The Sedona Conference Journal

Volume 8 2007

The TREC Legal Track: Origins and Reflections on the First Year

Jason R. Baron

Recommended Citation: Jason R. Baron, *The TREC Legal Track: Origins and Reflections on the First Year*, 8 SEDONA CONF. J. 239 (2007).

THE TREC LEGAL TRACK: ORIGINS AND REFLECTIONS ON THE FIRST YEAR

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

Origins

Imagine you are involved in a multi-party case, and have just been handed an FRCP 34 Request to Produce consisting of 1,726 paragraphs, the last paragraph of which expressly says that all of the prior paragraphs – some of which ask for documents going back 50 years – apply to your client institution's records. Further imagine that you are preserving on the order of 20 million email records that would need to be searched for responsive documents to the request to produce. You cannot review every email, and must rely on automated search software to narrow the possible universe of responsive documents. How would you approach that task? Would you use a set of keywords to apply against the email database (with or without consulting your opposing party)? How do you know that the keywords you have dreamed up adequately locate all, or even a substantial percentage of, relevant email records? Where would you go to find out if there is a better, more accurate, more efficient means to conduct the search? What objective criteria would you use to evaluate how much better an alternative method might be than using keywords alone?

The scenario outlined above was not merely a hypothetical, but a real world set of concerns faced by my agency, the National Archives and Records Administration, in 2002 and 2003.[2] In an effort to separate truly responsive emails from false positives, NARA enlisted 25 archivists and lawyers to manually search through the approximately 1% of "hits" from the 20 million White House emails resulting from simple keyword searches. Though NARA resources were severely strained, the experience was highly instructive on several scores: first, it convinced me that simple keyword searching was a less than adequate tool when applied against very large databases (and indeed, that even a manual review of 1% of a database an order of magnitude larger would simply prove to be impossible, given present agency resources); second, were a similar tsunami wave of litigation to wash over the agency in the future, I would be recommending using more sophisticated and alternative ways of searching for evidence, including methods drawn from notions of fuzzy logic, concept searching, and statistical techniques; third, I found that there was little in the way of present-day research showing what search and information retrieval methods were objectively better to use in a legal context.

The experience of conducting a massive search for White House e-mail records in turn gave birth to the TREC Legal Track. I first conducted a literature review, and became aware of the Blair

[1] The author serves as Director of Litigation at the National Archives and Records Administration (NARA). As NARA's representative to The Sedona Conference®, the author serves as Co-Chair of the Search and Retrieval Sciences Team in Working Group 1, and as Editor-in-Chief of the *Best Practices Commentary on The Use of Search and Information Retrieval Methods in E-Discovery*. Mr. Baron is also an Adjunct Professor at the University of Maryland College of Information Studies, College Park, Maryland. The author is indebted to Ellen Voorhees and Ian Soboroff at the National Institute of Standards and Technology, and to fellow track coordinators Doug Oard and Dave Lewis, for all of their very substantial contributions toward the successful launch of the first year of the TREC Legal Track. I also wish to thank Doug Oard for his substantive comments on an earlier draft of this article, although I take full responsibility for any errors that remain uncorrected notwithstanding Doug's considerable efforts to keep me honest in fairly and accurately reporting out our scientific results. I also wish to thank Ezequiel Berdichevsky, George Washington University Law School (J.D. candidate, '08), for his expert editorial assistance. The views expressed herein are the author's alone, and do not necessarily reflect the official views of NARA, or the views of any other person or institution.

[2] For a more extensive discussion of NARA's response to the discovery production in *U.S. v. Philip Morris et al.* (D.D.C.), *see* Jason R. Baron, "Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used In E-Discovery," 6 *Sedona Conference Journal* 237 (2005).

and Maron study.[3] At the same time, I became aware of the existence of TREC through discussions with NARA staff, and in late 2004 I approached the computer scientist at NIST who serves as project manager for TREC, Dr. Ellen Voorhees, to gauge NIST's interest in sponsoring research into evaluating search methods in a legal context. She was very open to the idea of a new legal track, and even encouraged NARA's last-minute participation in the then-current 2005 track year, at least on a pilot basis, if a suitable database and resources could be found.[4] At about the same time, I also approached my University of Maryland colleague Dr. Douglas Oard, who holds a joint appointment in the Institute for Advanced Computer Studies and the College of Information Studies, telling my tale of NARA's experience and the huge, practical problem facing lawyers. I hoped he would find the challenge interesting, and luckily he did. Later in 2005, we drew up the first draft of a proposal that would eventually be approved by Ellen Voorhees and her colleagues on the TREC steering committee.

I understood that developing a proposal for a new NIST research track devoted to e-discovery concerns in the abstract would only be the first step; executing a plan to have the track actually come to fruition meant seeking the assistance of others who would be prepared to volunteer considerable time and industry to the effort. Accordingly, I approached Richard Braman, Executive Director of The Sedona Conference®, with my ideas for a TREC Legal Track and what kind of resources I would need. He immediately recognized that this was a project that would dovetail nicely with the newly formed Search and Retrieval Sciences Team (SRS Team) within the Working Group on Best Practices for Document Retention and Production (WG1), which I had agreed to co-chair, inasmuch as the mission of the SRS Team includes assisting in future research efforts aimed at enhancing the quality of the legal process.

With the backing of The Sedona Conference®, the proposal Doug Oard and I submitted to TREC in the fall of 2005 consisted of the following three main elements: first, a sketch of possible choices for a possible test collection, including the Enron public database of email,[5] and the State Department cables database (*see* n.4, *supra*); second, our plan to draw on a small group of Sedona Conference® members to draft hypothetical complaints and topics; and third, our proposal for using the auspices of The Sedona Conference® to solicit lawyers, legal assistants, law clerks, law students, and other professionals with ties to the legal sector, to volunteer their time to "assess" documents for responsiveness. The proposal was accepted with the idea that we would further flesh it out at a November 2005 planning workshop, held during the 2005 annual TREC conference in Gaithersburg, Maryland.

It was David D. Lewis, Ph.D., a private consultant and long-time member of the TREC steering committee, who, prior to the planning workshop, made a strong case for us using a collection of nearly seven million publicly available documents from the "Master Settlement Agreement" database – a collection of tobacco documents produced in discovery proceedings related to lawsuits filed by the Attorneys General of several U.S. states against seven major tobacco organizations (including five tobacco companies and two research institutes).[6] The version of the MSA repository proposed for our use came from the Illinois Institute of Technology (IIT) Complex Document Information Processing (CDIP) 1.0 Collection, consisting of 1.5 terabytes of scanned document images, metadata, and optical character recognition (OCR) output. As described in the TREC 2006 Legal Track Overview paper ("Track Overview"),[7] the main strength of the collection consists of its wide range of document genres, including letters, memos, budgets, reports, agendas, minutes, plans, transcripts, scientific articles, and some email - where document length varies from

[3] *See* David C. Blair & M.E. Maron, "An Evaluation of Retrieval Effectiveness for a Full-text Document-Retrieval System," *Communications of the ACM* 28:3 (1985), 289-299.

[4] I initially presented NIST with an idea of using a database of 400,000 State Department cable telegrams from the Nixon Administration (circa 1973-74), which at the time was being prepared by the State Department and NARA for public release through a portal on the NARA web site. Although this collection represents, at best, a legacy form of test collection, *sans* email and other forms of modern-day "ESI," I was confident it would nevertheless be endlessly topic-rich and thus would be of interest for use in any long-term research project. However, a number of factors conspired against participating in the 2005 round of TREC, including technical problems NARA was experiencing with getting the database ready for public access; the cables database was, however, eventually loaded for public access in 2006, and remains a fascinating repository of government documents available to the public, as well as a possible test collection for future research. *See* http://www.archives.gov/press/press-releases/2006/nr06-71.html.

[5] The public version of the Enron collection of documents comprises over 500,000 emails with over 130,000 unique addresses; other related databases exist as well. *See* Tamer Elsayed and Douglas W. Oard, "The Enron and W3C Collections," powerpoint presentation at ICAIL 2007, DESI Workshop (n.13, *infra*), *available at* http://www.umiacs.umd.edu/~oard/desi-ws/.

[6] For the text of the MSA and related documents, *see* http://ag.ca.gov/tobacco/msa.php.

[7] *See* Jason R. Baron, David D. Lewis, Douglas W. Oard, "TREC 2006 Legal Track Overview," *available at* http://trec.nist.gov/pubs/trec15/t15_proceedings.html (#3).

one to many thousands of pages. It was also understood from the beginning that the MSA document collection was itself the collective *output* of prior legal discovery proceedings, and thus, in some sense, is unrepresentative of a pure heterogeneous corporate collection of ESI - even one found on one of the tobacco defendant's own present-day servers. Nevertheless, as stated in the Track Overview, "[o]ur worries about that point [were] mitigated to some extent, however," by the fact that the MSA documents originated in seven different organizations in response to hundreds of document requests in multiple cases.[8] Speaking for myself, I perhaps was more confident than my colleagues that whatever possible bias might exist could be controlled for through the use of a range of topics as part of the evaluation exercise that, in the main, were substantially different than past MSA discovery requests.

At the November 2005 planning workshop, Doug requested that Dave Lewis and I demonstrate how lawyers might proceed to "negotiate" a search protocol consisting of a Boolean string of keywords. So far as I am aware, prior to the December 2006 amendments to the Federal Rules of Civil Procedure, it was the rare, exceptional case that involved such open, transparent negotiations over what keywords would be used by the responding party. Anecdotally speaking, I believe this to remain the case at least as of the first year after the new Rules' adoption - however, it is also my understanding that more sophisticated discussions at least amongst parties and their own consultants are also increasingly taking place as to what constitutes the proper construction of search string keywords and syntax. (Nevertheless, it remains the case, as of this writing, that there is no reported case law where parties have asked a court to intervene in connection with deciding the form of a properly-constructed Boolean string of search words to be used.) The planning workshop also established an ideal target of 50 topics for the lawyers to construct for research purposes.

Notes from Year 1

One of the first tasks for the legal track consisted of constructing a usable set of topics modeled on real-world civil discovery requests. The five hypothetical complaints and associated topics in the form of requests to produce that were developed are described in more detail in the TREC 2006 Legal Overview paper. The complaints consisted of (i) an investigation by the so-called Federal Watchdog Commission into a fictional tobacco company's improper campaign contributions; (ii) a federal consumer protection lawsuit challenging a fictional tobacco company's product placement decisions with respect to television, film, and theatre shows watched by children; (iii) an insider trading securities lawsuit involving fictional tobacco executives; (iv) a state court antitrust lawsuit involving the movement of a fictional company's tobacco-related commerce in the San Diego, California area; and (v) a federal product liability lawsuit involving defective surgical devices as demonstrated in animal testing.[9] Care was taken to ensure that all of these complaints were readily seen as fictional in nature, even if the ultimate aim of topic development would be to find real-world documents in the MSA collection from actual companies. My thanks must go to Sedona Conference® colleagues Ryan Bilbrey, Conor Crowley, Joe Looby and Stephanie Mendelsohn, who all volunteered in drafting the hypothetical complaints and dreaming up potential topics for use in the research, and who thereafter participated in the Boolean negotiations as described below.

For each of the 43 topics used in the first year effort, I engaged in "mock negotiations" with my fellow complaint drafters, in which we essentially agreed to play the respective of opposing counsel on the propounding and receiving end of e-discovery requests. To this end, the final XML topic file that was produced as part of the track included an initially proposed Boolean query, followed by a rejoinder from the opposing party which, unless stated in the documentation, became the final agreed-upon negotiated Boolean query.

Thus, for the antitrust complaint, a typical example of a topic in the form of a request to produce and subsequent negotiations was as follows:

[8] *Id.*, Section 2.

[9] For a complete description of the topic set used in the TREC 2006 legal track, *see* http://trec-legal.umiacs.umd.edu/ (scroll down to heading "Materials from the TREC-2006 Legal Track," subheading "Test Collection," and filename "Evaluation topics").

Topic 27: All documents discussing or relating to the placement of product logos at events held in the State of California.

Defendant counsel's initial proposal for a Boolean string of search terms for this topic was: ("product placement" AND logos AND California). Plaintiff counsel countered with [("product placement" OR advertis! OR market! OR promot!) AND (logo! OR symbol OR mascot OR marque OR mark) AND (California OR cal. OR calif. OR CA)].

Clearly, many other terms could have been substituted into the proposed string. However, one must start somewhere, and for purposes of the TREC Legal Track the outcome of these negotiations over what constituted one form of a reasonable Boolean string was intended to serve the purpose of providing a baseline, both for track coordinators as well as the participants in the track, as to how one might approach the more general search problem posed by each topic query. For the track coordinators, it was essential that someone perform a "Boolean run," to obtain what were deemed "baseline" results from the MSA document collection using the final negotiated Boolean string; theoretically, each participant in the track would then be free to utilize and build upon the expressly stated Boolean negotiated query, in performing their own alternative ways of searching through the test collection for potentially responsive documents.

Thereafter, the success of the first year of the Legal Track was dependent on (a) attracting a nucleus of participating institutions to submit "runs" based on whatever home-grown search methodologies they wished to use; (b) pooling the results of the submitted runs so as to create properly sized "judgment pools" for the assessment phase of the track; and (c) engaging a sufficient number of volunteer assessors to assist in reviewing actual documents. Six institutions from around the world (Hummingbird, now Open Text; the National University of Singapore; Sabir Research; the University of Maryland; the University of Missouri-Kansas City; and York University), submitted a total of 31 ranked runs with no more than 5,000 documents per topic. In addition, a run known as the "expert manual searcher run" was also employed, consisting of a research assistant performing her own searches of the test collection through automated means, leading to production of up to 100 documents per topic from the collection that she felt would be unlikely to be retrieved by fully automated systems. Computer scientists at NIST, working with Dave Lewis, performed the necessary statistical stratification to come up with combined judgment pools per topic; these consisted of on average about 800 documents per topic, collected across all of the submitted runs.

The assessment phase of the track began in earnest during the first week of August 2006, in which 35 volunteers, including participants from Sedona Conference® member law firms and legal technology companies, NARA staff, and participating law schools, assessed a total of 32,738 documents for 40 completed topics. The volunteers included eight lawyers, ten law students, three paralegals, one professional archivist, one historian, and several individuals with degrees in science or finance. The affiliations of volunteers included Sedona Conference® members Bank of America, FTI Consulting, H5 Technologies Inc., Lewis & Roca LLP, Preston Gates LLP, as well as additional volunteers from NARA, George Washington University Law School, George Mason University School of Law, Reasonable Discovery LLC, the New Mexico State Attorney General's Office, and three private individuals. The Overview Paper provides a more complete summary of the problems and issues that arose in the assessment phase, which ran fairly smoothly over the course of approximately six weeks. Individual assessors were given a "how to" guide for working with an online judging platform, and individual questions were responded to by track coordinators on an ad hoc basis. Although some effort was put into defining relevancy, assessors were basically instructed that a document should be considered relevant when the reference to a topic was found in the document, and to use their best judgment. A special rule that required leaving the document unjudged if relevance could not be easily determined upon cursory examination was imposed for the occasional document that surpassed 300 pages.

A passage from the Overview paper describes the experience of assessors in somewhat more detail:

> Some of the assessors went beyond the text of the topic (the complaint, the production request, and the Boolean queries) to perform additional legal research which they viewed as helpful to the exercise. For example, the assessor for Topic 30 researched at greater length what the numbered statutory code provisions were corresponding to the California Cartwright Act, to ensure that all documents containing such references, with or without reference to the Cartwright Act itself, would be marked as responsive. The assessor on Topic 10 performed independent research into the ban on tobacco advertising, as an aid to understanding what documents might be expected to be found in response to a topic involving tobacco product placement in television or film. One assessor asked for assistance on the definition of one of the keywords in the topic, leading to additional research conducted on the Internet.
>
> Some differences were observed in how liberally or narrowly assessors viewed the scope of their discretion to find responsiveness. In some exceptional cases, assessors were willing to find responsiveness even where a key term might be missing, if the document was otherwise sufficiently generic and might yet be viewed as responsive with the aid of further research. For example, the assessor for Topic 9 ("All documents discussing, referencing or relating to payment of compensation to 20th Century Fox Corporation for placement of products and/or brands in a film production"), marked certain documents as relevant even if the film company was not expressly mentioned, where the context indicated that the company might be involved. In most cases, however, assessors appeared to adopt relatively restrictive interpretations on what met the mark for relevance.
>
> Assessors reported some confusion as to whether they should exclude documents that might be within the literal scope of a production request when read in isolation, but which weren't relevant to the main thrust of the associated complaint (i.e., the document had nothing to do with the causes of action in the lawsuit or investigation). The question of scope arose in particular for production requests associated with the one complaint that on its face did not involve allegations against the tobacco industry (but which was instead about medical devices). Topic 49, which coupled that complaint with a production request for "[a]ll documents created between 1962 and 1999 referencing or including warnings or draft warnings used in the United States," proved to be particularly problematic because it was read by the assessor as being aimed at warnings for faulty medical devices. Not surprisingly, no relevant documents were found for topic 49. It was therefore removed from the evaluation because topics with no known relevant documents can not be used to compare the effectiveness of alternative system designs. Results are therefore reported for the remaining 39 topics.
>
> As is often the case, assessors found some unintended ambiguity in the topics, either due to grammatical construction of the topic (e.g., what did the word "their" refer to), or due to inherent ambiguity embedded within words or concepts (e.g., what constitutes "lobbying efforts," "advertising," "marketing," and "promotion"). For one assessor, the word "event" (in a topic asking for all documents relating to the placement of product logos at events held in California), prompted them to consult the Random House Dictionary, where the word is defined as "something that occurs in a certain place during a particular interval of time." Therefore, in this assessor's view, documents that mentioned such activities as the America's Cup Race, speed skiing, auto racing, Hispanic Cultural events, Swing jam weekend, an antiviolence campaign, a country music festival, and an anti-smoking campaign called "Tobacco is Whacko," were all properly within the scope of the topic.

> Another miscellaneous concern of one or more assessors involved how to deal with documents containing foreign language text. The track coordinators instructed assessors to make judgments based on English portions of documents, or otherwise mark the document as unsure.
>
> In general, assessors took their jobs very seriously. A number of assessors made a second pass through their document set to resolve anomalies or to revisit judgments based on knowledge gained on the first pass.[10]

Recruitment for assessors remained a substantial concern through the months of July, August and September 2006, especially towards the end when my fellow track coordinators made the surprise announcement that we really should also be performing a "dual assessment" round, meaning that additional volunteers would be needed to review a subset of the documents already assessed by earlier volunteers. However, my anecdotal experience from the first year confirms that lawyers and law students are actually more than willing to perform otherwise tedious document review if they believe - correctly, in this case - that they are advancing legitimate scientific research. They even reportedly (based on post-assessment surveys) have some amount of fun in doing so. I certainly have been in no position to correct them or suggest otherwise![11] I trust that the experience bears out for the second year of the track, where we are initially targeting law students to volunteer in performing assessment duties for a new round of 50 new topics based on four new hypothetical complaints.

Scientific Findings

The first year of the TREC Legal Track has been the subject of the aforementioned Overview paper, four additional written submissions by track participants to the 15th TREC Proceedings,[12] two research papers presented at a recent conference on search and retrieval issues in Palo Alto,[13] and at least two other Ph.D. students are known to be currently working on additional scholarship. Subject to an important caveat, I will here venture to summarize some of the findings from the first year. The caveat: the Overview paper the track coordinators initially produced does not purport to represent a comprehensive analysis of the first year data. Rather, at best, it constitutes a preliminary stab at evaluating the first year's results after performing some data analysis. Accordingly, the present piece should not be viewed as anything near a comprehensive report on the findings of the TREC Legal Track. Much more data analysis and revisiting of what actually has taken place in year 1, possibly combined with the results of year 2 and future years, will be needed, before more definitive conclusions emerge from the results of the overall study. That said, my summary follows:

First, none of the participating institutions submitting automated runs based on alternative ways of performing searches was found to have "the" answer to the search problem faced by lawyers. To the contrary, the results accompanying Figure 3 of the Overview paper show that the best runs from three of the participating institutions were nearly indistinguishable under one widely-reported statistical measure (mean R-precision),[14] and one of those runs (submitted by Hummingbird), was artificially limited in its role as the *baseline* Boolean run. Indeed, it might well have been surprising if one automated method was better at performing searches in a statistically significant way than all others. The extent to which no individual run did better than the baseline Boolean run is, however, somewhat curious, counter-intuitive, and deserves of further reflection.[15] The Overview paper makes

[10] TREC 2006 Legal Track Overview, *supra*, n.7, at Section 4.2.

[11] The experience may be perhaps likened to the "reward" of performing endless Bluebook citation checking, a task more-or-less enthusiastically embraced by generations of newly recruited members of law review.

[12] *See generally* TREC Legal Track web page for research papers, *available at* http://trec-legal.umiacs.umd.edu/ (under "2006 Results").

[13] *See* written submissions to the Eleventh International Conference on Artificial Intelligence and Law, Supporting Search and Sensemaking for Electronically Stored Information in Discovery Proceedings (ICAIL 2007, DESI Workshop), Palo Alto, CA, June 4, 2007, *available at* http://www.umiacs.umd.edu/~oard/desi-ws/ (under "Research Papers"). *See also* Jason R. Baron and Paul Thompson, "The Search Problem Posed By Large Heterogeneous Data Sets in Litigation: Possible Future Approaches to Research," ICAIL 2007 Conference paper, *available at* http://www.umiacs.umd.edu/~oard/desi-ws/ (under "Related Paper presented at Main Conference").

[14] "Precision" is the measure of the relevant documents that have been retrieved, as a percentage of the total number of documents that have been retrieved, at any given point or "rank" in a search. As defined in the Overview paper, "mean R-precision" is computed as an average across all topics of the density of relevant documents at rank R, where R is the number of known relevant documents for each topic. TREC Legal Track Overview, Section 5.2 ("R-Precision").

[15] For further discussion of this point and an excellent overall technical analysis of year 1 of the Legal Track, *see* Stephen Tomlinson, "Experiments with the Negotiated Boolean Queries of the TREC 2006 Legal Discovery Track," http://trec.nist.gov/pubs/trec15/t15_proceedings.html (#57).

the point that with only six participating institutions, "we are nowhere near exhausting the design space for search techniques, so ways may yet be found to achieve improvements that are not available to a Boolean system."[16]

A second result from the same data shows that the "expert manual searcher" outperformed all of the automated runs under a mean-R precision measure. While an important finding, one must be careful not to over-interpret this result: the manual search being referred to was one research assistant's interactive attempts to construct 100 documents per topic, using an automated interface for conducting searches, and based on her superior knowledge of the MSA collection gathered from prior research. The outcome of her efforts was emphatically *not* the same as if someone with no *a priori* knowledge of the collection performed a straight "manual" review of some subset of documents. As lawyers, we must be very careful to understand how the term "expert manual searcher" has been deployed.

A third, most intriguing result, emerges in the Overview paper's section entitled "Uniques analysis." As the paper relates, "[o]ne way of characterizing the results of different approaches to searching is to examine the contribution of each approach to the total set of known relevant documents."[17] Figure 2 of the Overview paper proceeds to show that "on average across the 39 topics, 57% of the known relevant documents were found by the reference Boolean query (i.e., either uniquely by the reference Boolean system, or by the reference Boolean system and also one or more other systems)."[18] Another 11% of relevant documents were found by the expert manual searcher. This leaves 32% as the combined average amount of known relevant documents across all 39 topics that were found by systems other than the reference Boolean system.

What are we to make of this figure, that I have elsewhere called "dark matter" potentially representing the relevant documents not found by simple keyword searches or other traditional means?[19] Coming out of Year 1 of TREC, the figure represents a combined result - a pooling of *all* of the other search methods employed by Year 1 participants beyond the "baseline" Boolean method - and thus is useful only when understood in that important context. That is, the figure does not show that any particular alternative method will guarantee that a relevant document can be found. On the other hand, if anything, the figure represents a lower bound on the number of undiscovered relevant documents that might possibly exist were still other means employed to search for them.

My fellow Legal Track coordinators would caution that any strong claims regarding what this data represents are at best premature, based on the level of analysis conducted to date. Nevertheless, I wish to suggest that the legal community should sit up and take notice of what this set of empirical data from Year 1 may yet prove to show. Lawyers historically have accepted results obtained through the use of simple searches, based on keywords alone. More recent scholarship has pointed the way towards showing that despite the power of keywords, there are known deficiencies in over-reliance on simple keyword searches alone as the method of choice.[20] The 32% figure represents a universe of potentially relevant evidence not found in a concededly large haystack of ESI - even after parties had engaged in a relatively sophisticated set of negotiations with the outcome of the reference Boolean string of terms. This suggests, at a minimum, that parties should be open to experimenting with alternative forms of search methods and techniques, and using them in a combined fashion, so as to maximize the rate of recall to be obtained, *i.e.*, maximizing the proportion of all relevant documents found as the result of any particular search or searches.[21]

[16] TREC Legal Track Overview, Section 5.2. The paper goes on to describe the fact that not all Boolean runs are created equal, which reflects the fact that, in one case, the Boolean run used the initial rather than the final queries, and in other run there was incomplete or inconsistent automated support for implementing extended or non-Boolean operators (such as where the topic requests used truncated terms, *i.e.*, "!" to denote all possible further suffixes to a particular keyword or portion thereof, and/or where the topics employed the device "w/3," *i.e.*, "within 3"). *Id.*

[17] *Id.*, Section 5.1.

[18] *Id.*

[19] *See* Jason R. Baron, "Thinking Outside The Boolean Box: Metastrategies for Conducting Legally Defensible Searches In An Expanding ESI Universe," powerpoint presentation at ICAIL 2007, DESI Workshop, *available at* http://www.umiacs.umd.edu/~oard/desi-ws/.

[20] *See* "The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery," Section 3, 8 *Sedona Conf. J.* --- (2007) ("Sedona Search Commentary"); George L. Paul and Jason R. Baron, "Information Inflation: Can the Legal System Adapt?," 13 *Rich. J. Law & Tech.* 10 (2007), at Paragraph 37, *available at* http://law.richmond.edu/jolt/v13i3/article10.pdf.

[21] *See* Sedona Search Commentary, *supra*, n.20, at Section 4; G. Paul & J. Baron, *supra*, n.20, at Paragraph 41.

Much greater exploration of this form of "dark matter" is obviously necessary, particularly in terms of what kinds of searches are most likely to produce unique relevant documents not otherwise found by traditional keyword searches and/or Boolean techniques. In doing so, we may take a significant step towards performing the type of objective benchmarking of competing search technologies that I have elsewhere suggested is imperative, given a fundamental desire on the part of one or both parties in litigation to understand how well e-discovery obligations are being met, as well as in providing assistance to the trier of fact to determine what has constituted a "reasonable" search for responsive evidence.[22]

The TREC Legal Track also has produced empirical data on two "process"-oriented issues that may be of general interest. First, some attempt was made to assess the effects of differing assessor interpretations, through the use of what is known as "dual assessments." Based on a small sample of 50 documents per topic, the study showed a moderate rate of overall agreement between assessors (where so-called "kappa values" can range from -1 for complete disagreement to +1 for complete agreement, the mean value of kappa for the legal track was +0.49). In the case of Topic 24 ("All documents referencing the Federal Election Commission dated subsequent to 2001"), there was actually a slight negative kappa figure (-0.037), indicating more disagreement than agreement as between two assessors on their respective relevance judgments. The entire matter of the divergence of views among law students, paralegals, junior, and senior lawyers in judging relevant documents for relevance is a subject wide open for future research.

Finally, another form of "process" metric was tabulated from post-assessment survey reports, namely, how much time assessors spent in performing their document reviews. (Recall that documents could range from 1 to over 300 pages, with the 300 page figure representing a cut off for performing the normal expected full-text review.) Based on time data from 16 of the 35 participants, representing 39% of the overall assessment effort (12,743 out of the 32,738 assessments), the reported review rate per hour ranged from a low of 12.33, to a high of 67.5, with an average of 24.7 documents per hour. This figure certainly has great value for estimating the commitment levels volunteers will need to make in future iterations of the legal track, but may also be of interest to a greater community of researchers.

The data gathered during the first year of the track will be subject to continued study for years to come, both on its own and in conjunction with data gathered in the second and any future years of the existence of the track. I welcome all such efforts. Indeed, the value of the TREC Legal Track exercise is that the test collection, protocols, fundamental methodology, and research results are all transparent and openly available to the world to replicate (and improve upon).

Additional Observations

As my computer science colleagues informed me near the end of the first year's cycle of work, the first year of a new TREC track is intended to function as an experimental pilot, where the bugs and kinks are worked out, in order to perform solid research in second and succeeding years. (Had I only known!) Although we were able to push out a 2006 Legal Track Overview paper, as stated above, substantive scholarship analyzing the results of the first year of the track, combined with the results obtained in the second and any succeeding years, largely remains to be undertaken.

As the Overview paper notes, "[p]erhaps the greatest accomplishment of the TREC 2006 Legal Track is that it happened at all. More than 50 volunteers contributed to assembling and distributing the collection, creating topics, developing systems, managing submissions, creating pools, judging relevance, developing metrics, creating scoring software, analyzing results, and coordinating those activities. This has yielded the results that we would hope for from any TREC track in its first year: (1) a reusable test collection to support future research, (2) a set of baseline results to which future research can be compared, and (3) a community of researchers who bring a variety of perspectives to these important challenges."[23] The fact that the Legal Track exists, and that this type of

[22] *See* J. Baron, *Toward A Federal Benchmarking Standard, supra*, n.2. G. Paul and J. Baron, *Information Inflation, supra,* n.20, at Paragraph 46.
[23] TREC Legal Track Overview, *supra*, n.7, at Section 6.

study holds the potential to aid in objectively modeling present practices in the area of civil e-discovery, continues to be enormously professionally rewarding.

I do have a continuing concern about the methodology of pooled relevance assessments employed as part of TREC. As stated in my paper with Paul Thompson,

> The assumption was made that if a large number of systems participated in TREC that by judging for relevance the top n documents retrieved by each system this pooled set of judged documents would provide a representative sample of the document collection as a whole, particularly if many systems with diverse ranking algorithms participated. This assumption seemed reasonable, but is now breaking down as collections are becoming increasingly large. The problem is made worse if there are also a large number of relevant documents for each query. For example, assume that a document collection has one billion documents and that around one million of these may be relevant for a given query. Judging the top several hundred documents retrieved for each query for each of 20 or 30 participating systems will not give a good estimate of recall for the collection as a whole.[24]

In other words, I continue to hope that, either as part of TREC or in additional research to come, the real-world concerns of lawyers facing the need to find responsive ESI in exponentially greater universes of data will continue to be adequately modeled. Nevertheless, as I have made clear, I believe the TREC legal track project itself holds the potential to produce useful and important results of benefit to the legal profession as a whole, as we all struggle to make sense of litigation demands and obligations in this brave new world of ESI.

[24] *See* J. Baron and P. Thompson, *The Search Problem Posed By Large Heterogeneous Data Sets in Litigation*, *supra*, n.13, at Section D.