Fifth DESI Workshop on Standards for Using Predictive Coding, Machine Learning, and Other Advanced Search and Review Methods in E-Discovery

DESI V Workshop, June 14, 2013, Position Paper
Presented by: Dan Regard and Tom Matzen, iDiscovery Solutions, Inc.

A Re-Examination of Blair & Maron (1985)

Abstract

David Blair and M. E. Maron wrote a seminal paper on full-text retrieval in 1985. That paper concluded that full-text searching was more expensive and less satisfactory than manually indexed databases.

Today this paper is often cited for the position that search terms cannot be more effective than they were for Blair and Maron. These citations are wrong to the extent that (a) they often appear to misstate the original conclusions, (b) they typically fail to fully consider the specific methods and objectives of the original researchers, and (c) subsequent testing has demonstrated that search terms can be very effective in a litigation context when developed in a methodical manner.

Detail

In 1985, David C. Blair and M. E. Maron published a paper in the journal "Communications of the ACM".[1] Since then their paper, entitled "An Evaluation of Retrieval Effectiveness for a Full-Text Document Retrieval System" has been cited many times in the realm of information retrieval and, in particular, in U.S. based electronic discovery. Like a famous musician, the study's name has been shortened as the notoriety has increased. It is now often referred to as simply "Blair & Maron." For the purpose of this paper, and to distinguish from subsequent papers that bear light on their original work, I will refer to that research as Blair & Maron (1985) and as the authors, themselves, as Blair and Maron. The authors, themselves, often refer to the paper as the "STAIRS study" because they were using the IBM STAIRS/TLS full-text retrieval system.[2]

Blair & Maron (1985) is famous for being an early, well-funded, substantial study of the use of search terms and their ability to provide "satisfactory levels of retrieval

---

[1] Blair D, Maron ME (1985) An evaluation of retrieval effectiveness for a full-text document-retrieval system. Communications of the ACM 28(3):289–299.

[2] **ST**orage and **I**nformation **R**etrieval **S**ystem/**T**hesaurus **L**inguistic **S**ystem ("STAIRS/TLS")

effectiveness." In fact, they were actually studying prior papers from 1960 and 1970 to determine if full-text retrieval really was better than manually coded indexes.

They eventually concluded that full-text searching was *not* cost effective when compared to searching manually assigned index terms. They based their conclusion on their perceived high cost and poor performance.

Costs were different in 1985

The high cost assessment was based on the scarcity of full text and the cost of storage. The poor performance assessment was based on *recall* and *precision*.

In 1985, full-text was scarce. Documents were largely available as paper documents, not as full-text electronic files. Paper to text conversion, via optical character recognition ("OCR") was in its infancy. At the time, it was estimated to cost 20-times as much to type and verify full text documents as it did to type and verify manually assigned index terms.

In 1985, storage was expensive. In fact, it was over 100,000,000% (100 million percent!) more expensive than today.[3] At the time, Blair & Maron (1985) estimated that full-text databases required 20-times as much data storage and 50-times as much index storage as manually coded abstracts.[4]

As a result of these practical issues, cost was considered a very real component of whether or not to even attempt full-text retrieval across large data sets.

The low performance assessment was based on the average *recall* rate of 20%. The research attorneys also had an average *precision* rate of 79%. *Recall* measures how well a system retrieves all the relevant documents. *Precision* measures how well the system retrieves only the relevant documents.[5] Notably, the study never ran comparison retrievals using manually assigned index terms.[6]

Understanding those results and putting them into the context of the goals, techniques and tools of the experiment are the bases for the rest of this paper.

Shifting conclusions

Whereas Blair & Maron (1985) asked *if* we should consider full-text searching (they concluded the answer to be "no"), today we grapple instead with *how* and *when*, rather than *if*.

---

[3] In 1984, a 20 megabyte Tandem hard drive cost $2,399. That equates to $100,000 *per gigabyte*. Today, you can purchase a 1 terabyte Dell hard drive for $79.99. That equates to 7.8¢ *per gigabyte*. By that comparison, today's GB's are 1,280,160 times cheaper.

[4] Blair & Maron (1985), at 298.

[5] Id at 290.

[6] Blair & Maron (1990), at 441, ("…the STAIRS study did not examine how full-text techniques could be used to retrieve abstracts as opposed to complete documents")

Blair & Maron (1985) eventually concluded that "full text searching does *not* operate at satisfactory levels"[7] and "the system did not work well in the environment in which it was tested and that there are theoretical reasons why full-text retrieval systems applied to large databases are unlikely to perform well in any retrieval environment."[8]

In fact, because they found it un-satisfactory and too expensive, Blair and Maron (1985) concluded that full-text searching should be abandoned:

> *Full-text searching is one of those things, as Samuel Johnson put it so succinctly, that ". . . is never done well, and one is surprised to see it done at all."*[9]

In the intervening years, there have been many advances both in the technology and methodology of using search terms for litigation purposes. Getting access to full-text documents no longer requires manual entry. Instead, there is an over-abundance of full-text documents available. It is accepted almost as a given that full-text searching will be as satisfactory and will be more cost effective than manual indexing.

As Blair wrote in 1996:

> *Because of the widespread use of word processing, most documents now begin in computer-readable form. As a result, simple full-text retrieval techniques can be implemented very easily, becoming the method of choice by default.*[10]

As a result, today we seem to be more concerned as to whether full-text retrieval can outperform exhaustive human review then as to whether it can outperform manual indexing and retrieval.[11]

In fact, due to advances in information retrieval science, advances in computer performance, and a marked increase in the level of interest in information retrieval, there have been many other papers written analyzing Blair & Maron (1985), and the conclusions therein.[12,13,14] Even Blair and Maron have written follow-up and explanatory papers to their own work.[15,16,17]

---

[7] Blair & Maron (1985) at 297.

[8] Id. At 298.

[9] Id.

[10] Blair (1996) at 5.

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

[12] Dabney DP (1986), "The cures of Thamus: An analysis of full-text document retrieval", Law Library Journal, 87:5-40.

[13] Salton G (1986) "Another look at automatic text-retrieval systems", Communications of the ACM,

Nevertheless, many modern papers cite to Blair & Maron (1985) as the definitive work in the area of keyword search, particularly in a litigation context:

> *"Classic studies in the field of Information Retrieval which outline the perils and inherent accuracy of manual review processes have found new audiences. (see, e.g. Blair & Maron, 1985)."*[18]
>
> *"Keyword based search has been the bread and butter method of searching, but its limitations have been well understood and documented in a seminal study by Blair & Moran [sic]."*[19]
>
> *"The Blair Moran [sic] study illustrates this using an event which the victim's side called the event in question an "accident" or a "disaster" while the plaintiff's side called it an "event", "situation", "incident", "problem", "difficulty", etc. The combination of human emotion, language variation, and assumed context makes the challenge of retrieving these documents purely on the basis of Boolean keyword searches an inadequate approach."*[20]
>
> *"The majority of e-discovery services and software products rely on the use of keywords to identify and organize documents in preparation for human review. The legal community is familiar with keyword searching, which forms the bases of case law and statutory law research in legal databases. In the e-discovery context, the limitations of this approach are well documented and understood (Baron, J. 2009; Blair, D.C. & Maron, M.E. 1985, 1990; The Sedona Conference Working Group 2007; Tomlinson 2008)."*[21]

29:648-656.

[14] Salton G (1992) "The state of retrieval-system evaluation", Information Processing and Management, 28:441-449.

[15] Blair DC (1990), "Language and representation in information retrieval", Elsevier Science, Amsterdam.

[16] Blair DC and Maron ME (1990) "Full-text information retrieval: Further analysis and clarification", Information Processing and Management, 26: 437-447

[17] Blair DC (1996) "STAIRS Redux: Thoughts on the STAIRS Evaluation. Ten Years after", Journal of the American Society of Information Science, 47:4-22.

[18] Eli Nelson, A False Dichotomy of Relevance: The Difficulty of Evaluating the Accuracy of Discovery Review Methods Using Binary Notions of Relevance, May 2011.

[19] Jeremy Pickens, John Tredennick and Bruce Kiefer, Process Evaluation in eDiscovery as Awareness of Alternatives, May 2011.

[20] Pickens, Tredennick and Kiefer (2011).

[21] Thomas Barnett, Svetlana Godjevac, Jean-Michel Renders, Caroline Privault, John Schneider and Robert Wickstrom, Xerox Research Center Europe and Xerox Litigation Services, "Machine Learning Classification for Document Review", May 2009.

> *"Thus the work is generally monotonous, and this when combined with the often poor working conditions of the review teams [1] means that the work is likely to be highly error prone [2]"*[22]

Not only is the work still cited, some authors have adopted the interpretation that search terms can *only be effective* 20% of the time or that search terms are *per se* ineffective*:*

> *Predictive coding has been shown to identify 70-80% responsive documents as compared to the search term method used by Defendants, which has shown to only be 20-30% effective in identifying responsive documents.*[23]

> *[K]eyword searches usually are not very effective. In 1985, scholars David Blair and M. Maron collected 40,000 documents from a Bay Area Rapid Transit accident, and instructed experienced attorney and paralegal searchers to use keywords and other review techniques to retrieve at least 75% of the documents relevant to 51 document requests. David L. Blair & M. E. Maron, An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System, 28 Comm. ACM 289 (1985). Searchers believed they met the goals, but their average recall was just 20%. Id. This result has been replicated in the TREC Legal Track studies over the past few years.*[24]

> *Using keywords to cull and filter documents, and thereby failing to produce 80% of relevant documents is not acceptable by any standard.*[25]

> *Nor is it a typical practice to review the documents left behind by a keyword search, even though upwards of 80% of the relevant documents typically are missed.*[26]

> *"Predictive Coding" is superior, more efficient and less expensive than the Search Term methodology. Consequently, the defendant's premature and unilateral selection of the Search Term method deprived the Plaintiffs of as much as 80% of the relevant documents. When correctly "trained" at the outset of a case, Predictive Coding software successfully identifies approximately 75-*

---

[22] Jacki O'Neill, Caroline Privault, Jean-Michel Renders, Victor Ciriza, and Gregory Bauduin, Xerox Research Center Europe, "DISCO: Intelligent Help for Document Review", May 2009, where footnote [2] cites Blair & Maron (1985).

[23] Declaration of Daniel S. Robinson, Fosamax / Alendronate Sodium Drug Cases, JCCP 4644, April 12, 2013.

[24] DaSilva Moore v. Publicis Group, No. 11 Civ. 1279 (ALC) (AJP) (SDNY, February 24, 2012)

[25] Declaration of Douglas E. Forrest, Fosamax / Alendronate Sodium Drug Cases, JCCP 4644, April 12, 2013. Relying on earlier citation of Blair & Maron (1985).

[26] Global Aerospace Inc., et al. v. Landow Aviation, L.P., Consolidated Case No. CL 61040, Circuit Court for the Loudon County, Virginia, Memorandum in Support of Motion for Protective Order Approving the Use of Predictive Coding, April 9, 2012, at 18.

> *95% of all relevant documents, while the antiquated "search term method" results in only approximately 20%.*[27]
>
> *Predictive Coding software successfully identifies approximately 75-95% of all relevant documents, while the antiquated "search term method" results in just approximately 20%.*[28]
>
> *Even if cooperatively developed, Search Terms are inherently unreliable—yielding a paltry 20% of relevant documents.*[29]
>
> *The use of Search Terms that were selected by Biomet likely resulted in a less than 20% relevancy rate.*[30]
>
> *The Blair and Maron study (discussed below) reflects that human beings are less than 20% to 25% accurate and complete in searching and retrieving information from a heterogeneous set of documents (i.e., in many data types and formats). The importance of this point cannot be overstated, as it provides a critical frame of reference in evaluating how new and enhanced forms of automated search methods and tools may yet be of benefit in litigation.*[31]
>
> *Simple key word searching has been known to be ineffective for decades. Back in 1985, Blair and Maron conducted a study of the effectiveness of key word searches and found they only retrieved 20% of the relevant documents.*[32]

In the intervening 30 years of new techniques and new technology, we have been able to break free of the presumed boundaries of full-text retrieval and have shown that 20% is not a limit.[33]

---

[27] *In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation (MDL 2391)*, Plaintiff's Memorandum in Support of Collaborative Predictive Coding, April 1, 2013.

[28] Id at 7.

[29] Id at 11.

[30] Id.

[31] The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery, August 2007, Public Commentary Version, The Sedona Conference Journal, 2007, at 199.

[32] Tredennick J, 2011, New Model E-Discovery Order for Patent Cases Turns Fishing Expeditions into Games of "Go Fish", E-Discovery Search Blog: Technology, Techniques and Best Practices.

[33] This is not to say that a given set of search terms must surpass 20% recall to be deemed effective. The practical application of search terms for inclusive and exclusive analysis for preservation, collection, culling, review and production; the value of negotiated search terms; and the use of search terms in phased discovery are but a few of the many ways that search terms can be used and calibrated for different goals, all the while bearing case specific cost, burden and proportionality in mind.

One immediate example of this, at least in the litigation space, is the result of the 2008 TREC Legal Track where the submission by H5, using an iterative development of search terms, achieved over 75% *recall* and 75% *precision*.[34] The same team, H5, repeated this again in 2009.[35]

What some have stated as fact, therefore, is simply not. Instead, it is a matter of debate and interpretation. Who better to understand that advocates may bend debate into pseudo fact than a judge? As Magistrate Judge John Facciola stated with respect to the tendency of lawyers to misstate issues of debate:

> *I recently commented that lawyers express as facts what are actually highly debatable propositions as to efficacy of various methods used to search electronically stored information. United States v. O'Keefe, No. 06-CR-249, 2008 WL 44972, at *8 (D.D.C. Feb. 18, 2008).*[36]

In fairness, some papers do clarify the study. As a recent glossary of technology-assisted review terms stated in reference to Blair & Maron (1985):

> ***Blair and Maron****: Authors of an influential 1985 study (David C. Blair & M.E. Maron,* An Evaluation of Retrieval Effectiveness for a Full-Text Document-Retrieval System*, 28 COMMC'NS ACM 289 (1985)), showing that attorneys supervising skilled paralegals believed they had found at least 75% of the Relevant Documents from a Document Collection, using Search Terms and iterative search, when they had in fact found only 20%. That is, the searchers believed they had achieved 75% Recall, but had achieved only 20% Recall. In the Blair and Maron study, the attorneys and paralegals used an iterative approach, examining the retrieved Documents and refining their search terms until they believed they were done. Many current commentators incorrectly distinguish the Blair and Maron study from current iterative approaches, failing to note that the Blair and Maron searchers did in fact refine their search terms based on their review of the Documents that were returned in response to their queries.*[37]

---

[34] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2008 Legal Track. In The Seventeenth Text REtrieval Conference (TREC 2008), 2009.

[35] Douglas W. Oard, Bruce Hedin, Stephen Tomlinson, and Jason R. Baron. Overview of the TREC 2009 Legal Track. In the Eighteenth Text REtrieval Conference (TREC 2009), 2010.

[36] *Equity Analytics v. Lundin* No. 07-2033, (D.D.C. Mar. 7, 2008).

[37] Cormack G, Grossman M, "The Grossman-Cormack Glossary of Technology-Assisted Review", Federal Courts Law Review, Vol. 7", Issue 1, 2013.

Arguably, the best two descriptions of the study appear to be the following:

> *"In a seminal study on the efficacy of human document review, humans thought they were retrieving 75% of the relevant documents, when in reality they were retrieving less than 20% of the relevant documents."* [38]

> *"The seminal study evaluating the success of text retrieval methods in an e-discovery setting was conducted by Blair & Maron over 20 years ago [2]. That study identified a serious gap between the perception on the part of lawyers that using keywords they would retrieve on the order of 75% of the relevant evidence to be found in a collection of 40,000 documents gathered for litigation purposes, whereas the researchers were able to show on the basis of using additional keywords that only 20% of relevant documents had in fact been found. These results have been replicated in other settings, and from them we can conclude that estimating recall is simply a hard task for people to do."*[39]

Although subsequent studies question the meaning and even the reliability of the paper, and even though there are serious differences between the manner in which search terms were used in the STAIRS system versus the manner in which they can be (have been) used in modern systems, Blair & Maron (1985) continues to be cited for the proposition that search terms can only be effective 20% of the time.[40]

Ironically, while we should reject the notion that the STAIRS study is the high water mark for search terms, the actual recall was probably worse than 20%. The authors later opined that their *recall* estimates were actually the maximum upper threshold for *recall*. In other words, they felt that the actual *recall*, had every document been manually examined rather than estimated using their estimation method, would likely have been lower.[41]

Results Will Vary

In the real world, litigation driven information retrieval is a complicated equation. It relies not only on *precision* and *recall*, but also on cost, complexity, resources, need, time constraints, reasonability and proportionality.

---

[38] Sonya Sigler, Cataphora, Inc., "Are Lawyers Being Replaced by Artificial Intelligence? Beyond Keyword Search: An Introduction to Advanced Search & Retrieval Technologies", May 2009.

[39] Feng C. Zhao, U. Washington, and Douglas W. Oard, University of Maryland & Jason R. Baron, National Archives and Records Administration, "Improving Search Effectiveness in the Legal E-Discovery Process Using Relevance Feedback", May 2009.

[40] See *Global Aerospace v. Landow Aviation*, No. CL 61040 (Va. Cir. Ct., Loudon County April 23, 2012).

[41] One conclusion of the STAIRS study was implicit in the reported evaluation, but deserves to be made explicit. That is, the value for Recall, although low, represents a maximum value because it was based on estimating Recall for small subsets of the document collection, not the entire database. If we examined the entire database we probably would have found more unretrieved relevant documents. The "actual" value for Recall, if it could be calculated, would be significantly lower. Blair and Maron (1990).

As a result, despite the strong rhetoric and the anecdotal stories, the results that are achievable with various technologies appear to vary according to the user, the facts, the corpus, the stage in the litigation lifecycle, the goals and, of course, the process employed. This appears particularly true with search terms.

For one, the process is important. There has been much published by EDRM about the advantages of non-linear review and iterative processes that result in better results.[42]

For another, there is not currently a standard value for *recall* in U.S. litigation. In a recently published opinion in Indiana Federal Court, the Court cited the *recall* that Biomet achieved using search terms against a corpus of over 19m (million) documents. In that matter, when one does the math and examines the worst possible scenario and the best possible scenario based on the estimated numbers of responsive documents retrieved by search terms and the estimated numbers of responsive documents not retrieved, search terms had a *recall* rate between 47% and 73%.[43]

Why does this matter?

With the advent of technology-assisted-review (TAR), why should we care about search terms? Why should we take time today to re-assess a study from 30 years ago?

The first simple answer is because we still use search terms in almost every matter that involves ESI in litigation. They are used to collect by custodian. They are used to filter by date ranges. They are used to cull broad collections of documents from archives. They are used to find documents to train TAR systems. They are used in numerous ways to deal with documents and processes in the litigation lifecycle. And, they appear to be here to stay.

The second simple answer is that when studies such as Blair & Maron (1985) are not put into proper context, the uninformed reader can develop an unreasonable bias against the use of search terms. This may then lead to unnecessary suspicions, delays, confusion or motion practice. This manifests as a bias against search terms when it should be a bias against a non-principled approach to search term development.

That was then, this is now

There are many differences that become apparent when comparing the research of Blair & Maron (1985) to the development and application of search terms today.

---

[42] http://www.edrm.net/resources/diagram-elements

[43] *In Re: Biomet M2a Magnum Hip Implant Products Liability Litigation (MDL 2391)*, Order Regarding Discovery of ESI, April 18, 2013.

These differences are one of the reasons that the study is so often incorrectly applied to modern e-discovery processes.

While this list is not exhaustive, consider the following:

- The data corpus
- The examination of false negatives
- The measurement of recall
- The complexity of the search terms
- The methodology used to develop search terms

The Data Corpus

We have little information on how the corpus of documents used in Blair & Maron (1985) was created. We are told that it consists of 40,000 documents used in the defense of a large corporate lawsuit. Although the complete manner of document selection was not disclosed, it appears that the documents had already been preserved, collected, and hand selected for litigation purposes (i.e. culled):

> *[T]he documents STAIRS provided access to were personally selected (from a larger set of documents) by the two lawyers and two paralegals who participated in the study. This selection process spanned a 1-year period and produced a set of documents that were all germane to the various issues in the lawsuit.*[44]

The application of search terms to a hand picked document set to identify narrow issues for defense purposes is fundamentally different than their application to preservation, collection, culling, or even review for production purposes. Since these are all different exercises with potentially different emphases on *precision* and *recall*, we should be careful when comparing the use of search terms (or any technology or methodology) in one stage of discovery to the use in another stage of discovery.

The Examination of False Negatives

False negatives are those documents that would be considered responsive to an information request, but are not labeled responsive by the actual query executed. These are incorrect negative labels or "false negatives".

Typically, in a modern, iterative development process, false negatives are examined to understand why a set of search terms might be missing a known responsive document. This technique, when available, can be used to identify new query strings

[44] Blair (1996) at 17.

that will add to the overall *recall* of a set of search terms. This is often used to calibrate search terms and can lead to *recall* improvement.[45]

At the end of the experiment, the controllers sampled the document set and examined false negatives. During the experiment, however, the research attorneys *did not* have access to false negatives.

Without this access, the technique used by the researchers should not be compared to modern exercises where such false negative are considered.

The Measurement of Recall

The research attorneys had prior familiarity with the suit and the documents. They had the advantage of examining the search term results. They had the ability to make an assessment of the *precision* of a given search merely by examining the search results. Because they did not have even an estimate of the total number of issue responsive documents in the total corpus, however, they could not calculate the *recall*.

As such, any subjective estimate of the *recall* of their searches can only be that – an estimate. Further, the estimates in Blair & Maron (1985) appear to be bad estimates.

Blair & Maron estimated their recall in the following manner. They felt that they needed 75% of the documents for their purposes, they searched until they had enough documents for their purposes, and then assumed that they had 75% of the documents. This can be analogized to fishing in a pond with an unknown stock of fish. The fisherman assumes that he needs to catch 75% of the fish in order to prepare dinner. He catches enough fish for dinner and, therefore, assumes that he caught 75% of the fish. But there is zero basis for that assumption. He has no independent measure or estimate of the actual number of fish in the pond originally (the universe) or remaining (the null set).

Perhaps the best take-away from Blair & Maron (1985) is the statement by Sigler (2009) and of Zhao, Oard and Baron (2009): "estimating recall is simply a hard task for people to do."

This may be one of the reasons that the TREC Legal Track and other information retrieval studies have focused on learning how to measure *precision* and *recall* (*but see* David Blair, "Some Thoughts on the reported results of TREC", Elsevier Science Ltd., 2002).

[45] The research lawyers in Blair & Maron did consider *false positives* by default every time they reviewed a retrieved set of documents. This technique is useful for improving precision. Blair & Maron reported an average 79% precision in their study.

The Complexity of the Search Terms

One of the biggest challenges in comparing one technique to another or one technology to another is that there are so many variables to the equation. One often-overlooked variable in the STAIRS study is the complexity of the search terms used.

A simple search is a Boolean search string that has no operator or extension. It is a stand-alone term. Standard modern operators include AND, OR, NOT, parentheses "(" or ")", proximity such as WITHIN or NEAR, and multi-character "*" or single multi-character "?". A complex search is a Boolean search string that does have an operator or an extension.[46]

From various sources, we know that the IBM STAIRS system was capable of complex searches.[47, 48] But, while they did not publish a list of the actual terms, it can be inferred from subsequent publications by Maron (1988) and Blair & Maron (1990) that the original study used fairly simple terms. As the original authors wrote four years later in 1989 to support the proposition that it would be difficult to do better than the *recall* in their 1985 study using a "simple full-text retrieval model" with a large document database:

> *In order for a simple full-text system to retrieve effectively, the user/searcher must be able to predict (and use as his query terms) those words, phrases and word combinations that occur in most of the relevant documents, and which do not occur in most of the non-relevant documents. (See also Maron, 1988.) If a searcher can construct such a query, we shall call that an "effective query." We see that there are two interrelated parts to an effective query; predicting A, the words, word combinations, etc., that occur in the relevant documents and then B, reducing that set of terms by excluding those word or word combinations which are likely also to occur in non-relevant documents.*[49]

The decision to use simple or complex Boolean search terms is governed by the facts and needs of a given case and the nature of documents being searched. Yet it stands to reason and common experience that the *ad hoc* development of complex search terms is a difficult proposition. In practice, longer and more complex search strings,

---

[46] If a set of Boolean search terms has more than a single search term, there is already an implicit "OR" operator in that list.

[47] The full name is **STAIRS/TLS** ("**ST**orage **A**nd **I**nformation **R**etrieval **S**ystem / **T**hesarus **L**inguistic **S**ystems". It provided for exact match, proximity (adjacent, within sentence, within paragraph), expansion ("narrower than", "broader than", "related to", "synonymous with"), hit frequency, and "automatic phrase decomposition". Blair & Maron (1985), at 291.

[48] Bourne, Hahn, "A History of Online Information Services: 1963 – 1976", Massachusetts Institute of Technology, 2003, p. 130-132. (Detailing the development of the AQUARIUS and re-branded STAIRS systems).

[49] Blair D.C., Maron M.E., "Full-Text Information Retrieval: Further Analysis and Clarification", Information Processing & Management, Vol. 26, No. 3, pp 437-477, 1990.

and sets of search strings, require an algorithm, a discipline, a methodology to be done properly.

The Methodology Used to Create Search Terms

The methodology used to create search terms has been recognized as a potential source of success or failure:

> *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. Magistrate Judge John Facciola (2008).*[50]

> *This Opinion should serve as a wake-up call to the Bar in this District about the need for careful thought, quality control, testing, and cooperation with opposing counsel in designing search terms or "keywords" to be used to produce emails or other electronically stored information ("ESI"). Magistrate Andrew Peck (2009).*[51]

This was not a new idea in 2008 or 2009. Eero Sormunen proposed a systematic method of developing high recall search strings in 2000 and 2001. He describes a method of creating more complex searches ("optimal queries") in a paper studying the trade-off of precision and recall using "query tuning". Sormunen studied precision and recall in a larger database (52,000 documents) than Blair & Maron (1985) (40,000 documents). He also studied the effect of search terms on a "large & sparse database" and on a "large and dense database".

By using query optimization, which included the examination of false negatives and the iterative, but methodical, modification of search strings, Sormunen was able to create "optimal queries" that achieved 99.3% recall.[52]

Again, the TREC Legal Track studies and the contribution of H5 demonstrated that a sound methodology based on linguistic and document analysis provides the ability to choose, in a principled manner, appropriate search elements. In addition, a sound methodology provides the the ability to choose, in a principled manner, appropriate

---

[50] United States v. O'Keefe, 537 F. Supp. 2d 14 (D.D.C. 2008).

[51] William A. Gross Constr. Assocs. v. Am. Mfrs. Mut. Ins. Co., 2009 U.S. Dist. LEXIS 22903 (S.D.N.Y. Mar. 19, 2009).

[52] 2001, Sormunen E, "Extensions to the STAIRS Study – Empirical Evidence for the Hypothesized Ineffectiveness of Boolean Queries in Large Full-Text Databases", Information Retrieval 4(3/4):257-274.

operators and operator distances when combining search elements into complex searches that lead to high performance.[53]

More recently, Ralph Losey, a well- respected author and thought leader in e-discovery, gave us a modern restatement of both Blair & Maron (1985) and Sormunen (2001) when he wrote, "[keyword search] 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."[54]

Although Blair & Maron (1985) often emphasizes the *efforts* expended by the research attorneys, the STAIRS study does not describe any particular methodology used by the research attorneys in developing their queries.

> *In the STAIRS study, we permitted the searchers to revise their original queries as many times as they liked, and to search until they believed that they had retrieved all the documents they wanted. No search consisted of a single query, and many queries went through 10 or more revisions, gathering more relevant documents during each iteration.*[55]

Only after the initial work was complete did the researchers use a variation of multi-modal searching while performing a post-experiment estimate of *recall*:

> *To measure recall, the experimenters used systematic generalizations of the queries to search promising subsets of the un-retrieved documents. Two primary techniques were used. First , the Boolean connectives of existing queries were modified to produce alternative but similar retrieval sets. For example, if an initial query had been "A and B and C", they would use the three modified queries, "A and B and not C", "A and C and not B", and "B and C and not A". Second, synonyms were substituted for the search terms to form related queries. In another, other parts of the collection were randomly selected. The experiment was conducted in such a way that the lawyers were unaware whether the retrievals they were evaluating were the result of their paralegals' queries, or of the experimenters' modified samples.*[56]

Thus, while not using the exhaustive rigor described by Sormunen, even a fairly simple multi-model methodology of mixing up search queries resulted in the identification of a greater number of relevant documents.

---

[53] See also, Brassil, Hogan, Attfield, "The Centrality of User Modeling to High Recall with High Precision Search", IEEE International Conference on Systems, Man and Cybernetics, 2009.

[54] Losey R, 2011, "Secrets of Search – Part One".

[55] Blair (1996), at 15.

[56] Rose, "A Symbolic and Connectionist Approach to Legal Information Retrieval", Lawrence Erlbaum Associates, Inc., 1994, at 73.

To What Purpose "Recall"?

What if recall was never estimated? Would that have made the exercise a failure? In the STAIRS study, the experienced attorneys felt that they had enough documents for each of the 51 issues to try those issues:

> *In this case, the lawyers who were to use the system for litigation support stipulated that they must be able to retrieve at least 75 percent of all the documents relevant to a given request for information, and that they regarded this entire 75 percent as essential to the defense of the case... the information-request and query-formulation procedures were considered complete only when the lawyer stated in writing that he or she was satisfied with the search results for that particular query (i.e., in his or her judgment, more than 75 percent of the "vital," "satisfactory," and "marginally relevant" documents had been retrieved)... the lawyers were encouraged to continue requesting information from the database until they were satisfied they had enough information to defend the lawsuit on that particular issue or query.*[57]

While this list of differences is not intended to be exhaustive, it is instructive. It seems that Blair & Maron (1985) was really an exercise for experienced attorneys to find enough documents for them to defend the lawsuit and not necessarily designed to produce to opposing counsel.

Unfortunately, no list of the actual searches employed in Blair & Maron (1985) is available. It is unknown if they used wildcard or fuzzy search. There is some literature that suggests that STAIRS did have a basic form of proximity search.[58] Based on this, it seems that Blair & Maron (1985) was limited to whole-word exact matches with limited proximity – what might be referred to as "simple keyword searches." They also concluded that making compound searches to increase *recall* can have the effect of decreasing *precision*.

Sormunen (2001) concluded that Blair & Maron (1985) were right in observing that *recall* tends to decrease and *precision* tends to increase as Boolean search strings become more complex. It is generally accepted in information retrieval that there is a trade-off between *precision* and *recall*. This was documented by Michael Buckland and Fredric Gey in 1994.[59] Buckland and Gey (1994) stated that "[w]ith very large databases and/or systems with limited retrieval capabilities there can be advantages to retrieval in two stages: initial retrieval emphasizing high Recall,

---

[57] Blair D, Maron ME (1985) at 291.

[58] STAIRS queries were formulated as *Boolean expressions* of desired terms. In addition to the normal Boolean functions of AND, OR, and NOT, STAIRS recognized such modifiers as *adjacent to* or *in the same paragraph as*. Plain text documents could also contain so-called *formatted fields*, which could be used for additional selection. These might contain fixed information such as a date or state name. Wikipedia, May 10, 2013. http://en.wikipedia.org/wiki/IBM_STAIRS.

[59] Buckland M, Grey F, "The Relationship between Precision and Recall", Journal of the American Society for Information Science, 45(1):12-19, 1994.

followed by more detailed searching of the initially retrieved set, can be used to improve both Recall and Precision simultaneously."[60]

This was, essentially, exactly the technique used in *Kleen Products*, *Biomet*, and in other cases where search terms are used to cull the initial data set and some other techniques (additional queries, computer-assisted-review, or responsiveness review) are used for the next stage of review or for fine-tuning.

If there is no subsequent fine-tuning step, however, parties may attempt to maximize both *precision* and *recall* in a single step. These competing interests, as well as human nature, affect success. Blair & Maron (1985) examined this, as well. They described the basis of the low recall they observed as related to *prediction criteria (PC)*, *futility point criteria (FPC)*, and the *anchoring effect*.[61, 62, 63]

At the risk of repetition: a party seeking documents for litigation production may be more focused on recall, while a party seeking documents for their own use may be more interested in precision.

This is exactly the scenario in Blair & Maron (1985). In that study, the researchers (case experienced defense attorneys) were not seeking to fulfill a legal obligation to produce documents to opposing counsel (high *recall*), but rather were seeking only to find sufficient highly relevant documents to defend their case (high *precision*). If that is the case, then a methodology that resulted in high *precision*, but not high *recall* may have been sufficient for their actual needs, regardless of their unsupported estimates of *recall*.

In fact, it is often the case that parties end up needing fewer documents for actual merits resolution than for production in discovery.

In 2011, Microsoft Corporation ("Microsoft") submitted a letter to the Advisory Committee on Civil Rules. In that letter, Microsoft detailed their internal studies of preservation, collection, review, production and usage of exhibits in trial, based on the then current litigation portfolio of 329 matters and 14,805 active holds. Microsoft found that their average documents counts were as follows:

Action	Document count	Percentage
Preserved	48,431,250	100.0000%
Collected	12,915,000	26.6667%
Reviewed	645,750	1.3333%
Produced	141,450	0.2921%
Used as Exhibits	142	0.0003%

[60] Id at 12.
[61] Blair & Maron (1985).
[62] Blair (1990).
[63] Somunen (2001).

This means that Microsoft observed that about 0.1% of documents produced were actually used as exhibits.[64] This correlates with another study finding that Fortune 200 companies experience a 1,000:1 ratio (i.e. 0.1%) for documents produced versus documents used as exhibits.[65]

It is not speculative to say that the documents that the attorneys from Blair & Maron (1985) found were "sufficient". That is exactly how the research attorneys described them. As David Blair explained a decade later:

> *We also left the endpoint of the search up to the lawyers. We simply told them to search until they found enough useful documents, in their estimation, to conduct the defense of the lawsuit. Since the lawyers were the ones who would defend the lawsuit, they were naturally the only people to know when to stop.*[66]

As a result, it becomes clear that the goal of the researchers in Blair & Maron (1985) (i.e., the defense of the lawsuit using culled documents) may vary considerably from the goals of a party applying search terms for the purpose of preservation, collection, culling or the production of documents for a given litigation.

Conclusion

Once all of the various aspects of Blair & Maron (1985) are studied and compared to what we see parties actually doing in litigation today, it becomes clear that while we can learn a lot from this early study, we cannot compare the results achieved in this study almost 30 years ago with the potential results in a different matter today, using different tools and different techniques.

However, we can draw some reliable observations:

- Estimating recall is simply a hard task for people to do.[67]
- There are information retrieval techniques that can be used to calibrate our tools and techniques and to avoid the guesswork of developing search terms.
- Not all search terms are created equal.
- Not all search methodologies are created equal.
- For their own use, attorneys may stop and may be satisfied with less than 100% or even less than 75% of the relevant documents.

---

[64] Howard, Palmer and Banks, 2011, http://www.bricker.com/documents/attachments/microsoft.pdf, at 4.

[65] "Litigation Cost Survey of Major Companies", submitted by Civil Justice Reform Group, Lawyers for Civil Justice, U.S. Chamber Institute for Legal Reform, 2010 Conference on Civil Litigation, Duke Law School, 2010. http://www.uscourts.gov/uscourts/RulesAndPolicies/rules/Duke%20Materials/Library/Litigation%20Cost%20Survey%20of%20Major%20Companies.pdf

[66] Blair D, "STAIRS Redux: Thoughts on the STAIRS Evaluation, Ten Years after", Journal of the American Society for Information Science 47(1):4-22, 1996.

[67] Roitblat, "Search and Information Retrieval Science", Sedona Conference Journal, Fall 2007.

- The fact that the Blair & Maron (1985) attorneys *did not* find more documents cannot be equated with the statement that they *could not* find more documents. They merely stopped when they had *enough for their purposes*.

As we stand on the eve of a new round of amendments to the Federal Rules of Civil Procedure, we should expect the new rules to encourage more decisions of discovery based on proportionality. This may also invite more review and criticism of techniques used in discovery – including the use of search terms.

In general, many techniques, technologies and processes are candidates for review and even criticism. But that review and that criticism should stand – in fact must stand - on the merits of the actual results, rather than association with Blair & Maron (1985).

Otherwise, the comparison is often clouded by the many differences between how, when and why that exercise was conducted in 1985 compared to how, when and why search terms are deployed today.

The use of faster computers, cheaper storage, improved Boolean capabilities, now-standard information retrieval measurement techniques, and robust methodologies dramatically improves the ability to develop and apply search terms in a litigation context. However, you still have to exercise that ability. As the Sedona Conference has put it,

> *"There is no magic to the science of search and retrieval: only mathematics, linguistics, and hard work."*[68]

We should not ignore or reject Blair & Maron (1985). Instead, we should recognize the study for what it actually represents and acknowledge the authors for heralding the use of information retrieval science in a litigation context. Conversely, we should not draw unsupported conclusions, ill-matched comparisons or performance absolutes from their work of 30 years ago.[69]

---

[68] The Sedona Conference Best Practices Commentary on the Use of Search & Information Retrieval Methods in E-Discovery, August 2007, Public Commentary Version, The Sedona Conference Journal, 2007, at 208.

[69] Blair & Maron (1990), at 446, ("One of the important points of the STAIRS study […] was, in a large part, a conscious attempt to raise the standards and methodological rigor of information retrieval evaluations to a level comparable to other more established empirical disciplines.")