



WHITE PAPER

E-Discovery Search: The Truth, the Statistical Truth, And Nothing But the Statistical Truth+

By Nick Brestoff, M.S., J.D.

Western Regional Director, Discovery Strategy &
Management International Litigation Services,
www.ilsTEAM.com

+ This Article was first published by the American Bar
Association's E-Discovery and Digital Evidence Journal, Vol. No.
1, Issue No. 4 (Autumn 2010)

Introduction

This article is a call to revisit Rule 26(g)(1) of the Federal Rules of Civil Procedure, which requires attorneys to certify “to the best of the person’s knowledge, information, and belief formed after a reasonable inquiry” that disclosures are “complete and correct.” Given the exponentially growing mountain of electronically stored information (ESI), and the incompleteness and statistical nature of search technologies, which this article will explain, no attorney can honestly so “certify.” One day, this gap, a loophole between the law of yesterday and the technology of today, will cause a monumental waste of judicial, attorney, and client resources.

Most of us know the meaning of a “loophole.” These days, when one seeks a definition, or perhaps an example, we look online and, more often than not, we turn to Wikipedia. According to Wikipedia, “[a] loophole is a weakness or exception that allows a system, such as a law or security, to be circumvented or otherwise avoided. Loopholes are searched for and used strategically in a variety of circumstances, including taxes, elections, politics, the criminal justice system, or in breaches of security.”

Wikipedia mentions the “criminal justice system.” But to this entry we must add our system of “civil justice,” and, in particular, the giant middle of every lawsuit, discovery. As most attorneys are now aware, what used to be thought of as “discovery” is now dominated by e-discovery.

But e-discovery is a hybrid, a confluence of slowly changing laws and rules, on the one hand, and rapidly changing computer-based technologies, on the other. In this dynamic context, which besets every system of justice in the world, loopholes may be expected. Here we explore a rather large disconnect (or loophole) in the U.S. system of justice which comes as a result of the new complexities of e-discovery.

Loopholes

Loopholes can be large or small. In 2005, for example, Wal-Mart proposed a large store in Calvert County, Maryland. Because Calvert County restricted the size of a retail store to 75,000 square feet, Wal-Mart's executives and attorneys proposed building two separate smaller stores, which, technically speaking, would not have violated the restriction. The plan was controversial, and Wal-Mart later withdrew it. Until Wal-Mart made the proposal, however, this legal loophole went undetected.

One further example will serve to demonstrate that when loopholes are exploited, money – big money -- is usually at stake. Ford imports a vehicle called Transit Connect from Turkey, but pieces of its interior are shredded when they arrive in Baltimore to circumvent the 1963 Chicken Tax, which imposes a 25% tariff on imported light trucks. Ford avoids this 25% tariff on its Transit Connects because it does not import these vehicles as light trucks; instead, they are imported as passenger vehicles with rear windows, rear seats and rear seatbelts, and are immediately converted into light trucks when they arrive, by replacing the rear windows with metal panels and by removing the rear seats. This change costs Ford hundreds of dollars, but it saves thousands in taxes.

In the context of e-discovery, lawyers have attempted to exploit what they thought were loopholes right from the start. Examples abound. In one case, for example, when the format for producing ESI was not specified and emails (and only emails) were requested, they were produced, but they were “divorced” from their attachments, which were not produced. In another case, a producing party converted searchable documents into nonsearchable TIFF files before producing the ESI.

These gambits revealed certain weaknesses in the system, and some of them have been addressed. Now, for example, the federal rules provide that when a party is seeking documents from an opposing party, or from a third party pursuant to a subpoena, the requesting party may specify the form or forms of the documents when they are produced. California's e-discovery statutes also provide that the requesting party may specify the “form or forms in which each type of [ESI] is to be produced,”

but, like the federal rules, the requesting party has only an opportunity to specify the forms once.

Even though its growth-rate is prodigious, the hallmark of e-discovery is the immense volume of ESI that must be addressed. In 2003, researchers at UC Berkeley published an update to their study, How Much Information? At that point in time (and now hopelessly outdated), they explained that each year almost 800 megabytes of recorded information was produced per person, and that 92% of that information was stored on computers or a computer-based storage system. Eight hundred megabytes is enough to fill a set of books stacked 30 feet high. Today, if each person generated only 25% more information than in 2003, or 1,000 megabytes, then each person would generate a gigabyte of data per year, and that amount is roughly equivalent to 75,000 pages, if printed. It is easy to imagine that today we generate much more than that. Indeed, it is often said that 98% or 99% of all the information generated today, by everyone in the world, is generated as ESI. Why? Because today the digital universe includes not only servers, desktops, laptops, cell phones, hard drives, flash drives, and photocopy/fax machines, the digital universe includes data from TV and radio transmissions, telephone calls received as emails, surveillance cameras, datacenters supporting “cloud computing,” and, of course, social networks.

So, in lawsuits, parties and attorneys must often deal not just with gigabytes of data, but with several terabytes of data, and a single terabyte is roughly equivalent to 75 million pages, if printed. Even if a requesting party asks for readily accessible data, meaning data in native format with metadata intact, there is still the problem of how to search through a much, much bigger haystack than lawyers ever faced when, e.g., 10,000 boxes of documents were produced.



Key Words and Boolean Searches

Now, how can anyone get their arms around this much data? They can't. The volume of data today is far greater than those times when parties attempted to hide the needle in the haystack by producing truckloads, or worse, warehouses full of boxes stuffed with papers. E-discovery expertise is partly the domain of an information technologist and partly the domain of lawyers. The technologist's approach is to cull the data by removing exact duplicates (de-duping) and system files. Culling will certainly reduce the size of the data set. Now the lawyer's task is to query that data set with key words and "field" terms, just as they did when searching opinion databases for applicable case law. Because they are familiar with key words, the receiving attorneys include key words describing the subject matter of the dispute and the names of the key players and employees who had "any involvement with the issues raised in the litigation or anticipated litigation." An oft-used field term is a date or a range of dates.

Indeed, in the context of online legal research, teams of lawyers and other law firm denizens have become "power users" of key words and field terms. It was not always so. Fifty years ago, lawyers relied on their memories and library tables populated with books. Their search technique was non-linear and depended on a more personal skillset. But once the published cases were uploaded and computers could be used to hunt through databases, key words, date ranges, and Boolean connectors (e.g., AND, OR, NOT, term X "within 7 of" term Y, etc.) were deployed. Lawyers have been using this technique for over 35 years.

But now the scope of the data is vastly increased and the problem is different. The problem is different because we are not querying databases of published opinions in which courts use familiar legal terms. In the e-discovery context, we are working within the context of the law, but we are not looking for it. We are trying to find the facts, and we are trying to find them in a mountain of data that is not only enormous, it is contained in numerous places. In this endeavor, opposite sides have different goals, especially because they treat the discovery process as an adversarial adventure and notwithstanding the platitudes spoken about cooperation.

For example, a requesting party may attempt to use key words to over-collect documents. In one recent case, for example, where, pursuant to a stipulated order the defendant had sole discretion to specify search terms, the defendant submitted 400 search terms. Over the producing party's objections based on cost (\$6 million), which the court denied because of the stipulated order, these 400 terms yielded 660,000 documents.

On the other hand, a producing party may attempt to under-produce. They may use key words to narrow the scope of the documents they must produce. In a different case, the plaintiff requested documents from the hard drives of 26 employees. The defendants used de-duplication to narrow the documents to be produced down from 423,835 to 129,000, and then used search terms to narrow the actual production down to 4,000 documents. The plaintiff objected, and wanted more, but the magistrate dismissed the plaintiff's objections, stating "To the extent Plaintiff contests the adequacy of the search terms, it has not set forth an alternative search methodology; moreover, no specific challenge to the search terms has been brought and briefed before the Court."

Ah, now there's a rub. Is there an alternative search methodology? Yes. But before describing it, let's stay with key words for a moment. The goal, after all, is to use automated, computer-based searches to find as many of the potentially relevant documents as we can. All non-privileged information relevant to a claim or defense must be produced.

But just how successful are key word searches? Test yourself. Here's the proposition: Key words using Boolean connectors will find only about 25% of the relevant documents. True or false?

True! One of the founders of the "information retrieval" field, M. E. Maron (now professor emeritus, UC Berkeley) reported as long ago as 1985 that attorneys were over-estimating the efficacy of their searches. The attorneys thought they were identifying 75% of the relevant documents, but they were wrong: they were finding only

about 20%. More recently, studies show that key word searches are, even today, only a little more successful. Tomlinson and others reported in 2008 that Boolean searches identified only 22% of the relevant documents, while Oard and others reported in 2009 that Boolean searches pinned only 24% of the relevant documents. (These reports come from the Legal Track of the Text Retrieval Conference (TREC), which is administered by the U.S. National Institute of Standards and Technology.)

Now for attorneys used to key word searches, these reports are not good news. As previously noted, in the process of “early disclosure” and responding to document requests, an attorney must certify that, “to the best of [their] knowledge . . . formed after a reasonable inquiry,” their response to a document request is “complete and correct.”

Is there an alternative methodology to key words and field terms? Yes. We come to it now: concept search.



Concept Search

What is concept search? Concept search is a way of finding patterns in unstructured data sets. It sounds technical, doesn't it? Yes, it is. It involves matrix algebra, formulas you don't want to see (ever), and statistical concepts you don't want to know about, but will be forced to learn anyway (note: more on this point, later).

Let's stick with key words for a moment. Key words approach a document collection in a simplistic way; either a document contains the key word (or a variation of it) or it does not contain that word. Let's say we have only two key words, w1 and w2, for our query, and that we find w1 in document 1, which we'll call d1, and w2 in document 2, or d2; but we do not find w1 in d2 and we do not find w2 in d1. In the four-square box at the end of this sentence, a “1” means that the word in question is present, while a “0” means that the same word is not present:

This simple “picture” is a hypothetical word-document matrix. It is clear that using w1 as “input” will result in d1 as “output,” but not d2. If we use w2 as input, we will get d2, but not d1. But if we are looking for a document with both w1 AND w2, we will get nothing.

Dox -->	d1	d2
w1	1	0
w2	0	1

But wait. This matrix is too simplistic. It consists of only two key words and only two documents. The documents in our collection, which will likely consist of gigabytes and terabytes of data, are certain to have many more than one word each. Here is the key to understanding what concept search engines do: they find with “co-occurrences” of words that are not used as search terms.

If a picture is worth many words, a bigger matrix should help. You can see what co-occurrence means by looking at the next matrix.

Dox	d1	d2	d3	d4	d5	d6	d7	d8
Word								
w1	1	0	0	0	0	0	0	0
w2	0	1	0	0	0	0	0	0
w3	1	1	0	0	0	0	0	0
w4	0	1	0	1	1	1	1	1
w5	0	1	0	0	1	1	1	1
w6	0	1	0	0	0	1	1	1
w7	0	1	0	0	0	0	1	1
w8	0	1	0	0	0	0	0	1

It starts in the upper left hand corner with the simple four square matrix of (w1, w2) and (d1, d2) that we first described. But then this matrix adds more words (w3 through w8) and more documents (d3 through d8).

Let's begin with w3. It appears in both d1 and d2. When we were considering the four-square matrix, inputting w1 AND w2 did not result in either d1 or d2; it resulted in nothing. In the matrix below, if we input w3, we will get d1 and d2, because it is contained in both documents.

Now look at w4. It is contained in d2 and d4, d5, d6, d7, and d8. Similarly, w5 is in d2 as well as in d5, d6, d7 and d8 (so one less; w5 is not in d4). And so on. Now we can make some observations about our collection (or corpus).

First, note that neither w1 nor w2 are in any of the other documents, d3 through d8, which is why, for the w1 and w2 rows, there are nothing but "0s" in the columns after d2. For both the rows for w1 and w2, the columns d3 through d8 are all zeros.

Also, no matter what word we use to query this matrix, will we ever get back d3? No. It has none of the words on the list.

Now let's look at words w4, w5, w6, w7, and w8. Notice that w4 shows up in d4 through d8. Fine, that word is used frequently. But frequency is not the test.

The big idea of concept search is to find documents (as output) that are responsive to a query (using key words as input), based on co-occurrences. As output, we want documents that have key words in them, but also the documents that do not contain any of the key words but which are nevertheless potentially related and, thus, potentially relevant. We are looking for patterns.

In this regard, patterns can be strong or weak. Which document exhibits the strongest pattern? It's d8. Although d8 does not have our input key words, w1 or w2, column d8 has five of the same content words contained in d2; that is, both d2 and d8 have words w4 through

w8 in common. The weakest pattern involves the most documents but the weakest link: d4 through d8 all share only one word – w4 – with d2.

Computers do not understand “patterns.” They go through a process (a series of steps) which eventually leads to a measurable threshold, a cut-off point. To scholars in the field of Information Retrieval, such steps, including the mathematical scissors, is called an “algorithm.” In our simplistic hypothetical, if we want all documents that are potentially relevant, we might choose a cut-off where there is only one matching co-occurrence, a low threshold. If we want to find a “smoking gun,” we might search again, this time adjusting our process (algorithm) to find only the strongest co-occurrences. In this example, if we want, say, more than four (4) co-occurrences, the search output would be only d8.

See how this works? With concept searching, computers are going through gigabytes and terabytes of data consisting of documents and words, using a strictly mathematical approach.

This search methodology is called Latent Semantic Indexing or LSI. This term is best understood “inside out.” The “Index” part is simple. You have seen indexes before. They are at the end of nearly every book. Indexes indicate which words are on which page. Here, the computer ingests all of the documents and all of the words, and creates an index of each word that is contained in each document. We have just done this with two hypothetical matrices, one with two words and two documents, the other with eight words and eight documents.

What does “Latent” mean? Roughly speaking, it means “hidden.” And “Semantic” means, again roughly, “meaning.”

So, the phrase is actually descriptive of what we are trying to accomplish: find the hidden meanings (patterns) in a collection of documents, not because of the specific words we choose as input, but because of the other words in the documents containing the words we did choose and their “co-occurrence” with words in other documents, documents which do not contain our search terms.

Let’s deepen our understanding. As we did with the documents themselves, culling out exact duplicates and system files, let us cull our words. In LSI, we discard articles (like “a” and “an”); prepositions, conjunctions, common verbs (like known, see, do, be); pronouns (e.g., it, they); common adjectives (like big, late, and high); pointer or frilly words (like thus, therefore, however, and albeit); any words that appear in every document; and any words that appear in only one document. Now we are down to the core words that have semantic value; they have “content.” It is with these words that we form the word-document matrix.

Now we do some “weighting” (think “handicapping”). Some content words appear more than once in a single document. They are given greater weight; and the process of giving them more weight is called “local weighting.” Still other words show up frequently throughout the entire set, and because of this, they are “commonplace.” Words that appear in only a small handful of documents may have special significance. They get greater weight. This is “global weighting.” And there is a scaling step, called “normalization,” which is just like handicapping in golf. Some documents may be long ones and have many key words. To keep them from overwhelming the shorter documents, the larger ones are penalized a bit, so that every document has, approximately, equal significance.

Because LSI is mathematical, it is a search engine that “likes” addressing large collections of data. The more words and documents in the set, the better LSI performs at finding documents responsive to a query. And, after a fruitful search puts some documents into a “shopping cart,” a human being can learn from the initial results and iterate the process. With this feedback, the input terms are more focused and the LSI search engine is likely to produce even better results.

LSI was not conceived to address the problem of search in the e-discovery context. But it has found application in the world of e-discovery. Moreover, because many business and governmental endeavors involve more than one language, LSI is useful because it does not pretend to understand anything about the words it is considering. The words are, in a sense, digitized; then LSI creates the word-document matrix, and seeks out the patterns

based on statistical co-occurrences. It is therefore as functional with words in Chinese, Korean and Japanese (or Arabic) as it is with words in English. Using LSI, “hot” documents across different languages can be identified. The next step is machine translation, which is not known for precision. So, the step after that is human review. And if certain documents appear to a human to be suitable for use in deposition, in a motion, or at trial, the final step is human translation, so that the translated documents can be certified and offered into evidence.

In the e-discovery context, you have likely seen LSI in action. You just didn’t know what was “under the hood.” Simply put, concept search based on LSI, or a variant of LSI, is now at the heart of programs that are offered by a number of different vendors, each of which has provided different “bells and whistles” to differentiate themselves.

Why is LSI powerful? Because, when LSI is used on unstructured data, such as business communications, LSI returns documents that may be highly relevant that even power key word searching would miss. Here’s an example. In a stock option back-dating case, an LSI-based search returned documents whose common denominator (pattern) was the phrase “Let it roll.” Why return these documents? Remembering that LSI is designed to seek out hidden meanings, the consultants involved in the case called the “Let it roll” group to the attention of the litigators. Sure enough, this phrase turned out to be the “go” signal the executives were using to authorize the back-dating. Unless a power key word searcher made a lucky guess, the “Let it roll” documents – the key needles in a very large haystack – would have gone undetected.

So LSI has proven to be more efficient than key words, even though key words are still used in the queries that are framed. But could you have explained LSI to a court, in case you were challenged by opposing counsel to do so?

It’s All Statistical

Now, finally, we come back around to whether an attorney can honestly sign off on the Rule 26 certification concerning the documents he or she has disclosed or produced. With a new appreciation for what goes into searching a collection for potentially responsive documents, the answer is “no.” We have a loophole. Attorneys are, by rule, being forced to certify to a degree of certainty that just is not there; and they put their licenses on the line when they sign.

Suppose we have collected 100 million documents; how many should be produced? A suitably sized random sample will accurately reflect the number of responsive documents to be produced, no matter how large the set may be. For a confidence level of 95%, with an error of plus or minus 5%, a random sampling of 1,537 documents must be examined. For a confidence level of 99%, with an error of plus or minus 1%, a sampling of 66,358 documents is needed. Thus, “if we have 100 million documents in the unretrieved set, we need to examine only 1,537 documents to determine within 95% confidence that the number of responsive documents in the unretrieved set is within the margin of error. If we find that there are 30 documents that were responsive in the unretrieved set, we can state that we have 95% confidence that the number of responsive documents in the sampled set is between 28 and 32 (rounding up the document count on the high end, rounding down on the low end). Extending that to the 100 million population, approximately 1,951,854 plus or minus 97,593 are responsive in the unretrieved set. [Para.] In the case of a review where errors are expensive (such as a review for privilege), 99% confidence with 1% error condition would require 66,358 samples. If we identify 200 privileged documents in such a sample, you will have 99% confidence that the number of privileged documents in the sample is between 198 and 202 privileged documents.”

Some Proposals and a Grand Conclusion

As previously mentioned, responding attorneys must currently certify that, “to the best of [their] knowledge . . . formed after a reasonable inquiry,” the disclosure or response to a document request is “complete and correct.” But in this digital era, attorneys must face up to understanding some of the math they hoped to avoid (forever) by going to law school, because attorneys are ill-equipped to flatly certify the “completeness” of their disclosures or responses. “[T]he assumption on the part of lawyers that any form of present-day search methodology will fully find ‘all’ or ‘nearly all’ available documents in a large, heterogeneous collection of data is wrong in the extreme.” So how can attorneys vouch for “completeness”? Clearly, attorneys who continue to sign off on Rule 26(g) certifications are over-promising. They are venturing into areas where an expert’s opinion is warranted, if not necessary. If a client is prejudiced when a court agrees, after some future battle over the alleged impropriety of an attorney’s certification, that “completeness” was promised but not achieved, will that attorney have fallen below the standard of care? Having likely ventured beyond his or her competence, will that attorney have violated a rule of professional conduct? Is a malpractice lawsuit in that attorney’s future?

We come now to four concrete proposals for change, and one grand conclusion:

- » Rule 26(g)(1)(A) should be changed to indicate (for example) that, with the assistance of experts, the document production is complete and correct, with a 95% confidence level and an error rate of plus or minus 5%;
- » Attorneys would be wise (as a matter of best practices) to sample for privileged documents, so that they are withheld with a 99% confidence level and an error rate of plus or minus 1%;
- » Malpractice insurers should be actively revising their applications for errors and omissions insurance to force attorneys to disclose the level of their e-discovery competence, and insurers should be monitoring, if not mandating, the continuing education of attorneys in e-discovery matters.
- » Besides being able to choose the format for the production of ESI, requesting parties should be able to designate the search methodologies used by the responding parties to search for potentially relevant documents. Otherwise, responding parties may use key words and search methodologies that under-produce to the requesting party.

The grand conclusion brings us back to loopholes. In an adversarial system, attorneys will exploit loopholes. And now you know that a large technical loophole besets our system. It besets every judicial system in the world, and we have not yet faced up to it.

We seek the truth. But now that there’s so much data, the best we can say about the truth is this: it’s statistical.

###

After graduating with a B.S. in engineering systems from the University of California at Los Angeles (U.C.L.A.), Nick Brestoff earned an M.S. in environmental engineering science from the California Institute of Technology (Caltech) and graduated from the Gould School of Law at the University of Southern California (U.S.C.). During his litigation career, Mr. Brestoff litigated business, employment, environmental, and other civil disputes in state and federal court. He is currently a consultant to businesses and attorneys through International Litigation Services (www.ilsTeam.com). Mr. Brestoff’s email address is nbrestoff@ilsTeam.com. He gratefully acknowledges editorial comments on drafts from Helen Marsh, attorney at law (California), Ken Rashbaum, attorney at law (New York), and Nicolas Nunez, P. Eng. (California).

References

¹ Rule 26(g)(1)(B) applies the certification to discovery responses, and requires a certification that is “consistent” with the rules, which includes Rule 26(g)(1)(A).

² The Wikipedia entry for “Loophole,” as modified on 27 July 2010, was viewed by the author on August 27, 2010.

³ Paley, Amit R. (May 17, 2005) “Wal-Mart Drops Plan for Side-by-Side Calvert Stores.” *The Washington Post*. <http://www.washingtonpost.com/wp-dyn/content/article/2005/05/16/AR2005051601271.html>.

⁴ Dolan, Matthew (September 22, 2009) “To Outfox the Chicken Tax, Ford Strips Its Own Vans.” *The Wall Street Journal*. <http://online.wsj.com/article/SB125357990638429655.html>.

⁵ See *PSEG Power N.Y., Inc. v. Alberici Constructors, Inc.*, No. 1:-05-CV-657 (N.D.N.Y. 2007) (producing party ordered to re-produce ESI at its cost).

⁶ See *Goodbys Creek, LLC v. Arch Ins. Co.*, No. 3:07-cv-947-J-34 HTS (M.D.Fla. 2008) (conversion held improper; producing party order to re-produce ESI); *L.H. v. Schwarzenegger*, 2008 U.S. Dist. LEXIS 86829 (C.D.Cal. 2008) (sanctions were imposed for the untimely (late) production of non-sortable PDFs).

⁷ Federal Rules of Civil Procedure, Rules 26(f)(3)(C) [discovery plan] and 34(b)(1)(C) [content of the request].

⁸ California Code of Civil Procedure §2031.030(a).

⁹ Lyman, Peter and Varian, Hal, *How Much Information?* (2003); see <http://www.sims.berkeley.edu/how-much-info-2003> (reviewed on August 28, 2010).

¹⁰ *Ibid.*

¹¹ Keteyian, Armen, “Digital Photocopiers Loaded with Secrets: Your Office Copy Machine Might Digitally Store Thousands of Documents That Get Passed on at Resale,” *CBS News* (New York, April 15, 2010); See <http://www.cbsnews.com/stories/2010/04/19/eveningnews/main6412439.shtml?tag=mncol;txt>.

¹² Gantz, et al., *The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011* (March 2008) (Executive Summary). See <http://www.idc.com>.

¹³ See *Pension Comm. Of the Univ. of Montreal Pension Plan v. Banc of Am. Sec., LLC*, 2010 WL 184312 (S.D.N.Y. Jan. 15, 2010; as amended May 18, 2010) (Scheidlin, J.)

¹⁴ Here is an example of a “power key word search” using Boolean operators (which were borrowed from computer programming): [successor /5 corporation] /p (toxic or hazardous or chemical or dangerous /5 waste) /p clean! and (aft 1/1/90). In plain language, this search is for cases where a successor corporation is liable for the cleanup of hazardous (toxic) waste. The sample Boolean search looks for the combination of successor within five words of corporation, in the same paragraph as the combination of toxic or hazardous or chemical or dangerous within five words of waste, within the same paragraph as clean or cleanup or cleans or cleaned or cleaning (the exclamation mark in clean! causes the computer to search for all words with clean as a root). Cases are limited to those dated after January 1, 1990.

¹⁵ See *In re Fannie Mae Secs. Litig.*, 552 F.3d 814, 818-819 (D.C.Cir. 2009).

¹⁶ See *In re CV Therapeutics, Inc. Sec. Litig.*, 2006 WL 2458720 (N.D.Ca. Aug. 22, 2006).

¹⁷ Federal Rules of Civil Procedure, Rule 26(b)(1); see *Zubulake v. UBS Warburg LLC*, 217 F.R.D. 309, 316 (S.D.N.Y. 2003); *SEC v. Collins & Aikman Corp.*, 256 F.R.D. 403, 417-418 (S.D.N.Y. 2009) (over objections based on cost, SEC ordered to produce emails; parties required to establish a reasonable search protocol).

¹⁸ Maron, M. E., *An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval Sys.*, 28(3) *Comm. of the ACM* 289 (1985).

¹⁹ Tomlinson, Stephen, et al., *Overview of the 2007 TREC Legal Track* (April 30, 2008).

²⁰ Oard, Douglas W., et al., *Overview of the 2008 TREC Legal Track* (March 17, 2009).

²¹ Federal Rules of Civil Procedure, Rule 26(g)(1)(A).

²² Two tests are “recall” and “precision.” Recall is the proportion of relevant documents identified out of the total number of relevant documents that exist. If the total number of relevant documents is 100, but a search identified 80, the recall rate is 80%. Precision is the percentage of identified documents that were actually relevant. If 100 documents were identified but only 75% of them were relevant, the precision would be 75%. Using LSI, recall and precision rates just under 90% have been achieved. Source: Content Analyst Company, LLC (“Content Analyst”) in Reston, Virginia (<http://contentanalyst.com>). Content Analyst is the original patent-holder of LSI.

²³ See Landauer, T. K. and Dumais, S. T., “Solution to Plato’s Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge,” *Psychological Review*, 104(2), 211-240 (1977).

²⁴ There are at least three hosted review platforms that have integrated an LSI solution from Content Analyst: Relativity (by kCura), iCONNECT, and Eclipse by IPRO. In addition, a variation of LSI called Probabilistic LSI is “under the hood” of Axcelerate by Recommind.

²⁵ For that matter, could you differentiate LSI from still other computer-based search approaches, including taxonomies, ontologies, and Bayesian classifiers? These topics are beyond the scope of this article.

²⁶ See *Qualcomm, Inc. v. Broadcom Corp.*, No. 05 Civ. 1958-B, 2008 U.S. Dist. (S.D.C at. Jan. 7, 2008); and *id.*, *Order Declining to Impose Sanctions, Etc.* (Document 998; filed Apr. 2, 2010).

²⁷ Search Guide, *Electronic Discovery Reference Model Draft v.1.17* at p. 79 of 83 (May 7, 2009).

²⁸ *Ibid.*

²⁹ *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D. 251, 262 (D.Md. 2008) (Grimm, J.) (quoting from “Information Inflation: Can the Legal System Adapt,” 13 *Rich. J. L. & Tech.* 10 (2007), at *38, 40). See *Mt. Hawley Ins. Co. v. Felman Prod., Inc.*, 2010 WL 1990555*10 (S.D.W.Va. May 18, 2010) (failure to sample in order to identify and remove privileged documents was “imprudent”).

³⁰ In several recent cases, courts have made statements supporting the proposition that a certification of completeness of a large document product by an expert should replace certification by an attorney. For example, in *United States v. O’Keefe*, 537 F.Supp.2d 14, 24 (D.D.C. 2008), the court stated, “Whether search

terms or 'keywords' will yield the information sought is a complicated question involving the interplay, at least, of the sciences of computer technology, statistics and linguistics Given this complexity, for lawyers and judges to dare opine that a certain search term or terms would be more likely to produce information than the terms that were used is truly to go where angels fear to tread." In *Equity Analytics, LLC v. Lundin*, 248 F.R.D. 331, 333 (D.D.C. 2008), the court stated, "Determining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a lay person (and a lay lawyer)" And in *In re Seroquel Prods. Liab. Litig.*, 244 F.R.D. 650, 660 n. 6, 662 (M.D.Fla. 2007), the court criticized the defendant's use of keyword search to select ESI for production, in particular because the defendant failed to provide information "as to how it organized its search for relevant material, [or] what steps it took to assure reasonable completeness and quality control," and noting that "while key word searching is a recognized method to winnow relevant documents from large repositories . . . [c]ommon sense dictates that sampling and other quality assurance techniques must be employed to meet requirements of completeness." (Emphasis added.)