would be retrieved with statistical document retrieval systems.")

[35] Howard Turtle, "Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance," *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval:* 212-220 (1994).

[36] 509 U.S. 579.

[37] *Id.* at 594.

[38] "Missed Documents" constitute the number of responsive documents missed by the search methodology; in the example above, $M=C/(A+C)$ (the inverse of the Recall rate).

[39] See http://www.daubertexpert.com/basics_daubert-v-merrell-dow.html.

[40] Cf. http://www.diagnosticstrategies.com/info_retrieval.htm (suggesting that problems of false positives and false negatives warrant more equal weight in the medical profession).

searching, a fundamental assumption should be that any proposed alternative must be able to meet or surpass the Recall rate associated with a particular keyword methodology for a given level of Precision. Conversely, establishing a hierarchy of methodologies with the "winner" being the method which generates the greatest rate of Precision for a given level of Recall provides an optimal result. In doing so, it is not to be underestimated that "the comparison of retrieval performance between ranked and unranked systems raises a number of difficult problems"[41] which will need to be addressed.

V. THE IDEA OF ESTABLISHING A FEDERAL INFORMATION RETRIEVAL BENCHMARK

As stated succinctly by Jennifer Trant representing Archives & Museum Informatics:

> Benchmarking is well established throughout the economy, in areas from automotive manufacturing to knowledge management to higher education. All benchmarking initiatives are committed to sharing information and to improving business processes or performance. Through shared measures, assessments can be conducted that provide comparable results in different contexts. The emphasis in benchmarking is on reliability and comparability.
>
> * * * *
>
> Benchmarking involves the comparison of the results of two or more different methods of performing a known task with a known result (the benchmark) in order to establish relative effectiveness. Benchmarking is a critical component of establishing best practices, key performance indicators, or performance metrics . . . .[42]

For our purposes, some of the questions raised with respect to benchmarking search methodologies would be:

- Can we validly compare keyword searches against alternative forms of search methodologies?
- What objective statistical measures of accuracy can be employed in making such comparisons?
- What levels of "Recall" and "Precision" associated with current search methodologies exist for particular types of legal searches, as applied to particular forms and content types of documents?
- What costs are associated with increasing either rates of Recall or Precision?
- How are new methodologies to be evaluated?[43]

Were there an agreed upon set of standards for using particular software products employing one or more search methodologies, opposing parties in litigation would theoretically be more likely to reach agreement in the form of stipulations at the onset of discovery concerning the use of those products for conducting wide-scale searches of e-records in their corporate or institutional databases. Such a process could easily be incorporated as part of the FRCP 26(a) framework or in FRCP 16 pre-trial orders. Moreover, such discussions dovetail with recent proposed changes to FRCP Rules 16 and 26, which would expressly require both disclosure of "electronic source materials" to opposing parties, and a discussion of any issues surrounding the manner in which such materials will

[41] Turtle, "Natural Language vs. Boolean Query Evaluation: A Comparison of Retrieval Performance," *supra* n.35, at 214. Turtle defines the two main problems as "(1) how can ranked versus unranked lists be compared," and "(2) how can differences in the size of the returned set be reconciled?" In performing his evaluation of most relevant cases found by competing methodologies against "F. Supp" and "F.2d" Westlaw collections, he elected to deem the Boolean set of results as in fact ranked by date (since lawyers depend most on more recent cases). Thus, his method of comparison may not appear to map to the typical unstructured corporate or institutional data set typically subject to e-discovery. Turtle does, however, suggest a number of statistical measures, including measures of average precision and precision against arbitrary rank cutoffs, that may be useful in conducting the contemplated evaluation and bechmarking of competing methodologies. *Id.* at 215.

[42] Jennifer Trant, "Image Retrieval Benchmark Database Service: A Needs Assessment and Preliminary Development Plan," Report Prepared for the Council on Library and Information Resources and the Coalition for Networked Information (revised draft January 2004), Sections 4.1, 4.2 available at http://www.clir.org/pubs/reports/trant04/tranttext.htm.

[43] *Id.,* Section 4.1.

be preserved and made subject to further discovery (*i.e.*, accessed).[44] In this manner, settling on a particular search methodology relatively soon after the onset of litigation would greatly further the aspirational goal expressed in FRCP 1, namely, "to secure the just, speedy, and inexpensive determination of every action."

VI. PROPOSAL FOR TESTING BY AN ACCREDITED STANDARDS BODY

A host of accredited national and international standards bodies exist which could be tapped to perform a "testbed" service for evaluating competing products (and their methodologies), with results transparent to the legal community (and the public) as set out on the organization's website. Sponsoring organizations could include such accredited standards organizations as AIIM,[45] ARMA,[46] or the ISO,[47] working with the American Bar Association or some other legal affiliate.

A successful model for this type of process is run by the Joint Interoperability Test Command (JITC) of the Department of Defense, for the testing of electronic recordkeeping products meeting certain functional requirements.[48] DoD 5015.2-STD[49] "sets forth mandatory baseline functional requirements for Records Management Application (RMA) software used by DoD Components in the implementation of their records management programs. It defines required system interfaces and search criteria to be supported by the RMAs, and describes the minimum records management requirements that must be met based on current [NARA] regulations."[50] JITC's website[51] consists of a "Compliant product register," which lists products that have been verified to comply with the DoD standard's requirements, mandatory and otherwise. The JITC site provides test configurations and summary reports on each software product.

Similarly, there is no reason in principle why search software utilizing varying retrieval methodologies could not be made subject to evaluation and comparison testing with respect to parameters such as Recall and Precision, cued to the needs of the legal community.

VII. NEXT STEPS

A consensus needs to exist that benchmarking competing search products and methodologies as used in the civil discovery context would be a useful exercise, given the IR "state of the art." Assuming a consensus exists, it will be incumbent on interested parties to seek out legal institutions as well as an accredited standards organization which might act as sponsors of a general research effort. The academic information retrieval community should be apprised of the problems faced by lawyers in e-discovery, in order that they weigh in on relevant ongoing research and to make future proposals.

NARA is currently sponsoring a variety of lines of research on the matter of the efficacy of information retrieval. For the past several years, as part of its overall Electronic Record Archives initiative,[52] NARA has partnered with the Georgia Tech Research Institute on the "PERPOS Project" - a Presidential Electronic Records Pilot Operations System, aimed at analyzing software tools which support archival functions such as access and preservation of presidential records, including an evaluation of natural language-based search and retrieve tools used in responding to Freedom of Information Act requests.[53] More recently, NARA has collaborated with staff at the National Institute of Standards and Technology to look into evaluating information retrieval products and methodologies. NARA also is co-sponsoring a project with the Center for Information Policy[54] at the

[44] See Ken Withers, "Two Tiers and a Safe Harbor: Federal Rulemakers Grapple with E-Discovery, *The Federal Lawyer,* vol. 51, No. 8, at 29 (Sept. 2004).

[45] See http://www.aiim.org.

[46] See http://www.arma.org.

[47] See http://www.iso.org/iso/en/ISOOnline.openerpage.

[48] See http://jitc.fhu.disa.mil/recmgt/pp.htm.

[49] See http://www.dtic.mil/whs/directives/corres/pdf/50152std_061902/p50152s.pdf.

[50] See "Records Management Application Compliance Test and Evaluation Process and Procedures," Defense Information Systems Agency, Joint Interoperability Test Command, Fort Huachuca, Arizona, at Section 1.2 (June 2004), available at http://jitc.fhu.disa.mil/recmgt/pap2.doc.

[51] See http://jitc.fhu.disa.mil/recmgt/register.htm.

[52] See http://www.archives.gov/electronic_records_archives/about_era.html.

[53] See http://perpos.gtri.gatech.edu. The research paper by William Underwood cited *supra*, n.16, is one product of this ongoing effort.

[54] See http://www.cip.umd.edu.

University of Maryland's College of Information Studies to survey what constitute current best practices in the area of information retrieval methodologies.

For its part, the Sedona Conference®, under the direction of Richard Braman, has as of early 2005 formalized a Search and Retrieval Sciences special project team. The planned research agenda of this subgroup is to work initially on designing a comprehensive study to test the hypothesis that automated search and retrieval tools can meet or beat existing human search and retrieval techniques. A second contemplated phase of such a project (with the involvement of independent academic and public-private partners) might ultimately lead to the type of testing of current (and future) technologies under objective benchmarked criteria consistent with the proposal set out in this paper. All of the above initial efforts will hopefully lead to a better understanding of how to use search and retrieval products in the future when lawyers (both inside and outside of government) are subject to the next tidal waves of massive e-discovery.