Secrets of Search

Ralph C. Losey[1]
Jackson Lewis, LLP

Secrets of Search: Part One

This is the first installment of a three-part article revealing the secrets of *legal ESI search experts*. I am referring to the few technophiles, lawyers, and scientists in the e-discovery world who specialize in the search for relevant electronic evidence in large chaotic collections of ESI such as email. This exposé will include a secret deeply hidden in shadows, one only half-known by a few. Before I can get to the dark secret, I must lay bare a few other search secrets that are not so hidden. (Note: this article originally appeared on my blog: e-DiscoveryTeam.com.)

A Secret of Search Already Known to Many

The first secret of search here exposed is the same kind of secret as those revealed in *Spilling the Beans on a Dirty Little Secret of Most Trial Lawyers.* You probably have heard it already, especially if you have read Judge Peck's famous *wake-up call* opinion in *William A. Gross Construction Associates, Inc. v. American Manufacturers Mutual Insurance Co.*, 256 F.R.D. 134, 136 (S.D.N.Y. 2009). He repeated it again recently in his article *Predictive Coding: Reading the Judicial Tea Leaves*, (Law Tech. News, Oct. 17, 2011), that I wrote about in *Judge Peck Calls Upon Lawyers to Use Artificial Intelligence and Jason Baron Warns of a Dark Future of Information Burn-Out If We Don't*. Despite these writings and many CLEs on the subjects, most of your less informed colleagues in the law still don't know these things, much less litigants or the public at large. It would seem that Jason R. Baron's dark vision of a future where no one can find anything is still a very real possibility.

The wake-up call on search has a long way to go before it is a *shot heard round the world*. I am reminded of that on almost a daily basis as I interact, usually indirectly, with opposing counsel in employment cases around the country. They often insist on antiquated search methods. So bear with me while I begin by repeating what you may have already heard before. I promise that the exposé of these more common secrets will also set the stage for revealing the *seventh step* of

---

Ralph C. Losey is a partner of Jackson Lewis, LLP, where he lead's the firm's Electronic Discovery practice group. The opinions expressed in this article are his own, and not necessarily those of his law firm, clients, or University. Ralph can be reached at Ralph.Losey@JacksonLewis.com. Ralph is the author of four books on electronic discovery published by West Thomson and the ABA, and is the publisher and principle author of the *e-Discovery Team Blog* at http://e-discoveryteam.com, where this article first appeared. Ralph has limited his legal practice to electronic discovery since 2006, with a special interest in effective and economical search and review of electronic evidence in large, chaotic ESI collections. Ralph is also an Adjunct Professor of Law at the University of Florida where he teaches introductory and advanced electronic discovery, both in-person and online. A practicing attorney who supports the 700 attorneys in his law firm in 48 offices around the country, Ralph has been involved with computers and the law since he began private practice in 1980. His full biography may be found at RalphLosey.com.

© Ralph Losey 2012

incompetence causality that I mentioned in last week's blog, *Tell Me Why?*, and the one deep dark search secret that you probably have not heard before. Yes, the one is related to the other.

The First Secret: Keywords Search Is Remarkably Ineffective at *Recall*

First of all, and let me put this in very plain vernacular so that it will sink in, *keyword search sucks*. It does not work, that is, unless you consider a method that misses 80% of relevant evidence to be a successful method. Keyword search alone only catches 20% of relevant evidence in a large, complex dataset, such as an email collection. Yes, it works on Google, it works on Lexis and Westlaw, but it sucks in the legal world of evidence gathering. It only provides reliable recall value when used as part of a multimodal process that uses other search methods and quality controls, such as iterative testing, sampling, and adjustments. It fails miserably when used in the *Go Fish* context of blind guessing, which is the negotiated method still used by most lawyers today. I have written about this many times before and will not repeat it here again. *See eg. Child's Game of "Go Fish" is a Poor Model for e-Discovery Search*.

Keyword Search Still Has a Place in Best Practices

Keyword search still has a place at the table of Twenty-First Century search, but only when used as part of a multimodal search package with other search tools, and only when the multimodal search is used properly with iterative processes, real-time adjustments, testing, sampling, expert input and supervision, and other quality control procedures. For one very sophisticated example of what I mean, consider the following description by Recommind, Inc. of their patented Predictive Coding process that is embedded in their software review tool, *Axcelerate*. Their software uses highly advanced AI guided search processes, but keywords are still one of the many search tools used in that process:

> The Predictive Coding starts with a person knowledgeable about the matter, typically a lawyer, developing an understanding of the corpus while identifying a small number of documents that are representative of the category(ies) to be reviewed and coded (i.e. relevance, responsiveness, privilege, issue-relation). This case manager uses sophisticated search and analytical tools, **including keyword, Boolean** and concept search, concept grouping and more than 40 other automatically populated filters collectively referred to as Predictive Analytics™, to identify probative documents for each category to be reviewed and coded. The case manager then drops each small seed set of documents into its relevant category and starts the "training" process, whereby the system uses each seed set to identify and prioritize all substantively similar documents over the complete corpus.7 The case manager and review team (if any) then review and code all "computer suggested" documents to ensure their proper categorization and further calibrate the system. This iterative step is repeated … (emphasis added)
>
> The final step in the process employs Predictive Sampling™ methodology to ensure the accuracy and completeness of the Predictive Coding process (i.e. precision and recall) within an acceptable error rate …

© Ralph Losey 2012

Sklar, Howard, *Using Built-In Sampling to Overcome Defensibility Concerns with Computer-Expedited Review,* Recommind DESI IV Position Paper.

Here is the diagram that Recommind now uses to describe their overall process, which they were kind enough to give me permission to use:

Note that keyword search, including Boolean refinements, is used as part of the seed set generation step, which they call the first Predictive Analytics step in their multimodal process. By the way, as I will explain when I reveal the second search secret in a minute, that 95%-99% accuracy statement you see in their chart should be taken with a very large grain of salt. Still, aside from the dubious percentages claimed in this chart, the actual search methods and processes used are good.

Proof of the Inadequacies of Keyword Search When Not Used as Part of a Multimodal Process

Want scientific proof of the incompetence of keyword search alone when **not** used as part of a modern multimodal process? Look at the landmark research on Boolean search by information scientists David Blair and M.E. Maron in 1985. The study involved a 40,000 document case (350,000 pages). The lawyers, who were experts in keyword search, estimated that the Boolean searches they ran uncovered 75% of the relevant documents. In fact, they had only found 20%. Blair, David C., & Maron, M. E., *An evaluation of retrieval effectiveness for a full-text document-retrieval system*; Communications of the ACM Volume 28, Issue 3 (March 1985).

Delusion is a wonderful thing, is it not? *We are confident our search terms uncovered 75% of the relevant evidence.* Really? Still, no one likes the fool who points out that the emperor is naked, especially the emperor and his tailors who frequently pay all of the bills. Still, here I must go, where angels fear to tread. I must point out what science says.

Please join me in this Quixotic quest. Spread the word. Somebody has to do it. We must all continue to tell the unpopular truth, lest Baron's dark vision of a future world comes true. A world of injustice where relevant evidence is lost in ESI skyscrapers of junk, where cases are decided on false testimony and whim. We don't want that world. We have worked way too hard over centuries to build our systems of justice to let a few billion terabytes of ESI destroy them. But destroy them they will, if we are complacent. Baron's dystopian nightmares are real.

Want more recent scientific proof of the Emperor's old clothes? See the research conducted by the National Institute of Standards and Technology TREC Legal Track. It has again confirmed that keyword search alone still finds only about 20%, on average, of relevant ESI in the search of

© Ralph Losey 2012

a large data-set. In batch tests in 2009 of negotiated keyword terms they did much worse. Hedlin, Tomlinson, Baron, Oard, 2009 TREC Legal Track Overview, TREC legal track at §3.10.9. The Boolean searches had a mean precision ratio of 39%, **but recall averaged less than 4%**. Yes! You read that right. The negotiated keywords missed 96% of the documents. *Oopsie*. I wonder how many times lawyers have done this in practice and never known it? *We are confident our search terms uncovered 75% of the relevant evidence.*

Please note this awful 4% recall came out of what they called the *batch tasks*, where there were no subject matter experts, testing, or appeals. These safeguards were present only in the *interactive tasks*. The batch tasks are thus like my *Go Fish* scenario, where people simply guess keywords in the blind, and never test, sample, refine and iterate.

The same research also shows that alternative multimodal methods do much better. They still use some keyword based search tools, but also use predictive coding and other artificial intelligence algorithms with seed-set iteration and sampling methodologies. I wrote about these new methods in *Judge Peck Calls Upon Lawyers to Use Artificial Intelligence and Jason Baron Warns of a Dark Future of Information Burn-Out If We Don't* and before that in *The Information Explosion and a Great Article by Grossman and Cormack on Legal Search*.

Want still more recent proof? The final report on the 2010 TREC tests has not been completed, but many participants reports are final. I have done some deep digging and read most of them, and the draft summary report, in order to try to bring to you the latest evidence on search. *See* http://trec-legal.umiacs.umd.edu/. The 2010 tests once again confirm our little secret on the *absurd ineffectiveness* of keyword search alone. The confirmation comes inadvertently from the tests done by a fine team of information science graduate students from the Indian Statistical Institute, Kolkata, in West Bengal, India. They participated in the 2010 TREC Legal Interactive task in Topic 301 and Topic 302. (Yes, science is very international, including information science and TREC Legal Track.) They performed what proved to be an interesting (to me) experiment, although for reasons other than what they intended.

The Indian Statistical Institute had an AI predictive software coding tool using clustering techniques that they wanted to test. But the software could not handle the high volumes of email involved in the 2010 test: 685,592 items. So they had no choice but to cull down the amount of email somewhat before they could use their software. For that reason they decided to use keywords to cull down the *corpus* (a term information scientists love to use) before running their AI clustering software. Here is their own description of the process:

> We attempted to apply DFR-BM25 ranking model on the TREC legal corpus. We chose Terrier 3.0 as this toolkit has most of the IR methods implemented within. But as we received the TREC legal data set we realized that it would be difficult to manage such a large volume of data. So, we decided to reduce the corpus size by Boolean retrieval. We chose Lemur 4.11 as it supports various useful Boolean query operators which would suit our purpose. On the set obtained from Boolean retrieval we decided to apply ranked retrieval techniques. … The use of Boolean retrieval has the disadvantage that it will limit further search to the documents retrieved at this stage and have an adverse effect on our recall performance. But it would scale down the huge corpus size considerably (see Table 1) and enable us to perform our experiments on a smaller set which would reduce processing time.

That use of keyword Boolean as an upfront filter turns out to have been a mistake, at least in so far as any quest for good recall was concerned. Who knows, maybe they thought their keywords would be better than the lawyer derived keywords in the famous Blair Maron study. I see this kind of mistake made by opposing counsel all of the time. *We are confident our search terms will uncover 75% of the relevant evidence.*

© Ralph Losey 2012

They think their keywords are so good that they could not possibly miss 80% of all relevant document in the corpus. They have an almost superstitious belief in the magical power of keywords, and think that their Boolean is better that your Boolean. Hogwash! All keyword search sucks, no matter who you are, or how many lawsuits you've won, or Google sites you've found.

The computer algorithms used in the 1985 Blair Maron test are essentially the same used today for keyword search. Keyword search is pretty simple index matching stuff. Antiquated software really. It works fine in academic settings with artificiality controlled data sets or organized databases, but it does not survive contact with the real world where words and symbols are chaotic and vague, just like the people who create them. In real world email collections the meaning of documents is hidden in subtle, and not-so-subtle, word and phrase variations, misspellings, *abbreves*, slang, *obtusity*, etc. In reality, when large data sets are involved, no human is smart enough to guess the right keywords.

Getting back to the 2010 TREC study, in topic 301 the use of Boolean retrieval allowed the scientists from India to reduce the initial corpus from 685,592 to 2,715. Then they ran their sophisticated software on the whittled down corpus. The final metrics must have been disappointing. The TREC judges found that their *precision* in topic 301 was pretty good. It was 87% (meaning 87% of the items retrieved were determined to be relevant after an appeal process). But their recall was simply terrible, only 3% (meaning their method failed to retrieve an estimated 97% of the relevant documents in the original 685,592 collection). Random guessing might have done as well in the recall department, maybe even in the *F1* measure (the harmonic mean of *precision* and *recall*).

In their other interactive task topic 302 the results were comparable. They attained a precision rate of 69% and a recall rate of 9%. Again this means that they left 91% of the relevant documents on the table and only managed to find 9% of the relevant documents.

The Second Search Secret (Known Only to a Few): The *Gold Standard* to Measure Review is Really Made Out of Lead

The so-called *gold standard* used to judge recall and precision rates in information science studies is human review. This brings up an even more important secret of search, a subtle secret known only to a few. Experiments in TREC conducted well before the legal track even began showed that we humans are very poor at making relevancy determinations in large data sets. This is a very *inconvenient* truth because it puts all precision and recall measurements in doubt. It means that the recall and precision measures we use are more like rough estimates than calculations. It may be the measurements can be improved by expensive remedial, three-pass expert human reviews, and other methods, but even that has yet to be proven. *But see* Cormack, Grossman, *Inconsistent Assessment of Responsiveness in E-Discovery: Difference of Opinion or Human Error?* (2011) (humans can agree and create a gold standard if relevance is defined clearly enough to reviewers and if objective mistakes by reviewers (as opposed to subjective disagreements) are identified and corrected).

This secret of human inadequacy and resulting measurement vagaries in large data-set reviews has been known in the information science world since at least 2000. I understand from inquiring of Doug Oard, a well-known information scientist and one of the TREC Legal Track founders, that the problem of the "fuzziness" of relevance judgments remains an important and ongoing discussion among scientists. Apparently the "fuzziness" issue is far less of a problem when simply trying to compare one system with another, and determine which one is better, than it is when trying to report a correct ("absolute") value for some quantity such as recall or precision. I corresponded with Doug Oard on this issue and he advised me that:

> The Legal Track of TREC has generated quite a lot of attention to the problem of absolute evaluation simply because the law, properly, has a need for that information. But the law also has a need for relative evaluation (which can help to answer questions like

© Ralph Losey 2012

> "did you use the best available approach under the circumstances"), and "fuzziness" is well-known to have only limited effects on such relative comparisons.

So even though our measurements are too fuzzy to ever really say with any assurance that there is 95%-99% accuracy, it can tell us how one method compares with another. For instance, we can know that keyword search sucks when compared with multimodal, we just cannot know exactly how well either of them do.

The fuzziness of recall measurements may explain the wide divergences in measurements of search effectiveness. For instance, it could explain how the 2009 batch tests of keywords only measured a remarkably low 4% recall rate. 2009 TREC Legal Track Overview, TREC legal track at §3.10.9. It may have been better than that, more in line with the usual 20% recall rates that other experiments have shown, but we do not really know because the gold standard measurements can fluctuate wildly. Again this is all because average one-pass human review is known to be unreliable.

William Webber

The fuzziness issue is one of several important topics addressed in an interesting paper written this year by a young information scientist, William Webber, entitled *Re-examining the Effectiveness of Manual Review*. Webber, shown right, is an Australian now doing his post-doctoral work with Professor Oard. His paper arose out of an e-discovery search conference held this year in China of all places, the *SIGIR 2011 Information Retrieval for E-Discovery (SIRE) Workshop*, July 28, 2011, Beijing, China. You may have heard about this event from some of its other attendees, including Jason R. Baron, Patrick Oot, Jonathan Redgrave, Conor Crowley, Bill Butterfield, Doug Oard, and David Lewis. Anyway, Webber in his China paper explains:

> It is well-known that human assessors frequently disagree on the relevance of a document to a topic. Voorhees [2000] found that experienced TREC assessors, albeit working from only sentence-length topic descriptions, had an average overlap (size of intersection divided by size of union) of between 40% and 50% on the documents they judged to be relevant. Voorhees concludes that 65% recall at 65% precision is the best retrieval effectiveness achievable, given the inherent uncertainty in human judgments of relevance. Bailey et al. [2008] survey other studies giving similar levels of inter-assessor agreement.

Can anyone validly claim absolute recall or precision rates in large data set reviews that is more than 65% when the determinations are made by single pass human review? Apparently not. Maybe double or triple pass review can create a true gold standard. I know that is what TREC is now striving for using sampling and an appeals process in the experiments since 2009. But has that been proven? I don't think so, and least that is my impression after reading Webber's work.

Webber's China paper goes on to explain the well-known study by Roitblat, Kershaw, and Oot, *Document categorization in legal electronic discovery: computer classification vs. manual review*. Journal of the American Society for Information Science and Technology, 61(1):70–80, 2010.

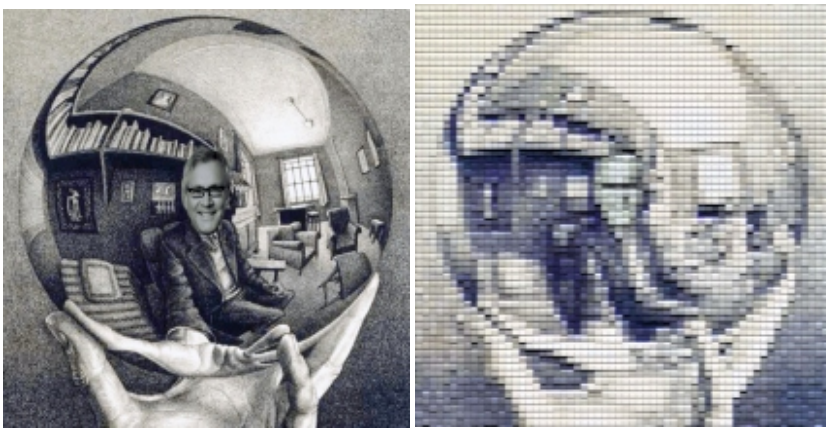
> For their study, the authors revisit the outcome of an earlier, in-house manual review. The original review surveyed a corpus of 2.3 million documents in response to a regulatory request, and produced 176,440 as responsive to the request; the process took four months and cost almost $14 million. Roitblat et al. had two automated systems and two manual review teams review the documents again for relevance to the original request. The automated systems worked on the entire corpus; the manual review teams looked at a sample of 5,000 documents. Roitblat et al. (Table 1) found that the overlap between the

© Ralph Losey 2012

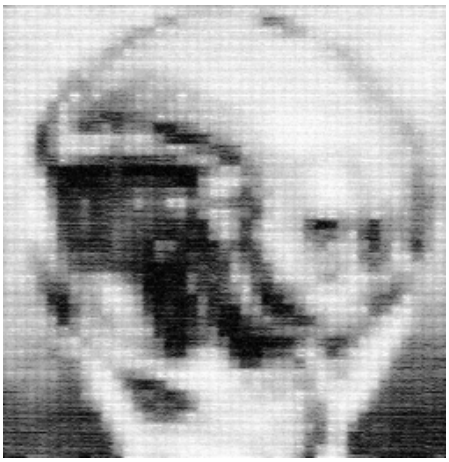
> relevance sets of the two manual teams was only 28%, even lower than the 40% to 50% observed in Voorhees [2000] for TREC AdHoc assessors. The overlap between the new and the original productions was also low, 16% for each of the manual teams, and 21% and 23% for the automatic systems. …
>
> The effectiveness scores calculated on the original production seemingly show that the automated systems are as reliable as the manual reviewers. However, as Roitblat et al. note, the original production is a questionable gold standard, since it likely is subject to the same variability in human assessment that the study itself demonstrates. Instead, the claim Roitblat et al. make for automated review is a more cautious one; namely, that two manual reviews are no more likely to produce results consistent with each other than an automated review is with either of them.
>
> Given the remarkably low level of agreement observed by Roitblat et al., their conclusion might seem a less than reassuring one; an attorney might ask not, which of these methods is superior, but, is either of these methods acceptable? More importantly, the study does not address the attorney's fundamental question: does automated or does manual review result in a production that more reliably meets the overseeing attorney's conception of relevance?

Think about that. Lawyers are on average even worse than non-lawyers in making relevancy reviews. We only agree 28% of the time, compared to earlier non-lawyer tests noted by Voorhess showing 40% agreement rates. The 40% agreement rates showed that the best retrieval effectiveness achievable, given the inherent uncertainty in human judgments of relevance, was only 65% recall and 65% precision. *See* Ellen M. Voorhees, *Variations in relevance judgments and the measurement of retrieval effectiveness*, 36:5 Information Processing & Management 697, 701 (2000). I wondered what an even lower 28% agreement rate as found in the Roitblat et al. study meant? In private correspondence with Webber to prepare this essay, he advised me that a 28% agreement rate produces a mean precision and recall rate of 44%.

It seems to me as if Webber and Voorhees are saying that on average the best that lawyers can ever do using the *so-called gold standard* of human review for measurement is something like 65%-44% recall? Any measurements higher than that are suspect because the gold standard itself is suspect. I think Webber, Voorhees, and others are saying that the human relevancy determinations lens we are using to study these processes is too fuzzy, too out of focus, to give us any real confidence in exactly what we are seeing, but the fuzzy lens does allow us to compare one method against another.



© Ralph Losey 2012



The Triple Pass Solution

Although I do not understand the math on the fuzziness issue, I understand it in an intuitive way from over thirty years of arguing with other attorneys and judges over relevancy. I also know from the thousands of vague requests for production I have read and tried to respond to. In the law we use a kind of triple pass quality control method based on disagreements of experts. The triple-pass method has evolved in the common law tradition over the past few centuries. We never simply rely on one tired lawyer's opinion. One lawyer expresses their view on relevance, then another lawyer, opposing counsel, uses their independent judgment to either agree or disagree, and, if they disagree, to object. A third expert, a judge, then hears argument from *both* sides and makes a final determination. Without such triple expert input and review the determination of the legal relevance of evidence in legal proceedings would also be unreliable.

TREC has been trying to use such a triple pass method since 2009 to buttress the accuracy of its findings. The first reviewers make their determinations, then the participants make theirs. If the participants disagree, then the participants can ask for a ruling from the subject matter expert who had been guiding the participants with up to ten hours of consults. The first review team has no such appeal rights and far less guidance. Also, the first pass reviewers cannot present their side of the arguments to the judge. Not surprisingly under these conditions, if and when the participants appeal, the reports show that the expert judges usually rule with the participants. They have, after all, had ongoing *ex parte* communications with them and don't hear from the other side. Not exactly the same triple play as in the real world of American justice, but it is far better than the flawed single human review that Voorhees initially studied. Moreover, it is improving each year as TREC's experiments are refined. To get closer to real world practice would require a lot more money for the experiments.

In my view the inherent fuzziness (or not) of human relevance capacities is a significant problem that needs a lot of further study. Think of the implications on our current legal practice. (*Hint* - this has something to do with the *seventh insight* into trial lawyer resistance, as I will explain in Part III of *Secrets of Search*.)

Not Too Fuzzy To Allow Valid Comparisons

Although the measures are fuzzy, they are not too fuzzy to make comparisons between reviews. So, for instance, you can compare two human reviews and use the differences to show just how vague and inaccurate human review really is. This would be

© Ralph Losey 2012

a comparison to establish the fuzziness of the gold standard you use to make recall, precision and other measurements.

The study by Roitblat *et al.* sponsored by the Electronic Discovery Institute (EDI) did just that. It proved the incredible inconsistencies of single pass human review in large data-sets. This study examined a real world event where Verizon paid $14,000,000 for contract reviewers to review 2.3 million documents in four months. (This is, by the way, a cost of $6.09 per document for review and logging only, a pretty good price for those days.) A second review by other reviewers commissioned by the study only agreed with 16% of the first determinations. Yes, there was only a 16% agreement rate. Incredible. Does that not suggest likely error rates of 84%?!

Surely this study by EDI is the death-blow to large-scale human reviews that are not in some way computer assisted to at least cull out documents before review. Why should anyone spend $14 Million for such a poor quality product after seeing this study? (Yet, I'm told they still do this in the world of mergers and acquisitions and second reviews.) This is especially true when you consider that machine assisted review is much faster and less expensive. Further, as the studies also show, the computer assisted review is at least as reliable as *most* of the human reviewers (but maybe not *all*, as will be explained (that is yet another search secret)).

With these limitations of human review and measurements in mind consider the paper by Maura R. Grossman and Gordon V. Cormack, which analyzed the 2009 TREC legal track studies on this issue. *Technology-assisted review in e-discovery can be more effective and more efficient than exhaustive manual review*. Richmond Journal of Law and Technology, 17(3):11:1–48, 2011. I have written about their paper before: *The Information Explosion and a Great Article by Grossman and Cormack on Legal Search*. Grossman and Cormack found that:

> [T]he 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 review companies – had they conducted a manual review of the entire document collection.

Id. at 4. This was good research and a great paper as I've noted before, but the gold standard was again *just* human reviewers and so subject to the vagaries of fuzzy measurement when it comes to calculating absolute values. As mentioned, TREC is working on this issue with their appeals process, but due to economic constraints, it still differs from actual practice in several ways as mentioned. The first reviewers have relatively limited upfront instruction and training on the relevance issues, only limited contact with subject matter experts during the review, no testing or sampling feedback, and no appeal rights.

Also, the human review in TREC 2009 did not meet minimum ethical standards of supervision established by most state Bar associations that have considered the propriety of delegated review to contract lawyers. Most Bar associations require direct supervision of contract lawyers by counsel of record, and, in my opinion, that requires direct, ongoing contact to supervise. Aside from the supervision issue, the statistics were skewed by a one-sided appeals process where the judge only heard one-side of a relevancy argument from the party they trained in relevance. It reminds me of a secret for getting an "A" in law school from some Professors: just tell them what you think they want to hear, not what you really think.

For that reason the win observed by Grossman and Cormack may not say as much about technology as it does about methodologies. Also, the paper focuses on the *two* technology-assisted processes that were better. What about the *other* technology-assisted processes that were not better?

Aside from these methodology concerns, as Webber points out, none of the studies so far, by TREC or anyone else, have addressed the key issue of concern to lawyers:

> … which is **not** how much different review methods agreed or disagree with each other (as in the study by Roitblat et al. [2010]), nor even how close automated or manual review methods turn out to have come to the topic authority's gold standard (as in the study by Grossman and Cormack [2011]). Rather, it is this: which method can a supervising attorney, actively involved in the process of production, most reliably employ to achieve their overriding goal, to create a production consistent with their conception of relevance. (emphasis added)

Let's Spend the Money Necessary to Turn Lead Into Gold

How can the studies and scientific research ever give us an answer to Webber's question, to our question, if the measuring device, the gold standard, is too fuzzy to make absolute measurements, just comparative ones? It seems to me the solution is a series of multimillion dollar scientific experiments, instead of the shoe-string-budget type projects we have had so far. We need experiments where the three-pass gold standard developed by the law is employed, where time-consuming quality controls are employed for both automated and manual reviews, and for various types of combined multimodal methods. We need to transform our lead standard into a bonafide gold standard. Yes, that means expensive relevancy determinations made by three-expert, triple-pass, statistically checked, state of the art reviews. But we have to go for the gold. We need absolute measurements we can trust and bank on to do justice.

These kind of scientific experiments will be expensive, but I think we should do it. Gold for gold. But it is worth it. After all, billions of dollars in fees are spent each year on e-discovery review. Trillions of dollars more ride on the outcome of litigation. What if a method we employ does not work as well as we think, and a key privileged document is overlooked? It could be game over. What if we end up looking at way too many of the wrong documents? How much money is lost already each year doing that? What if the 50% recall measurement you made is rejected by the court as too low, when it fact it was a 95% recall rate? What if the 95% recall measurement is really just 50%? With these constraints on measurements, how much recall should be considered legally sufficient? Should these measurements be used at all? Or should we just use methods that compare well with others, that use best practices, and not try to quantify precision and recall?

We need to really know what we are doing. We cannot just be alchemists playing with quicksilver. We need real science to verify exactly how accurate our methods are, not just compare them. We need to know more than comparative values. We need absolute measures. Heisenberg be damned, we need certainty in the law, or at least a lot more of it than we have now. Sure we know that computer assisted review is faster, cheaper and at least as good as average human review. But what recall rates do any of them really achieve? Sure we know that *keyword search sucks*, that multimodal is comparatively much better. But how much better? Is the true rate of recall for keywords 20% or 4%, or is it 44%? What is the true rate of recall for our top multimodal search techniques today, the ones like Recommind's that uses keywords, Predictive Coding and a variety of other tools and methods? Is it 97% or is it 44%, or less? We need hard numbers, not just comparisons.

Law and IT alone cannot give us the answers. The e-discovery team also needs scientists. We need to know what kind of recall rates and precision rates we are capable of measuring with a confidence level in the 90s, not just 44% to 65%. Is plus or minus 44% recall really the best anyone can hope for? Is the confidence level such that a measure of 44% recall might actually be much higher, might actually be 98%. And visa versa? Are we just kidding ourselves with all of the

© Ralph Losey 2012

recall measures we now have? Apparently so. All we can tell for sure right now is which method is better than another. That is not enough for the law. We need much more certainty than that.

The secret is now out and we have to address it. We have to talk about it. We have to perform experiments and peer review these experiments. I personally think the law's triple-pass methods with the latest quality control techniques will produce significantly higher rates of agreement, maybe even in the 90s, but who actually knows until we pay for the experiments?

I think the research that TREC and EDI have done to date are a good start, but not the final word by any means. We need many more open scientific experiments. The testing must be improved and several more groups should join in. Our major information science universities worldwide should join in. So should the National Science Foundation and other charitable organizations. So too should the big companies that can afford to finance pure research. How about Google? IBM? Microsoft? EMC? HP? Xerox? How about your company? Every e-discovery company should have some skin in this game.

The budgets of the testing organizations need to be ramped up for all of these experiments. We need gold to make measurements with a true gold standard, to give us real answers, not just qualified comparisons. I will make a donation and participate in fundraisers for that kind of scientific research. Will you? Will your company or firm join in?

There is still more to the insights contained in Webber's research in *Re-examining the Effectiveness of Manual Review*. But this first installment of *Secrets of Search* is already too long, so that, my friends, will have to wait again for Part Two. Webber's work and the discussion so far sets the stage for an even deeper and darker secret of search, the one that ties into the *seventh insight* to lawyer resistance to e-discovery. That will come at the conclusion of next week's blog, *Secrets of Search – Part Two*.

Secrets of Search: Part Two

This is Part Two of the article on the *Secrets of Search*, which was in turn a sequel to two blogs that I wrote before that at e-discoveryteam.com: *Spilling the Beans on a Dirty Little Secret of Most Trial Lawyers* and *Tell Me Why?* In *Secrets of Search – Part One* we left off with a review of some of the analysis on *fuzziness* of recall measurements included in the August 2011 research report of information scientist, William Webber: *Re-examining the Effectiveness of Manual Review*.

We begin part two with the meat of his report and another esoteric search secret. This will finally set the stage for the deepest secret of all and the seventh insight into trial lawyer resistance to e-discovery. That comes in the third a final installment of *Secrets of Search*.

Summarizing Part One of this Blog Post and the First Two Secrets of Search

I can quickly summarize the first two secrets with popular slang: keyword search sucks, and so does manual review (although not quite as bad), and because most manual review sucks, most *so-called objective* measurements of precision and recall are unreliable. Sorry to go all negative

© Ralph Losey 2012

on you, but only by outing these *not-so-little* search secrets can we establish a solid foundation for our efforts with the discovery of electronic evidence. The truth must be told, even if it hurts.

I also explained that keyword search would not be so bad if it were not done blindly like a game of *Go Fish,* where it achieves really pathetic recall percentages in the 4% to 20% range (the TREC batch tasks). It still has a place with smarter software and improved, cooperation based *Where's Waldo* type methods and quality controls. In that same vein I explained that manual review can probably also be made good enough for accurate scientific measurements. But, in order to do so, the manual reviews would have to replicate the state-of-the-art methods we have developed in private practice, and that is expensive. I concluded that we should come up with the money for better scientific research so we could afford to do that. We could then develop and test a new gold standard for objective search measurements. Scientific research could then test, accurately measure, and guide the latest hybrid processes the profession is developing for computer assisted review.

Another conclusion you could also fairly draw is that since the law already accepts linear manual review and keyword search as reasonable methods to respond to discovery requests, the law has set a very low standard and so we do not need better science. All you need to do to establish that an alternative method is legally reasonable is to show that it does as well as the previously accepted keyword and manual methods. That kind of comparison sets a low hurdle, one that even our existing fuzzy research proves we have already met. This means we already have a green light under the law, or logically we should have, to proceed with computer assisted review. Judge Peck's article on predictive coding stated an obvious logical conclusion based upon the evidence.

You could, and I think *should*, also conclude that any expectation that computer assisted reviews have to be near perfect to be acceptable is misplaced. The claim that some vendor's make as to near perfection by their search methods is counter to existing scientific research. It is wrong, mere marketing puff, because the manual based measurements of recall and precision are too fuzzy to measure that closely. If any computer assisted or other type of review comes up with 44%, it might in fact be perfect by an actual objective standard, and visa versa. Allegedly objective measurements of high recall rates in search is, for the time being at least, an illusion. It is a dangerous delusion too because this misinformation could be used against producing parties to try to drive up the costs of production for ulterior motives. Let's start getting real about objective recall claims.

In any event, most computer assisted search is already better than average keyword or manual search, so it should be accepted as reasonable under the law without confidence inflation. We don't need perfection in the law, we don't need to keep reviewing and re-reviewing to try to reach some magic, *way-too-high* measure of recall. Although we should always try to get more and more of the truth, we should always try to improve, we should also remember that there is only so much truth that any of us can afford when faced with big data sets and limited financial resources.

As I have said time and again when discussing e-discovery efforts in general, including preservation related efforts, the law demands *reasonable* efforts not perfection. Now science buttresses this position in document productions by showing that we have *never* had perfection in search of large numbers of documents, not with manual, and certainly not with keyword, and, here is the kicker, it is not possible to objectively measure it anyway!

At least not yet. Not until we start taking our ignorance of the processes of search and discovery as a disease. Then maybe we will start allocating our charitable and scientific efforts accordingly, so we can have better measurements. Then with reliable and more accurate measurements, with solid gold objective standards, we can create more clearly defined best practices, ones that are not surrounded with marketing fluff. More on this later, but first let's move onto another secret that

comes out of Webber's research. I'm afraid it will complicate matters even further, but life is often like that. We live in a very complex and imperfect world.

The Third Search Secret (Known Only to a Very Few): e-Discovery *Watson* May Still Not Be Able to Beat Our Champions

Webber's report reveals that there is more to the *man versus machine* question than we first thought. His drill down analysis of the 2009 TREC interactive tasks shows that the computer assisted reviews were not the hands down victors over human reviewers as we first thought, at least not victors over many of the well-trained, *exceptional* reviewer men and women. Putting aside the whole fuzziness issue, Webber's research suggests that the TREC and EDI tests so far have been the equivalent of putting Watson up against the average Jeopardy contestants, you know, the poor losers you see each week who, like me, usually fail to guess anything right.

The real test of IBM's Watson, the real proof, didn't come until Watson went up against the champions, the true professionals at the game. We have not seen that yet in TREC or the EDI studies. But the current organizers know this, and they are trying to level the playing field with multi-pass reviews and, as Webber notes, trying to answer the question we lawyers really want to know, the one that has not been answered yet, namely which Watson, *which method can an attorney most reliably employ to create a production consistent with their conception of relevance.*

Webber in his research and report digs deep into the TREC 2009 results and looked at the precision and recall rates of individual first pass reviewers. *Re-examining the Effectiveness of Manual Review*. He found that while Grossman and Cormack were accurate to say that overall two of the top machines did better than man, the details showed that:

> Only for Topic 203 does the best automated system clearly outperform the best manual reviewer. As before, the professional manual review team for Topic 207 stands out. Several reviewers outperform the best automated system, and even the weaker individual reviewers have both precision and recall above 0.5.

This means the best team of professional reviewers who participated in Topic 207 actually beat the best machines! They did this in spite of the mentioned inequities in training, supervision, and appeal. Did you know that secret? I'm told that topic 203 was an easy one having to do with junk filters, but still, easy or not, the human team won.

There is still more to this secret. When you drill down even further you find that certain individual reviewers on each team topic actually beat the best machines on each topic in some way, even if their entire human team did not. That's right, the top machines were defeated by a few champion humans in most every event. Humans won even though they were disadvantaged by not having an even playing field. I guaranty that this is a secret you have never heard before (unless you went to China) because Webber just discovered it from his painstaking analysis of the 2009 TREC results. Chin up contract reviewers, the reports of your death have been greatly exaggerated. Watson has not beat you yet, in fact, Watson still needs you to set up the gold standard to determine who wins.

© Ralph Losey 2012

Webber's research shows that a competition between the best Watsons and best reviewers is still a very close race where humans often win. Please note this analysis assumes no time limits or cost limits for the human review, which are, of course, false assumptions in legal practice. This is why pure manual review is still, or should be, as dead as a doornail. The future is a team approach where humans use machines in a nonlinear fashion, not *visa versa*. More on this later.

Webber's findings are the result of something that is not a secret to anyone who has ever been involved in a large search project, that all reviewers are *not* created equal. Some are far better than others. There are many good psychological, intelligence, and project management and methodology reasons for this, especially the management and methodology issues. *See eg* the *must read* guest blog by contract review attorney Larry Chapin, *Contract Coders: e-Discovery's "Wasting Asset"?*

The facts supporting Webber's findings on individual reviewer excellence are shown in Figure 2 of his paper on the variability in review team reliability. *Re-examining the Effectiveness of Manual Review*. The small red crosses in each figure (except flawed task 205) show the computer's best efforts. Note how many individual reviewers (a bin is 500 documents that were reviewed by one specific reviewer) were able to beat the computer's best efforts in either precision, or recall, or both. They are shown as either to the right or above the red cross. If above this means they were more precise. If to the right, they had better recall.

© Ralph Losey 2012

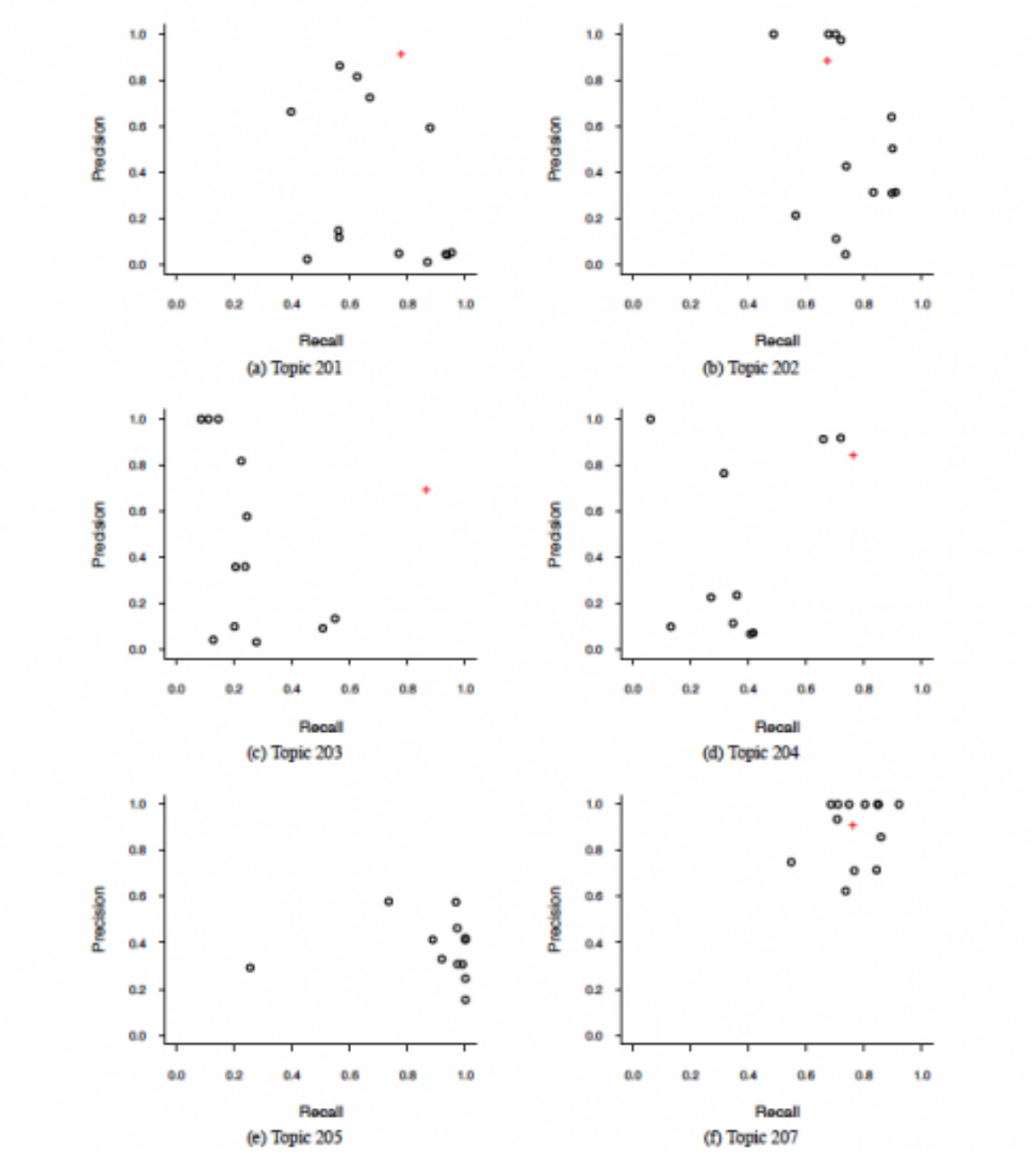


Figure 2: Assessor precision and recall, extrapolated to population, for Topics 201–205 and Topic 207. Each circle represents the reliability of a core bin. The red cross in each figure except that for Topic 205 gives the performance of the best automated retrieval effort, as listed in Table 2.

William Webber summarizes these findings in his blog recently by saying:

> The best reviewers have a reliability at or above that of the technology-assisted system, with recall at 0.7 and precision at 0.9, while other reviewers have recall and precision scores as low as 0.1. This suggests that using more reliable reviewers, or (more to the point) a better review process, would lead to substantially more consistent and better quality review. In particular, the assessment process at TREC provided only for assessors to receive written instructions from the topic authority, not for the TA to actively manage the assessment process, by (for instance) performing an early check on

© Ralph Losey 2012

> assessments and correcting misconceptions of relevance or excluding unreliable assessors. Now, such supervision of review teams by overseeing attorneys may (regrettably) not always occur in real productions, but it should surely represent best practice.

Webber, W., *How Accurate Can Manual Review Be?* IREvalEtAl (12/15/11). Better review process and project management are key, which is the next part of the secret.

How to Be *Better Than Borg*

Webber's research shows that some of the human reviewers in TREC stood out as *better than Borg*. They beat the machines. Does this really surprise anyone in the review industry? Sure, human review may be (should be) dead as a way to review *all documents* in large-scale reviews, but it is alive and well as the most reliable method for final check of computer *suggested* coding, a final check for classifications like privilege before production.

This is a picture of humans and machines working together as a *team*, as friends, but not as Borg implants where machines dictate, nor as human slaves where smart machines are not allowed. I know that George Socha, whom I quoted in *Tell Me Why?*, much like one of my fictional heroes, Jean Luc Picard, was glad to escape the Borg enslavement. So too would most contract lawyers who are stuck in dead-end review jobs with cruel employers. By this way, his embarrassing, unprofessional, *contract lawyers as slaves* mentality was shown dramatically by some of the reader comments to *Contract Coders: e-Discovery's "Wasting Asset"?* They report incredible incidents of abuse by some law firms. Some of the private complaints I have heard from document reviewers about abuse and mismanagement are even worse than these public comments. The primary rule of any relationship must always be mutual respect. That applies to contract lawyers, and, if they are a part of your team, even to artificial intelligence agents like Watson, Siri, and their predictive coding cousins. Get to know and understand your entire team and to appreciate their respective strengths and weaknesses.

Webber's study shows that the quality of the individual human reviewers on a team is paramount. He makes several specific recommendations in section 3.4 of his report for improving review team quality, including:

> Dual assessment, for instance, can help catch random errors of inattention, while second review by an authoritative reviewer such as the supervising attorney can correct misconceptions of relevance during the review process, and adjust for assessor errors once it is complete [Webber et al., 2010]. …
>
> [S]ignificant divergence from the median appears to be a partial, though not infallible, indicator of reviewer unreliability. A simple approach to improving review team quality is to exclude those reviewers whose proportion relevant are significantly different from the median, and re-apportion their work to the more reliable reviewers. …
>
> Fully excluding reviewers based solely on the proportion of documents they find relevant is a crude technique. Nevertheless, the results of this section suggest that this proportion is a useful, if only partial, indicator of reliability, one which could be combined with additional evidence to alert review managers when their review process is diverging from a controlled state. It may be that review teams with better processes, such as the team from Topic 207, already use such techniques. Therefore, they need to be considered when a benchmark for manual review quality is being established, against which automatic techniques can be compared.

Webber's conclusion summarizes his findings and bears close scrutiny, so I quote it here in full:

© Ralph Losey 2012

> 5. CONCLUSIONS. The original review from which Roitblat et al. draw their data cost $14 million, and took four months of 100-hour weeks to complete. The cost, effort, and delay underline the need for automated review techniques, provided they can be shown to be reliable. Given the strong disagreement between manual reviews, even some loss in review accuracy might be acceptable for the efficiency gained. If, though, automated methods can conclusively be demonstrated to be not just cheaper, but more reliable, than manual review, then the choice requires no hesitation. Moreover, such an achievement for automated text-processing technology would mark an epoch not just in the legal domain, but in the wider world.
>
> Two recent studies have examined this question, and advanced evidence that automated retrieval is at least as consistent as manual review [Roitblat et al., 2010], and in fact seems to be more reliable [Grossman and Cormack, 2011]. These results are suggestive, but (we argue) not conclusive as they stand. For the latter study in particular (leaving questions of potential bias in the appeals process aside), it is questionable whether the assessment processes employed in the track truly are representative of a good quality manual review process.
>
> We have provided evidence of the greatly varying quality of reviewers within each review team, indicating a lack of process control (unsurprising since for four of the seven topics the reviewers were not a genuine team). The best manual reviewers were found to be as good as the best automated systems, even with the asymmetry in the evaluation setup. The one, professional team that does manage greater internal consistency in their assessors is also the one team that, as group, outperforms the best automated method. We have also pointed out a simple, statistically based method for improving process control, by observing the proportion of documents found relevant by each assessor, and counseling or excluding those who appear to be outliers.
>
> Above all, it seems that previous studies (and this one, too) have not directly addressed the crucial question, which is not how much different review methods agreed or disagree with each other (as in the study by Roitblat et al. [2010]), nor even how close automated or manual review methods turn out to have come to the topic authority's gold standard (as in the study by Grossman and Cormack [2011]). Rather, it is this: which method can a supervising attorney, actively involved in the process of production, most reliably employ to achieve their overriding goal, to create a production consistent with their conception of relevance. There is good, though (we argue) so far inconclusive, evidence that an automated method of production can be as reliable a means to this end as a (much more expensive) full manual review. Quantifying the tradeoff between manual effort and automation, and validating protocols for verifying the correctness of either approach in practice, are particularly relevant in the multi-stage, hybrid work-flows of contemporary legal review and production. Given the importance of the question, we believe that it merits the effort of a more conclusive empirical answer.

The evidence shows that it is at least very difficult, *perhaps even impossible* (I await for more science to form a definite opinion), for us humans to maintain the concentration necessary to review tens of thousands of documents, day in and day out, for weeks. Sure we can do it for a few hours, and for 500 or so documents, but for 8-10 hours a day with tens or hundreds of thousands of documents for weeks on end? I doubt it. We need help. We need suggestive coding. We need a team that includes smart computers.

Know Your Team's Strengths and Weaknesses

The challenge to human reviewers becomes ridiculously hard when you ask them to not only make relevancy calls, but, at the same time, to also make privilege calls, and confidentiality calls, and, here is the worst, multiple case issues categorization calls, a/k/a, issue tagging. Experience shows that the human mind cannot really handle more than five or six case issues at a time, at

© Ralph Losey 2012

least when reviewing all day. But I keep hearing tales of lawyers asking reviewers to make ten to twenty case issue calls for weeks on end. If you think it is hard to get consistent relevancy calls, just think of the problem of putting relevant docs into ten to twenty buckets. Might as well throw darts. That is a scientific experiment I'd like to see, one testing the efficacy of case issue tags. How many categorizations can humans really handle before it becomes a complete waste of time?

I call on e-discovery lawyers everywhere to better understand their team members and stop asking them to do the impossible. Issue tagging must be kept simple and straightforward for the human members of your team to deal with it. The ten to twenty case-issue tags is a complete waste of time, perhaps with the exception of seed-set training, as thereafter Watson has no such limitations. But in so far as the final, out-the-door review goes, do not encumber your humans with mission impossible tasks. Know your team members, their strengths and weaknesses. Know what the humans do best, like catch obvious bloopers beyond the kin of present day AI agents, and do not expect them to be as tireless as machines.

The review process improvements mentioned by Webber, and other safeguards touted by most professional review companies who truly understand and care about the strengths and weaknesses of their team, will certainly mitigate against the problems inherent in all human review. In my mind the most important of these are experience, training, mutual respect, good working conditions, motivation, and quality controls, including quick terminations or reassignments when called for. More innovative methods are, I believe, just around the corner, such as game theory applications discussed by Lawrence Chapin in *Contract Coders: e-Discovery's "Wasting Asset"?* But the bottom line will always be that computers are much better at complex repetitive drudgery tasks such as reviewing tens of thousands, or millions, of documents. Thankfully our minds are not designed for this, whereas computers are.

Reviewers Need Subject Matter Expertise and Money Motivation

Based on my experience as a reviewer and supervisor, the human challenges to make review determinations over large scales of data are magnified when the human reviewers are not themselves subject matter experts, and magnified even further when the reviewers have no experience in the process. This was not only true of all of the student volunteer reviewers at TREC, but is also sometimes true in real world practice as well. That is just invited error. Training is part of the solution to that.

It is also my supposition that in our culture the errors are magnified again when there is no, or inadequate, compensation provided. All TREC reviewers were unpaid volunteers except for the professional review team members. They were paid by the companies they work for, although those companies were not paid, and the rate of pay to the individuals is unknown. Still, can you be surprised that the top reviewers, the ones who beat the machines, were all paid, and only a few of the student teams came close? In our culture money is a powerful motivator. That is another reason to have better funded experiments that come closer to real world conditions. The test subjects in our experiments should be paid.

© Ralph Losey 2012

The same principle applies in the real world too. Contract review companies should stop competing on price alone and we consumers should stop being fooled by that. Quality is job number one, or should be. Do you really think the company with the lowest price is providing the best service? Do you think their attorney reviewers don't resent this kind of low pay, sometimes in the $15-$20 per hour range. Most of these lawyers have six-figure student loans to pay off. They deserve a fair wage and, I hypothesize, will perform better if they are paid better.

To test my *money-motivation theory* I'd love to see an experiment where one review team is paid $25 an hour, and another is paid $75. Be real and let them know which team they are on. Then ask both to review the same documents involving weeks of grueling, boring work. Add in the typical vagaries of relevance, and equal supervision and training, and then see which team does better. Maybe add another variation where there is a stick added to the carrot and you can be fired for too many mistakes. Anyone willing to fund such a study? A contract review company perhaps? (Doubtful!) Better yet, perhaps there is a tech company out there willing to do so, one that competes with cheap human review teams? They should be motivated by money to finance such research (why would most contract review companies want this investigated?). The research would, of course, have to be done by *bona fide* third-party scientists in a peer review setting. We don't want the profit motive messing with the truth and objective science.

Secret of Sampling

There is one more fundamental thing you need to understand about the TREC tests, indeed all scientific tests, one which I suppose you could also call a secret since so few people seem to know it, and that is, *no one*, I repeat, no person, ever sat down and looked at all of the 685,592 documents under consideration in 2010 TREC Legal Track interactive tasks. No one has ever looked at *all* of the documents in *any* TREC task. No person, much less a team of subject matter experts with three-pass reviews as I discussed in Part One, has determined the individual relevancy, or not, of all of these documents by which to judge the results of the software assisted reviews. *All* that happened (and I don't mean that as a negative connotation), is that a random sample of the 685,592 documents were reviewed by a variety of people.

I have no trouble with sampling and do not think it really matters that only a random sample of the 685,592 corpus was reviewed. Sampling and math are the most powerful tools in every information scientist's pocket. It seems like magic (much like the *hash* algorithms), but random sampling has been proven time and again to be reliable. For instance, a sample of 2,345 documents is needed to know the contents of 100,000, with a 95% confidence level and a +/-2 % confidence interval. Yet for a collection of 1,000,000 with the same confidence levels, a sample of only 2,395 is required (just 50 more to sample 900,00 more documents). If you add another zero and seek to know about 10,000,000 documents, you need only sample 2,400.

To play with the metrics yourself I suggest you see the calculator at http://www.surveysystem.com/sscalc.htm. For a good explanation of sampling see: *Application of Simple Random Sampling (SRS) in eDiscovery,* Manuscript By Doug Stewart, submitted to the Organizing Committee of the Fourth DESI Workshop on Setting Standards for Electronically Stored Information in Discovery Proceedings on April 20, 2011. Sampling is important. As I have been saying for over two years now, all e-discovery software should include a sampling button as a basic feature. (Many vendors have taken my advice, and I keep asking some of them to whom I made specific demands, to now call the new feature the *Ralph Button*, but they just laugh. Oh well:)

If the Human Review is Unreliable, Then so is the *Gold Standard*

The problem with average human review and the comparative measurements of computer assisted alternatives is not with the sampling techniques used to measure. The problem is that if

© Ralph Losey 2012

the sample set created by average Joe or Jane reviwer is flawed, then so is the projection. Sampling has the same weakness as AI agent software, including predictive coding seed sets. If the seeds selected are bad, then the trees they grow will be bad too. They won't look at all like what you wanted and the errors will magnify as the trees grow. It is the same old problem of *garbage in, garbage out*. I addressed this in Part One on this article, in the section, *The Second Search Secret (Known Only to a Few): The Gold Standard to Measure Review is Really Made Out of Lead*, but it bears repetition. It is a critical point that has been swept under the carpet until now.

Like it or not, aside from a few top reviewers working with relatively small sets, like the champs in TREC, most human review of relevancy in *large-scale reviews* is basically garbage, unless it is *very* carefully managed and constantly safeguarded by statistical sampling and other procedures. Also, if there is no clear definition of relevance, or if relevance is a constantly moving target, or both as is often the case, then the reviewers work will be poor (inconsistent), no matter what methods you use. Note this clear understanding of relevance is often missing in real world reviews for a variety of reasons, including the requesting party's refusal to clarify under *mistaken notions* of work product protection, vigorous advocacy, and the like.

Even in TREC, where they claim to have clear relevancy definitions and the review sets were not that large, I'm told by Webber that:

> TREC assessors disagree with themselves between 15% to 19% of the times when shown the same document twice (due to undetected duplication in the corpus).

That's right, the same reviewers looking at the same document at different times disagreed with *themselves* between 15% to 19% of the time. For authority Webber refers to: Scholer et al., *Quantifying Test Collection Quality Based on the Consistency of Relevance Judgements*. As you start adding multiple reviewers to a project the disagreement rates naturally get much higher. That is in accord with most everyone's experience and the scientific tests. If people cannot agree with themselves on questions of relevance, how can you expect them to agree with others? Despite a few champs, human relevancy review is generally very fuzzy.

Some Things Can Still Be Seen Through the Fuzzy Lenses

The exception to the fuzzy measurements problem, which I noted in Part One, is that the measures are not too vague for purposes of comparison, at least that is what the scientists tell me. Also, and this is very important, when you add the utility measures of time and money to review evaluation, which in the real world of litigation we must do, but has not yet been done in scientific testing, and do not *just rely* on the abstract measures of precision and recall, then *computer assisted* review must always win, at least in large-scale projects. We never have the time and money to manually review hundreds of thousands, or millions, of documents, just because they are in the custody of a person of interest. I don't care what kind of cheap, poor quality labor you use. As Jason Baron likes to point out, at a fast review speed of 100 files per hr, and a cost of $50 per hour for a reviewer, it would still take $500 Million and 10 Million hours to review the 1 Billion emails in the White House.

When you consider the utility measures of time and cost, it is obvious that pure manual review is dead. Even our weak, fuzzy comparative testing lens shows that shows manual and computer review precision and recall are about equal, and maybe the computer is even leading (hard to tell with these fuzzy lenses on). But when you add the time and costs measures, the race is not even close. Computers are far faster and should also be much cheaper. The need for computer assisted review to cull down the corpus, and then assist in the coding, is painfully obvious. The EDI study of a $14 Million review project by all too human contract coders with an overlap rate of only 28% proved that. Roitblat, Kershaw, and Oot, *Document categorization in legal electronic discovery: computer classification vs. manual review*. Journal of the American Society for Information Science and Technology, 61(1):70–80, 2010.

© Ralph Losey 2012

Going for the Gold

The old gold standard of average human reviewers, working in dungeons <smile>, unassisted by smart technology, and not properly managed, has been exposed as a *fraud*. What else do you call a 28% overlap rate? We must now develop a new gold standard, a new best practice for big data review. And we must do so with the help and guidance of science and testing. The exact contours of the new gold are now under development in dozens of law firms, private companies, and universities around the world. Although we do not know all of the details, we know it will involve:

1. *Bottom Line Driven Proportional Review* where the projected costs of review are estimated at the beginning of a project (more on this in a future blog);

2. High quality tech assisted review, with predictive coding type software, and multiple expert review of key seed-set training documents using both subject matter experts (attorneys) and AI experts (technologists);

3. Direct supervision and feedback by the responsible lawyer(s) (merits counsel) signing under 26(g);

4. Extensive quality control methods, including training and more training, sampling, positive feedback loops, clever batching, and sometimes, quick reassignment or firing of reviewers who are not working well on the project;

5. Experienced, well motivated human reviewers who know and like the AI agents (software tools) they work with;

6. New tools and psychological techniques (e.g. game theory, story telling) to facilitate prolonged concentration (beyond just coffee, $, and fear) to keep attorney reviewers engaged and motivated to perform the complex legal judgment tasks required to correctly review thousands of usually boring documents for days on end (voyeurism will only take you so far);

7. Highly skilled project managers who know and understand their team, both human and computer, and the new tools and techniques under development to help coach the team;

8. Strategic cooperation between opposing counsel with adequate disclosures to build trust and mutually acceptable relevancy standards; and,

9. Final, last-chance review of a production set before going out the door by spot checking, judgmental sampling (i.e. search for those attorney domains one more time), and random sampling.

I have probably missed a few key factors. This is a group effort and I cannot talk to everyone, nor read all of the literature. If you think I have missed something key here, please let me know. Of course we also need understanding clients who demand competence, and judges willing to get involved when needed to rein in intransigent non-cooperators and to enforce fair proportionality. Also, you should always go for confidentiality and clawback agreements and orders.

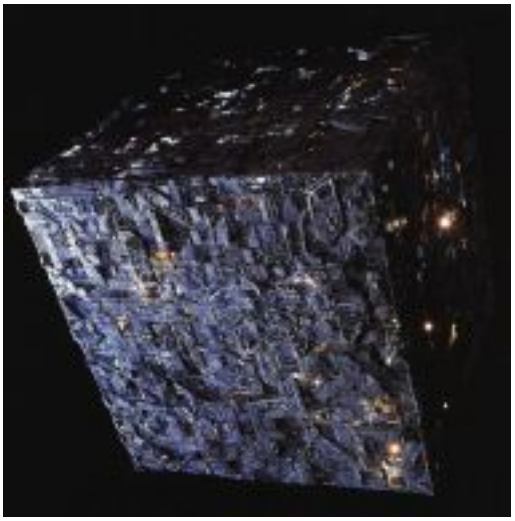
Technology Assisted Review

When I say *technology assisted review* in the best practices list above, which is now a popular phrase, I mean the same thing as *computer assisted review*. I mean a review method where computerized processes are used to cull down the corpus, and then again to assist in the coding. In the first step technology is used to cull out final selections of documents from a larger corpus

© Ralph Losey 2012

for humans to review before final production. The probable *irrelevant* documents are culled-out and not subject to any further human reviews, except perhaps for quality control random sampling. Keyword search is one very primitive example of that computer assisted culling. Concept search is another more recent, advanced example. There are many others. Think for instance of Axcellerate's 40 *automatically populated filters*, which they collectively refer to as their Predictive Analytics™ step that I described in Part One of *Secrets of Search*.

These days the software is so smart that technology assisted review can not only intelligently cull out likely irrelevant documents, it can also make predictions for how the remaining relevant documents should be categorized. That is the second step where all of the remaining documents are reviewed by software to predict key classifications like privileged, confidential, hot, and maybe even a few case specific issues. The software predicts how a human will likely code a documents and batches documents out in groups accordingly. This predictive coding, combined with efficient document batching (putting into sets of documents for human review), makes the human review work easier and more efficient. For instance, one reviewer, or small review team, might be assigned all of the *probable privileged* documents, another the *probable confidential* for redaction, a third the *probable hot* documents, and the remaining documents divided into teams by case issue tags, or maybe by date, or custodian, all depending on the specifics of the case. It is an art, but one that can and should be measured and guided by science.



I contrast this kind of *technology assisted review* with pure Borg type *computer controlled* review, where there is complete computer delegation, where the computer does all, with little or no human involvement, except for the first seed set generation of relevancy patterns. Here we trust the AI agent and produce all documents determined to be relevant and not-privileged. No human does a double-check of the computer's coding before the documents go out the door. In my opinion, we are still far away from such total delegation, although I don't rule it out someday. (Resistance is futile.) Do you agree?

Is anyone out there relying on 100% computer review with no human eye quality controls? Conversely, as to the opposite, is there anyone out there who still uses pure (100%) human review? Who has humans (lawyers or paralegals) review *all* documents in a custodian collection (assuming, as you should, that there are thousands or tens of thousands of documents in the collection)? Is there anyone who does not rely on some little brother of Watson to review and cull out at least some of the corpus first?

More Research Please

The fuzzy standard of most human review is an inconvenient truth known to all information scientists. As we have seen, it has been known to TREC researchers since at least 2000 with the study by Ellen Voorhees. *Variations in relevance judgments and the measurement of retrieval effectiveness*, 36:5 Information Processing & Management 697, 701 (2000). Yet I for one have not heard much discussion about it. This flaw cuts to the core of information science, because *without accurate, objective measurements, there can be no science*. For that reason scientists have come up with many techniques to try to overcome the inherent fuzziness of relevancy determinations, in and outside of legal search. I concede they are making progress, and TREC legal track is, for instance, getting better every year, but, like Voorhees and Webber, I insist there is still a long way to go.

© Ralph Losey 2012

Maybe the best software programs (whatever they are) are far better than our best reviewers under ideal conditions (that's what I think), maybe not. But the truth is, we don't really know what our real precision and recall rates are now, we don't really know how much of the truth we are finding. The measures are, after all, so vague, so human dependent. What are we to make of our situation in legal review where the Roitblat *et al* study shows an overlap rate of only 28%? Here is Webber's more precise information science language explanation that he made in reviewing my blog article in his blog:

> The most interesting part of Ralph's post, and the most provocative, both for practitioners and for researchers, arises from his reflections on the low levels of assessor agreement, at TREC and elsewhere, surveyed in the background section of my SIRE paper. Overlap (measured as the Jaccard coefficient; that is, size of intersection divided by size of union) between relevant sets of assessors is typically found to be around 0.5, and in some (notably, legal) cases can be as low as 0.28. If one assessor were taken as the gold standard, and the effectiveness of the other evaluated against it, then these overlaps would set an upper limit on F1 score (harmonic mean of precision and recall) of 0.66 and 0.44, respectively. Ralph then provocatively asks, if this is the ground truth on which we are basing our measures of effectiveness, whether in research or in quality assurance and validation of actual productions, then how meaningful are the figures we report? At the most, we need to normalize reported effectiveness scores to account for natural disagreement between human assessors (something which can hardly be done without task-specific experimentation, since it varies so greatly between tasks). But if our upper bound F1 is 0.66, then what are we to make of rules-of-thumb such as "75% recall is the threshold for an acceptable production"?

As Webber well knows, this means that such 75% or higher rules-of-thumb for acceptable recall are just wishful thinking. It means they should be disregarded because they are counter to the actual evidence of measurement deficiencies. The evidence instead shows that the maximum possible mean precision and recall rate measured objectively is only 44%. Demands in litigation for objective search recall rates higher than 44% fly in the face of the EDI study. It is an unreasonable request on its face, never mind the legal precedent for accepting keyword search or manual review. I understand that the research also shows that technology assisted reviews are at least as good as manual, but that begs the real question as to how good either of them are!

I personally find it hard to believe that with today's technology assisted reviews we are not in fact doing much better than 44% or 65% recall, but then I think back to the lawyers in the 1980s in the Blair Moran study: *We are confident our search terms uncovered 75% of the relevant evidence.* Well, who knows, maybe they did, but the measurements were wrong. Who knows how well any of us are doing in big data reviews? The fuzziness of the measures is an inconvenient truth that must be faced. The 44% max objective rate creates a *lack of confidence interval* that must be corrected. We have to significantly improve the gold standard, we have to upgrade the quality of reviews used for measurements.

This is one reason I call for more research, and better funded research. We need to know how much of the truth we are finding, we need a recall rate we can count on to do justice. Large corporations should especially step up to the plate and fund pure scientific research, not just product development. I trust you that it works, but, as President Regan said, I still want you to verify. I still want you to show me exactly how well it works, and I want you to do it with objective, peer-reviewed science, and to use a gold standard that I can trust.

Trust But Verify

As it now stands, the confidence rates and error margins are too low for me to entirely trust Watson, much less his little brothers. The computer was, after all, trained by humans, and they can be unreliable. Garbage in, garbage out. I will only trust a computer trained by several humans, checking against each other, and all of them experts, well paid experts at that. Even

© Ralph Losey 2012

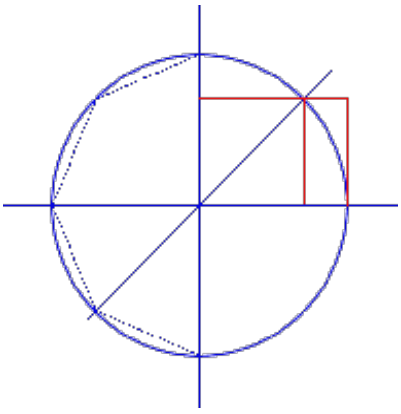
then, I'd like to have a final expert review of the documents finally selected for production before they actually go out the door. After all, the determinations and samples are based on *all too human* judgments. If the stakes are high, and they usually are in litigation, especially where privileges and confidential information are involved, there needs to be a final check before documents are produced. That is the true gold standard in my world.

Secrets of Search: Part Three

Now finally we come to the conclusion of this series on the *Secrets of Search* where all will be revealed. (Well, to be entirely honest, not *all* will be revealed. I'm still going to keep a few *trade secrets* up my sleeve for law partners and family.)

Recap of the First Three Secrets

Before I get to the fourth secret of search, I need to review the first three again and connect a few more dots. The first secret was already known to many. (Craig Ball said it was about as much of a *secret* as the square root of 256.) It was that keyword search, done alone, and as part of a blind *Go Fish* game of dueling attorneys, is *remarkably ineffective*. Keyword search only works when performed as part of an interactive, multi-modal process, one which uses constant sampling and review. Still, keyword search is yesterday's (1960s) technology. No matter how many Boolean bells and whistles and interactivity and quality controls you may add to keyword search, its only real strengths are familiarity and quick peeks. The future of legal search, the best promise for adequate precision and recall, lies in artificial intelligence software. By this I mean the called predictive coding algorithms where expert humans train computer agents, plus ever improving legal methods.



The second secret really *was* a secret, kind of like knowledge of the square root of two was in ancient Greece. This secret was little known outside of information science circles, who, in speech at least, tend to emulate the Pythagoreans in enigmaticalities. This secret is that the *gold standard* used to test precision and recall is, like keyword search, *remarkably ineffective*. That so-called *gold standard* is human review. This is a very imprecise, very fuzzy standard. The few studies we have on big data projects, ones where humans reviewed thousands of documents for days on end, reveal terribly inconsistent relevancy calls. (Not surprising when you consider how bleary-eyed and underpaid they were.) For instance, in the $14 Million Verizon project, human reviewers only agreed 28% of the time. This means that our yardstick for recall measurements has nothing smaller on it than a foot. All claims for precision within a few inches are bull. We really have no way of knowing that.

As information scientist William Webber notes, our maximum possible mean precision and recall rate ("***F1"***) measured objectively is only 44%, and other studies suggest an only slightly higher *F1* rate of 66%. This is very significant because it means there is no objective basis to ever demand a recall rate of better than 66%. A requesting party that asks for recall better than that is asking for something that cannot be reliably measured.

© Ralph Losey 2012

Logically, this also means random samples with 95% confidence levels +/- 2 are also unrealistically high. Plus or minus 5 might be more realistic considering the vagaries of our measurements and subjective determinations. I favor random sample buttons on software, but I want our use of them to be realistic and not budget busting. What is the point of such accuracy when the underlying data is so fuzzy? The demands of 99% confidence level, or plus or minus one confidence interval, are completely misplaced and illogical. Our measuring stick is too imprecise to justify such large sample sizes. The *experts* who ask for that kind of delusional certainty have not understood the second secret. Either that, or they are just trying to drive up the costs of the other side's quality control efforts.

Still, sampling is a powerful tool if used right, and if you understand what it can, and cannot do. For instance, it cannot by itself improve accuracy of search at all. It is just a tool to get an idea of how you are doing in your search processes. Since I am a strong proponent and have been urging all software providers to add a random sample generators to their programs for years, I decided to practice what I preach and figured out a way to add one on my blog. It can now always be found on the blog sidebar on the right, identified as a **Math Tool for Quality Control**.

The third secret is that even though humans are terrible at large-scale reviews, it is a completely different story when dealing with small-scale reviews. When reviewing small sets of data, in the 500-1,000 document range (this is the number of documents reviewed by the individual TREC reviewers), there were several professional reviewers in TREC who were ***more precise*** and had ***better recall*** than the best computer systems, even though they were not subject matter experts and had no access to such experts. Even a couple of the law students won a few times. Webber's analysis showed that the complete demise of human reviewers has been grossly exaggerated. *Re-examining the Effectiveness of Manual Review*.

Although pure manual review is good for a few hours, it is poor and inaccurate over large scales, as the second secret revealed. Even if it were not, manual review is far too expensive and slow for large-scale review projects. We cannot go it alone. We need the machines. But we also need to keep the arts alive, the special skills of persuasion and evidence evaluation that we lawyers have refined over centuries. (More on that in the fifth secret at the end of this blog.)

Requesters who demand production with only machine review, and any responders foolish enough to comply, have not understood the third secret. It is *way* too risky to turn it *all* over to the machines. They are not that good! The reports of their excellence have been grossly over-stated. Humans, there is need for you yet. The *Borg* be damned! Jobs may have passed away, but his work continues. Technology is here to empower art, not replace it.

Webber's research, and the common experience of our best law firms and vendor review teams nationwide, suggest that a hybrid multi-modal combination of both manual and machine review is the best approach. The new emerging gold standard uses the talents of both and a variety of automated tools. It also uses extensive interactivity between humans, and between humans and machines. In *Part Two of Secrets of Search* I suggested nine characteristics of what I hope may become an accepted best practice for legal review worldwide. I invited peer review and comments on what I may have left out, or any challenges to what I put in, but so far this list of nine remains unchallenged:

1. *Bottom Line Driven Proportional Review* where the projected costs of review are estimated at the beginning of a project (more on this in the next blog);

2. High quality tech assisted review, with predictive coding type software, and multiple expert review of key seed-set training documents using both subject matter experts (attorneys) and AI experts (technologists);

© Ralph Losey 2012

3. Direct supervision and feedback by the responsible lawyer(s) (merits counsel) signing under 26(g);

4. Extensive quality control methods, including training and more training, sampling, positive feedback loops, clever batching, and sometimes, quick reassignment or firing of reviewers who are not working well on the project;

5. Experienced, well motivated human reviewers who know and like the AI agents (software tools) they work with;

6. New tools and psychological techniques (e.g. game theory, story telling) to facilitate prolonged concentration (beyond just coffee, $, and fear) to keep attorney reviewers engaged and motivated to perform the complex legal judgment tasks required to correctly review thousands of usually boring documents for days on end (voyeurism will only take you so far);

7. Highly skilled project managers who know and understand their team, both human and computer, and the new tools and techniques under development to help coach the team;

8. Strategic cooperation between opposing counsel with adequate disclosures to build trust and mutually acceptable relevancy standards; and,

9. Final, last-chance review of a production set before going out the door by spot checking, judgmental sampling (i.e. search for those attorney domains one more time), and random sampling.

I have probably missed a few key factors. This is a group effort and I cannot talk to everyone, nor read all of the literature. If you think I have missed something key here, please let me know. You can reach me at ralph.losey@gmail.com.

You may note that I am herewith joining the call of other leaders in the field to develop best practice standards, notably including Jason Baron, and have overcome my initial reluctance to go there for a variety of reasons. *See* Jason R. Baron, *Law in the Age of Exabytes: Some Further Thoughts on 'Information Inflation' and Current Issues in E-Discovery Search*, XVII RICH. J.L. & TECH. 9, at 29-33. My concerns on arbitrary standards and unfounded malpractice claims remain, but I think we have no choice but to develop some basic industry standards. The nine characteristics of good document review outlined above constitute a first modest step in that direction.

The Fourth Secret of Search:
Relevant Is Irrelevant

Sorry to sound like one of Steve Jobs' Zen Masters, but a contradiction like *Relevant Is Irrelevant* has more impact than the technically more accurate statement, which is: *merely relevant* documents in big data reviews are irrelevant as compared to *highly relevant* documents. In other words, all that counts in litigation are the *hot* documents, the highly relevant ones with strong probative value, not the documents which are just relevant, not to mention just responsive. In fact, in big data collections, I could care less about merely relevant documents. Their only purpose is to lead me to highly relevant documents. Moreover, as we will see in the fifth and final secret, I only care about a handful of those.

In a case involving tens of thousands of documents, much less hundreds of thousands of documents, or millions of documents,

© Ralph Losey 2012

almost all of the documents that are merely relevant will *not* be admissible into evidence. (I'll explain why in a minute.) For that reason alone their discovery should be subject to very close scrutiny. The gathering of evidence for admission at trial is, after all, the only valid purpose of discovery. Discovery is never an end in itself, although many litigators (as opposed to true trial lawyers) and vendors often lose that track of that basic truth. Discovery is only permitted for purposes of preparation for trial. It is never permitted to extort one side into a settlement to avoid the costs of a document review, or to at least gain a strategic edge, although we all know this happens all of the time.

Why won't most *merely relevant* evidence be admissible as evidence you may wonder? For the same reason that most of the even *highly relevant* evidence won't be admissible. Even though relevant, this evidence is a **cumulative waste of time**, and for that reason is inadmissible under Rule 403 of the *Federal Evidence Code* and its state law equivalents. To refresh your memory on the Evidence Code:

> Rule 403. Excluding Relevant Evidence for Prejudice, Confusion, Waste of Time, or Other Reasons.
>
> The court may exclude relevant evidence if its probative value is substantially outweighed by a danger of one or more of the following: unfair prejudice, confusing the issues, misleading the jury, undue delay, wasting time, or needlessly presenting cumulative evidence.

Also see Rule 611. ("The court should exercise reasonable control over … presenting evidence so as to … (2) avoid wasting time")

The typical fact scenario used in law school to exemplify the principle of cumulative evidence is a situation where 100 witnesses see the same accident. Each would each give roughly the same description of the event and the testimony of each would be equally relevant. Still the testimony of 100 witnesses would never be allowed because it would be a waste of time, and/or a needless presentation of cumulative evidence, to have all 100 repeat the same facts at trial. The same principle applies to documentary evidence. If there are 100 emails that show essentially the same relevant fact, you cannot admit all 100 of them. That would be a cumulative waste of time.

The question of admissibility presented in Federal Rule of Evidence 403 requires a balancing of the costs and benefits of logically relevant evidence. This is sometimes referred to as the Rule 403 balancing test. This is similar to the balancing tests in Rule 26(b)(2)(C)(i) and (iii) of the *Federal Rules of Civil Procedure* between the benefits and burdens of discovery.

> 26(b)(2)(C) The frequency or extent of use of the discovery methods otherwise permitted under these rules and by any local rule shall be limited by the court if it determines that:
>
> (i) the discovery sought is unreasonably cumulative or duplicative, or is obtainable from some other source that is more convenient, less burdensome, or less expensive; … or
>
> (iii) the burden or expense of the proposed discovery outweighs its likely benefit, taking into account the needs of the case, the amount in controversy, the parties' resources, the importance of the issues at stake in the litigation, and the importance of the proposed discovery in resolving the issues.

New e-discovery Rule 26(b)(2)(B) has a similar balancing test for hard-to-access ESI. So too does Rule 26(g) that requires only a reasonable inquiry of completeness in a response to discovery. Perhaps more importantly, Rule 26(g)(1)(B) also prohibits any request for discovery made "for any improper purpose, such as to harass, cause unnecessary delay, or needlessly increase the cost of litigation" and prohibits any request that is unreasonable or unduly

© Ralph Losey 2012

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." All the rules point to reasonability in discovery, and yet in e-discovery we routinely engage in unreasonable, cumulative overkill. *See* Patrick Oot, Anne Kershaw and Herbert L. Roitblat, *Mandating Reasonableness in a Reasonable Inquiry*, Denver University Law Review, 87:2, 522-559, at 537-538 (2010).

The rules clearly state that cumulative evidence is not, or at least *should not*, be subject to discovery. It would be a waste of time and money. Thus even though the documents might be relevant, if they are unreasonably cumulative, repetitive, or duplicative, such that the burden outweighs the benefit, they are not only inadmissible as evidence, but they are, or *should be*, outside of discovery.

This is buttressed by the *prime directive* of the *Federal Rules of Civil Procedure*, Rule 1. It requires all of the other rules of procedure to be interpreted and applied so as to make litigation *just, speedy and inexpensive*.

In spite of the clear law against cumulative, over burdensome discovery, lawyers and judges faced with big data cases today still routinely engage in discovery overkill. A 2010 survey of large cases that went to trial in 2008 showed that on average, 4,980,441 pages of documents were produced in discovery, but only 4,772 exhibit pages were entered into evidence. Duke Litigation Cost Survey of Major Companies (2010) at pg. 3. That is a ratio of over one thousand to one! *Also see DCG Sys., Inc. v. Checkpoint Techs.*, LLC, No. C-11-03792 PSG, 2011 WL 5244356 (N.D. Cal. Nov. 2, 2011) (little benefit to justify burden of large scale email production because on average only ".0074% of the documents produced actually made their way onto the trial exhibit list" and in appeals "email appears more rarely as relevant evidence").

These are absurd numbers for a variety of reasons. The 4,772 admitted into evidence is ridiculous over-kill, as will be shown further in the fifth secret, and so is the number of documents produced. The producing parties, acting in concert and cooperation with the requesting parties, should do a better job of culling down the irrelevant documents and marginally relevant documents. They are not needed for trial preparation.

This so-called *Duke Survey,* which was commissioned by the *Lawyers for Civil Justice*, not Duke, also offered opinion convergent with my own that such discovery is excessive (although we disagree on causation):

> Whatever marginal utility may exist in undertaking such broad discovery pales in light of the costs. … Reform is clearly needed. A discovery system that requires the production of a field full of "haystacks" of information merely on the hope that the proverbial "needle" might exist and without any requirement for any showing that it actually does exist, creates a suffocating burden on the producing party. Despite this, courts almost never allocate costs to equalize the burden of discovery.

The Fifth Secret of Search: 7±2
Should Control All e-Discovery (But Doesn't)

We have already established that the purpose of discovery is to prepare for trial. But what is the purpose of a trial? We have to understand that to be able to grasp the fifth secret: **7±2**. We have to understand that the purpose of all trials is to *persuade*. It is a time and place, a level playing field, where lawyers try to persuade a judge and/or jury as to what happened and what should be done about it.

In this place of trial of humans, by humans, the rule of **7 ± 2** reigns supreme. It always has and, unless we allow robots as jurors, always will. Unfortunately, most litigators are unaware of this

rule of the transmission of information, or if they did know of it, most fail to see its connection to discovery and search. The rule of **7±2** now has little place in e-discovery analysis.

It is a secret, and because it is unknown, we have gone astray in e-discovery. Because this secret is unknown vast sums of money are routinely wasted in *the production of fields full of "haystacks" of information.* Because the secret has not yet been heard, and its clear implications have not been yet been understood, trial lawyers everywhere still scratch their head in disbelief at the sheer mention of *e-discovery*. Yes, this secret is also the key to the *seventh insight*. The insights into wide-spread lawyer resistance to e-discovery analyzed in *Tell Me Why?*

I have alluded to this *rule of seven* in a few past blogs, and discussed it at a few late night dinners. But this is the first time I have written at length on the *magic power of seven, plus or minus two.* I hesitate to go to this deep place of information transmission and cognitive limitations, but, in order to keep the search for truth and justice on track, we really have no choice. We must, like the Pythagoreans of old, consider the significance of the number seven and its impact on our work, especially on our conceptions of proportionality.

The fifth secret of search is based on the legal art of persuasion and the limitations of information transmission. **The truth is, no jury can possibly hold more than five to nine documents in their head at a time.**

It is a waste of time to build a jury case around more documents than that. Judges who are trained in the law, and are quite comfortable with documents, can do a little better, but not that much. In a bench trial you might be able to use eight to twelve documents to persuade the skilled judge. But even then, you may be pushing your luck. Judges, after all, have a lot on their mind, and your particular case is just one among hundreds (in state court make that thousands).

Computers Expand Document Counts, Not Minds

Even though the computerization of society has exploded the number of documents we retain a trillion-fold, the ability of the human mind to remember and process has remained the same. We still can only be persuaded by a handful of writings. That is all of the information we can retain. Presenting dozens of documents is a waste of time. The only reason to present more that five to nine documents at trial is to provide context and an evidentiary foundation. The few dozen other documents that you may need at trial are merely window dressing, a frame for the real art.

A computer can easily process and recall millions of documents, and can do so in minutes, but we cannot. Even fast readers are limited to about 500 words per minute or a skim-review rate of 1,000. No matter how much time we may have, and in legal proceedings the time is always constrained, our ability to read, understand, and comprehend relevant writings is limited. This is especially true in the high pressure and expedited schedule of a trial and formal presentation of evidence in court. That is why all experienced trail lawyers I have talked to agree that the average juror is likely to remember and be influenced by only a handful of documents. By the way this rule of seven in persuasion is a corollary to the *KISS* principle ("keep it simple, stupid"), well known to all persuaders, along with "tell-tell-and-tell."

Although most trial lawyers learn this just from hard experience, there is good theoretical support in psychology for such memory limitations. It is sometimes called *Miller's Law*, after cognitive psychologist George A. Miller, a professor at Princeton University. Professor Miller first described this limitation of human cognition in his 1956 article: *The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information*, Psychological Review 63 (2): 81–97. This is supposedly the most widely quoted psychology paper of all time. According to Wikipedia, Miller's paper suggests that seven (plus or minus two) is the magic number that characterizes people's memory performance on random lists of letters,

© Ralph Losey 2012

words, numbers, or almost any kind of meaningful familiar item. He essentially found that human beings were only capable of *receiving, processing and remembering* seven (plus or minus two) variables at any one time.

Professor Miller's ends his famous paper on the limits of our capacity to process information with this somewhat odd remark, especially considering his reputation as a scientist:

> What about the magical number seven? What about the seven wonders of the world, the seven seas, the seven deadly sins, the seven daughters of Atlas in the Pleiades, the seven ages of man, the seven levels of hell, the seven primary colors, the seven notes of the musical scale, and the seven days of the week? What about the seven-point rating scale, the seven categories for absolute judgment, the seven objects in the span of attention, and the seven digits in the span of immediate memory? For the present I propose to withhold judgment. Perhaps there is something deep and profound behind all these sevens, something just calling out for us to discover it. But I suspect that it is only a pernicious, Pythagorean coincidence.

George A. Miller, *The Magical Number Seven, Plus or Minus Two* (1956), 42-3.

Apparently some psychologists think Professor Miller overestimated the average human capacity when he said it was between 5-9. They think the limit is more likely to be from two to six, that the magic number is 4, not 7. Farrington, Jeanne, *Seven plus or minus two*, Performance Improvement Quarterly 23 (4): 113–6. doi:10.1002/piq.20099 (2011).

In any event, it is not hundreds of documents, much less thousands or millions. Yet in an average large case today 4,980,441 pages of documents are produced and 4,772 pages allowed into evidence. What is wrong with this picture? The discovery chase has lost track of the goal.

An experienced trial lawyer, who may use hundreds of exhibits in a very large trial for context and technical reasons, will still only focus on five to nine documents. They know jurors cannot handle more information than that. They know the rest of the documents that go into evidence will have little or no real persuasive value.

The limitations of the human mind thus provide a consistency and continuity with the trials and systems of justice of our past pre-computer civilizations. No matter how many more documents may exist today within the technical scope of legal relevance, our jurors' capacities are the same; the art of legal persuasion remains the same.

© Ralph Losey 2012

These mental persuasion limits provide a governor on the number of documents useful to a trial lawyer, judge, and jury. The human mind has its limits. Computer discovery must start to realize these limits and take them into consideration. This is a basic truth that we *e-discoverers* have lost sight of.
It is the core of why most old-time trial lawyers think the whole business of e-discovery is ridiculous. It is high time for the *secret of seven* to be outed and, more importantly, to be followed. The rule of seven should have significant consequences on our legal practice and scientific research.

Uneducated Searchers Will Never Find the Top 7±2

The location of these few highly relevant documents has always been a problem in the law. But in the low volume paper world it was never an overwhelming one. The paper document search and retrieval process was a relatively simple problem traditionally assigned to the youngest, most inexperienced lawyers. Today the search for the *smoking e-guns* is much more difficult than ever before, yet untrained young associates are still commonly given this task. Many are simply told to go *do e-discovery.* They are provided with little more training than attendance of a few CLEs, which, you should know by now, don't really teach you that much.

That is one compelling reason I took the time to make my law school training program available online to attorneys, paralegals, techs and students everywhere. e-DiscoveryTeamTraining.com. It provides over 75 hours of instruction, which is what it takes to really learn something. Just don't try to learn more than seven things at a time. Take your time and study online whenever it is convenient to you.

Lack of real education is the primary impediment to further progress in all e-discovery issues, including search. Patrick Oot, Anne Kershaw, and Herbert Roitblat explained it well in their excellent *Mandating Reasonableness* article:

> The problem is not technology; it is attorneys' lack of education and the judicial system's inattentiveness to ensure that attorneys have the proper education and training necessary for a proportional and efficient discovery process. Lack of attorney education aggravates the problem because uneducated litigators are unable to make informed judgments as to where to draw the line on discovery, thereby creating unrealistic expectations from the courts—particularly as to costs and burdens. For example, failing to understand how different methods of search methodology work, some judges will unnecessarily mandate traditional and expensive "brute force" attorney review. …
>
> Simply put, the legal system has a crisis of education. Both attorneys and judges need to better understand technology as it applies to the reasonable inquiry.

Mandating Reasonableness, supra at pg. 545, 547.

Just Give Me the *Smoking Guns*

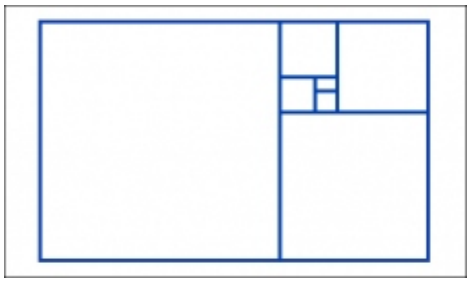
Since only a few documents are needed for analysis of a case and even less for persuasion at trial, the search of paper-only has sufficed, until recently, for most trial lawyers. They have found the few they needed in printouts. But these days are now all but gone. The few important documents found by paper searches, and even by ESI searches that are driven by old paper based systems, are not likely to uncover the best documents. The smoking guns will remain hidden in the data deluge. Lawyers will not find the top

© Ralph Losey 2012

seven needed for the judge and jury.

As the nature of documents changes, and the previously noted habits of witnesses to print key documents disappears, this problem will worsen. No one today says incriminating things in paper letters. Very few still even write paper letters. They say it in emails, text messages, instant messages, Facebook posts, blogs, tweets, etc., and almost no one prints these out and puts them in filing cabinets.

There is a key lesson for e-discovery in the trial lawyer wisdom of seven. To be useful discovery must drastically cull down from the millions of ESI files that may be relevant, to the few hundred that are useful, and the five or nine really needed for persuasion. Culling down from millions to only tens of thousands is not serving the needs of the law. It is a pointless waste of resources, a waste of client money. A production of tens of thousands of documents, not to mention hundreds of thousands, is unjust, slow and inefficient.
Many vendors today brag about how their smart culling was able to eliminate up to 80% of the corpus. They will tell you this is an excellent cull rate before you begin review. It is not. They may also tell you that it is unreasonable for you to try to cull out more than that. They are wrong. They have a financial motivation to take such conservative positions. The more documents you review, the more money they make. Some law firms see it that way too. But they won't last, the firm's clients will eventually catch on and switch their work away from the haystack builders.



Even if well-intentioned, many vendors (and lawyers) don't understand that the law requires only reasonable efforts, and proportional efforts, not perfect or exhaustive efforts. They don't understand the basic limitations of a trial or cumulative evidence. Many have never even seen a trial, much less tried one. Vendors are not supposed to give legal advice, yet I hear them do it all of the time when, for instance, they talk about how much you *should* review to meet your obligations under the law. Or they may say it would be *very risky* to try to cull out more than that. As if they could ever really eliminate risk, much less quantify risk. The only way to eliminate risk is by cooperation or court order. Not by following vendor best practice suggestions.

When you understand the fourth and fifth search secrets, you realize that a cull rate of *at least* 90% is proportional. It does not matter if you weed out a few *merely relevant* documents. If you have a million files, you should be able to weed out at least 90%, 900,000 documents, before you begin review. In fact, you should aim for elimination of 98%+ by using relevancy ranking, and only do a human hybrid review of the remaining 20,000 documents.

New e-discovery search and culling methods need to be perfected that limit the quantity of documents to a size that the human mind can deal with and comprehend. The processes should try to find all, or nearly all, of the *highly relevant* documents, even if a significant percentage of marginally relevant documents are missed. Who cares about these technically relevant documents? No one, except maybe those dazzled by recall stats who do not understand the natural speed limits of the mind. All that really matters are the hot documents. That is the lesson of the fourth secret of search, that *Relevant Is Irrelevant*.

The lesson of the fifth secret, ***7±2***, is that the true goal of e-discovery should be the five to nine of the hot documents that the triers of fact can understand. If your search finds those magic seven, and no others, it is a great success, regardless of all of its other misses. If your search finds a million relevant documents, and attains a precision and recall rate of 99%, but misses the top seven key documents, it is a complete failure. We have to change our search methods to focus on the top seven.

© Ralph Losey 2012

Change the Scientific Testing

We also have to redesign our scientific testing to measure what really counts, the 7±2, plus time and money. I suggest that the TREC Legal Track have a seeded test set next year where all searchers look to find seven planted *Easter eggs*. Whoever finds them all, or finds the most, *and* does so the fastest, and at the least expense, gets the highest score. In fact, for the tests to be fair and realistic, they should be time limited, and cost limited. Participants should no longer be allowed to keep that secret. In the law time and money matter. A search process is worthless that costs too much, or takes too long.

So far, all of the scientific experiments I have heard about in e-discovery have measured effectiveness, meaning how well or poorly a search performs, by *only* looking at Relevance measures, primarily precision and recall (or the harmonic mean thereof – F1). But in information science, *Relevance* is just one of the four basic measures of search effectiveness. The other three are *Efficiency, Utility*, and *User Satisfaction*. Sándor Dominich, *The Modern Algebra of Information*, Pgs 87-88 (Springer-Verlag, 2008). According to Dominich, the Efficiency measures are the costs of search and the time it takes. We need to start to include Efficiency measures in our tests, as well as provide heavy *ranking* to our Relevance measures.

In Law One Key Document is Worth a Million Relevant Documents

Too few *experts* in e-discovery today understand the fifth secret of search, namely the magic limiting power of seven. On the other hand, all experienced trial lawyers seem to know it well, even if they have never heard of Professor Miller. As a result of **7±2** being such a secret to many of my friends in e-discovery, they have erroneously focused on an effort to recall as many relevant documents as possible. They pride themselves in amassing large volumes of relevant documents, when in fact that is the last thing real trial lawyers want. They don't want ten thousand relevant documents; they want ten. They want just a handful of killer documents that will help persuade the jury, that will make their story clear and convincing. The failure of e-discovery proponents to focus on this is another reason, the 7th in fact, why so many lawyers think e-discovery is *stupid.*

Electronic discovery search is not an academic game to be played. It is all about finding evidence for trial. Statistics and methods are worthless unless they properly weigh recall statistics by persuasive impact. One highly relevant document can, and usually does, counteract ten million relevant ones. It is like one grand master at Chess playing a thousand amateurs. The amateurs don't have a chance. Because of this if your search is not designed to find the five to nine most persuasive documents, then your search is flawed, no matter what your precision and recall rates are.

High recall rates are only imperative for *highly relevant* documents, the *hot* documents. Nothing else matters, except for the costs involved, the time and money it takes to find evidence. If you don't focus your search on 7±2 hottest documents, you may never find them.

I know that some will argue you have to find *all* of the relevant documents in order to be able to find the top 7±2. That was true in the paper world of linear review of hundreds of documents, but is not true in large-scale electronic review. You can now use software that focuses its search on the highly ranked relevant documents. But you hve to adopt your methods accordingly.

© Ralph Losey 2012

New methods for ESI review should be used that focus on retrieval of *ranked relevancy*, not just relevancy. The methods should focus on finding the hot documents with the understanding that merely responsive documents are, due to their extreme number, of little importance. Relevant is irrelevant. The same ranking applies to identification of privileged and confidential ESI. If one hot privileged document is missed in a privilege review, it can be far more damaging that the inadvertent production of hundreds of marginally privileged ones.

Bottom line, to follow the fourth and fifth secrets we have examined in this blog, the key feature you should look for in search software is the ability to accurately *rank* the probable relevant documents. Ranking must be a far more sophisticated function than simply counting the number of times a keyword, or pattern, appears in a document. It should epitomize all of the criteria and indices used by the software *black box* – latent semantic, four-dimensional geometric, or otherwise.

The ideal e-discovery Watson computer must not only search and find, he must rank. Put the highest on top please. Watson may not be able to put the five you will use as the first five documents shown, but it is not too much to expect that the 7±2 will be in the top 5,000. The humans working with Watson will narrow them down, and the trial lawyers making the pitch will make the final selections.

Recap of All Five Secrets

To recap, in Part I we discussed the first two secrets. The first is that keyword search sucks and so most attorneys still using this old method are searching for ESI the wrong way. The second secret is that large scale linear manual reviews also sucks and this means we do not have a reliable gold standard by which to make precision and recall measurements. We do, however, know that a hybrid approach of man and machine, using keyword, predictive coding and other automated methods, is at least as accurate as manual review and far faster and less expensive.

In Part II we discussed the third secret that in small scale reviews of 500-1,000 documents professional reviewers are still better than our best automated methods, and it is foolhardy to take human review out of the final computer proposed production set. We need human review not only to instruct the computer, but for quality control and confidentiality protection. We also discussed the parameters for a new gold standard of hybrid, multimodal search and review.

In this Part III we discussed the fourth secret that *relevant is irrelevant*, meaning that smart culling that follows best practices is required by the rules to keep the time and cost of review proportional. The fifth secret gleaned from our friends the trial lawyers, 7±2, reminds us of the true goal of e-discovery and the need to heavily weight and constrain our relevancy searches.

The following graphic summarizes these thoughts using the symbol of the Pythagoreans, the five-sided polygon, or pentagon, who were, by the way, famous among the ancient Greeks for secret keeping and a relentless search for truth.

© Ralph Losey 2012

As you have no doubt guessed by now, my real goal here was not to give away *secrets*, but to lay the foundation for new standards of search and review. The pentagon shows the first five steps, but there is still one more. In the next blog I will discuss that step and use the six-sided figure, a hexagon, to show my current understanding of best practices.

Conclusion

Way back in 1947 the Supreme Court in *Hickman v Taylor*, the landmark case on discovery, stated that "[m]utual knowledge of all the relevant facts gathered by both parties is essential to proper litigation." 329 U.S. 495, 507 (1947). The opinion was written by Justice Frank Murphy (1890-1949) shown right. Today his statement is obsolete in so far as it says ALL the relevant facts gathered should be shared. This statement was reasonable when written in 1947, but not today. In those days, the forties, all of the relevant facts could be found in a few dozen documents. In the sixties that became at most a few hundred. In the nineteen seventies and eighties, a few thousand.

Today, sixty-five years after *Hickman v. Taylor,* we now live in a completely different world. Today written words profligate and multiply with the help of computers in a way that our ancestors could never imagine. Now you can gather hundreds of thousands or millions of relevant documents in even small cases. Now we write all of the time, and our writings multiply and remain, albeit in electronic form only.

The sharing of marginally important knowledge is no longer *essential to proper litigation*. In fact, as we have seen, it is contrary to the rules, especially Rule 26, *Federal Rules of Civil Procedure*. Most merely relevant documents today are inadmissable. Rule 403, *Federal Rules of Evidence*. They are a cumulative waste of time. It is unreasonable to gather them, much less disclose them. Rule 1 prohibits such a waste of time and money. Moreover, it is unjust. For it is easy to bury the truth in mountains of technically relevant haystacks. Document dumps are a way to hide the truth *essential to proper litigation*.

We need to design our e-discovery to be reasonably calculated to lead to admissible evidence, which means non-cumulative. We need to focus on the hot documents. We need to remember

© Ralph Losey 2012

that all that really matters are the five to nine of the hottest documents. This is what the trial lawyers need to tell their story of prosecution or defense. The few other documents that you may want to put into evidence are just window dressing. The millions of other technically relevant documents are of little or no use in the preparation for trial, and of no use whatsoever in the conduct of a trial.

This means we need smart AI enhanced software tools; software that we can teach to find the hottest documents. Software that has ranking built in as a core function. It also means that we need informed e-discovery attorneys who understand the secrets of search. They can then bridge the gap that now exists with trial lawyers. Then maybe the current e-discovery strategy used by most lawyers today of avoidance will be abandoned. Then maybe all lawyers will adopt proportional e-discovery designed for trial. There is a new year coming. Let's all resolve to work together as a team to make it happen! Let's focus our efforts. As Pythagoras supposedly said: *Do not talk a little on many subjects, but much on a few.*

For questions or comments, please feel free to contact the author at:
Ralph.Losey@JacksonLewis.com.

© Ralph Losey 2012