PAPER:

TOWARD A NEW JURISPRUDENCE OF INFORMATION RETRIEVAL: WHAT CONSTITUTES A "REASONABLE" SEARCH FOR DIGITAL EVIDENCE WHEN USING KEYWORDS?

By **Jason R. Baron**

The December 2006 amendments to the Federal Rules of Civil Procedure have highlighted the importance of digital evidence in U.S. civil litigation by expressly recognizing a new term, 'electronically stored information' or 'ESI.'[1]

References incorporating the ESI term are now to be found in a variety of discovery contexts, including requests for production, interrogatories, pre-trial conferences, and what are known as initial 'meet and confers', where parties are expected to discuss all manner of issues surrounding the preservation of, formatting of, and access to digital evidence in their respective possession and control that might be relevant at a later stage of legal proceedings.[2] Under these new rules, both the expectations of judges and the behaviour of parties and their counsel are changing, including how parties are approaching the task of searching for relevant digital evidence in response to discovery obligations. In a remarkable series of recent court decisions, judges have questioned, challenged, and even applied sanctions against parties for their failure to act reasonably in conducting what are known as 'keyword' searches for ESI. This is a new development in the law, with far-reaching implications not only on the civil side of practice, but potentially also with respect to how digital evidence is gathered in at least certain forms of criminal proceedings in the U.S. as well.

Only a few years ago, the idea that there would be a jurisprudence devoted to analyzing the strengths and limitations of keyword searching would be unheard of, for at least two reasons. First, until recent times, discovery practice, even in the largest and most complex cases in the U.S., consisted entirely of paper productions of documents, sometimes in admittedly massive quantities. For example, just short of ten million documents in hard copy form have been amassed in a repository, pursuant to a Master Settlement Agreement, between a variety of State Attorneys General and the tobacco industry.[3]

Second, with the advent of proprietary computerized databases of case law, represented most prominently by Westlaw and Lexis, lawyers have become well versed in conducting keyword searches to find relevant case precedent for use in legal pleadings and briefs. The beauty of keyword searching in this context is that no

[1] *Fed. R. Civ. P 34(a), 2006 Advisory Note, in which the definition of ESI is said to be 'expansive and includes any type of information that is stored electronically. A common example often sought in discovery is electronic communications, such as e-mail. The rule covers — either as documents or as electronically stored information — information "stored in any medium," to encompass future developments in computer technology. [The Rule] is intended to be broad enough to cover all current types of computer-based information, and flexible enough to encompass future changes and developments.'*

[2] *Fed. R. Civ. P. 16, 26, 33, and 34.*

[3] *For example, see, People of the State of California v. Philip Morris, et al., Case No. J.C.C.P. 4041 (Sup. Ct. Cal.) (December 9, 1998 consent decree incorporating terms of Master Settlement Agreement or 'MSA'). These documents have for the most part been digitized in using Optical Character Recognition (OCR) technology, and are available online on various web sites. See the Legacy Tobacco Collection, available at http://legacy.library.ucsf.edu/. The OCR portions of the MSA collection have been used in conjunction with the TREC Legal Track.*

lawyer wishes or needs to read more than a handful of cases as an aid in stating a legal position in writing, except in the rare instance where more exhaustive searches are necessary to be performed. In contrast, the limitations of keyword searching become more apparent as the task changes from finding case precedent to finding 'all' relevant evidence related to the discovery topic at hand. This has become especially clear with the exponential growth of databases and data stores of all kinds, especially with respect to electronic mail.[4]

The Sedona Conference's commentary on search and information retrieval methods explains the strengths and weaknesses of keyword searching this way:

> Keyword searches work best when the legal inquiry is focused on finding particular documents and when the use of language is relatively predictable. For example, keyword searches work well to find all documents that mention a specific individual or date, regardless of context. However, . . . the experience of many litigators is that simple keyword searching alone is inadequate in at least some discovery contexts. This is because simple keyword searches end up being both over- and under-inclusive in light of the inherent malleability and ambiguity of spoken and written English (as well as all other languages).[5]

The Sedona Search Commentary describes how keywords have the potential to miss documents that fail to contain the word either because other terms with the same meaning have been used, or due to common or inadvertently misspelled instances of the keyword term. The Commentary then goes on at length to describe a variety of alternative search methods that exist, starting with the use of Boolean operators to construct sophisticated search strings, to the use of fuzzy logic, statistical techniques, and taxonomies and ontologies.[6] The Commentary makes the point that lawyers 'are beginning to feel more comfortable' using these forms of alternative search tools, based on anecdotal evidence from a small (but increasing) number of companies and law firms.[7]

Although the limitations of keyword searching have been well-known to the library and information science communities for decades, and while as early as 1985 the Blair & Maron study[8] underscored the disconnect between what lawyers believe they retrieve and what is actually retrieved when using keywords,[9] nevertheless, U.S. jurisprudence has only recently begun to grapple with the problems inherent in the task of conducting a reasonable search for all the needles in what turn out to be very large e-haystacks.

Thus, in only a short interval of time some courts have gone from extolling the power of keyword searching to questioning its efficacy (at least as articulated by counsel). Compare the case of *In re Lorazepam & Clorazepate Antitrust Litigation*, 300 F. Supp. 2d 43, 46 (D.D.C. 2004) ('[t]he glory of electronic information is not merely that it saves space but that it permits the computer to search for words or 'strings' of text in seconds,' to *U.S. v. O' Keefe*, 537 F.Supp.2d 14, 24 (D.D.C. 2008):

> Whether search terms of '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.

Until mid-2007, the overarching approach taken by a number of courts in this area has been to define the reasonableness of the search conducted by a party solely in terms of the number of keyword terms being requested as well as their relevance to the subject at hand. Thus, in the case of *In re Lorazepam*, the district court endorsed the employment of a number of search terms as a reasonable means of narrowing the production for relevant ESI. In another case, as few as four keyword search terms were found to be sufficient.[10] In certain decisions, the court has ordered a responding party (usually the defendant) to conduct searches using the keyword terms provided by plaintiff.[11] In other cases, judges that have taken a more activist approach have attempted to force parties to cooperate on reaching an

[4] *George L. Paul and Jason R. Baron, 'Information Inflation: Can The Legal System Adapt?,' 13 Richmond J. of Law & Tech. 10 (2007), available at http://richmond,edu/jolt/v13i3/article10.pdf.*

[5] *The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery (Sedona Search Commentary), 8 Sedona Conf. J. 189, 201 (2007), available at http://www.thesedonaconference.org/publications. See also George L. Paul and Jason R. Baron, 'Information Inflation: Can The Legal System Adapt?,' 13 Richmond J. of Law & Tech. 10 (2007), n. 4.*

[6] *Sedona Search Commentary, 8 Sedona Conf. J. at 202-03; 217 (Appendix describing alternative search methods at greater length).*

[7] *Sedona Search Commentary, 8 Sedona Conf. J. at 202-03.*

[8] *David C. Blair and M. E. Maron, 'An evaluation of retrieval effectiveness for a full-text document-retrieval system,' Communications of the ACM, 289 (1985) (discussed at further length in Sedona Search Commentary, 8 Sedona Conf. J. at 206.*

[9] *Sedona Search Commentary, 8 Sedona Conf. J. at 204-06.*

[10] *J.C. Associates v. Fidelity & Guaranty Ins. Co., 2006 WL 1445173 (D.D.C. 2006).*

[11] *For example, see Medtronic Sofamor Danck, Inc. v. Michelson, 229 F.R.D. 550 (W.D. Tenn. 2003).*

agreement for a reasonable search protocol, including that the use of certain search terms.[12]

On June 1, 2007, U.S. Magistrate Judge John Facciola issued an opinion in the case of *Disability Rights Council of Greater Washington v. Metropolitan Transit Authority*,[13] in which for the first time in published case law a judge suggested that parties contemplate the use of an alternative to merely reaching a set of keywords by consensus. The dispute in question involved disabled individuals and an advocacy group bringing an action against a local transit authority alleging that inadequacies in para-transit services amounted to disability discrimination. The plaintiffs moved to compel the production of electronic documents residing on backup tapes in the defendants' possession. After engaging in a routine balancing analysis of the considerations set out in Rule 26(a), the court ordered that some form of restoration of the backup tapes be ordered to recover relevant documents. It is at this juncture that the opinion breaks new ground: Facciola J expressly required that counsel meet and confer and prepare for his signature a 'stipulated protocol' as to how the search of the backup tapes would be conducted, and pointed out 'I expect the protocol to speak to at least the following concerns,' including both 'How will the backup tapes be restored?', and

> Once restored, how will they be searched to reduce the electronically stored information to information that is potentially relevant? In this context, I bring to the parties' attention recent scholarship that argues that concept searching, is more efficient and more likely to produce the most comprehensive results.[14]

Following this decision and the publication of the Sedona Search Commentary in August 2007, was the case of *U.S. v. O' Keefe*,[15] another decision written by Facciola J, including a discussion on the use of search protocols. The *O' Keefe* case involved the defendant being indicted on the charge that as a State Department employee living in Canada, he received gifts and other benefits from his co-defendant, in return for expediting visa requests for his co-defendant's company employees. The district court judge in the case had previously required that the government 'conduct a thorough and complete search of both its hard copy and electronic files in a good faith effort to uncover all responsive information in its possession custody or control.'[16] This in turn entailed a search of paper documents and electronic files, including for e-mails, that 'were prepared or received by any consular officers' at various named posts in Canada and Mexico, 'that reflect either policy or decisions in specific cases with respect to expediting visa applications.'[17]

The defendants insisted that the government search both active servers and certain designated backup tapes. The government conducted a fairly well-documented search, as described in a declaration placed on file with the court, in which 19 specific named individuals were identified as being within the scope of the search, along with certain identified existing repositories by name and the files of at least one former member of staff. The declarant went on to describe the search string used was as follows:

> 'early or expedite* or appointment or early & interview or expedite* & interview.'[18]

Upon review of the results, only those documents 'clearly about wholly unrelated matters' were removed, for example, 'emails about staff members' early departures or dentist appointments.' Nevertheless, the defendants objected that the search terms used were inadequate. This led Facciola J to state that on the record before him, he was not in a position to judge whether the search was reasonable or adequate, and that given the complexity of the issues he did not wish 'to go where angels fear to tread.' He went on to note, citing to the use of 'expert' testimony under Federal Rule of Evidence 702:

> This topic is clearly beyond the ken of a layman and requires that any such conclusion be based on evidence that, for example, meets the criteria of Rule

[12] *Treppel v. Biovail, 233 F.R.D. 363, 368-69 (S.D.N.Y. 2006) (court describes plaintiff's refusal to cooperate with defendant in the latter's suggestion to enter into a stipulation defining the keyword search terms to be used as a 'missed opportunity,' and goes on to require that certain terms be used); see also Alexander v. FBI, 194 F.R.D. 316 (D.D.C. 2000) (court places limitations on the scope of plaintiffs' proposed keywords in a case involving White House e-mail).*

[13] *2007 WL 1585452 (D.D.C.).*

[14] *Disability Rights Council of Greater Washington v. Metropolitan Transit Authority, 2007 WL 1585452 (D.D.C.), 242 F.R.D. at 148 (with reference to George L. Paul and Jason R. Baron, 'Information Inflation: Can The Legal System Adapt?,' 13 Richmond J. of Law & Tech. 10 (2007), n. 4). In contrast to keyword searching, which relies on set-based searching using simple keywords or word combinations, with or without Boolean and related operators (such as 'and,' 'or,' '!'), concept searching involves language modeling and/or the use of probabilistic techniques to find relevant documents that nevertheless may not anywhere in them have the arbitrarily selected 'keyword.' See George L. Paul and Jason R. Baron, 'Information Inflation: Can The Legal System Adapt?,' 13 Richmond J. of Law & Tech. 10 (2007), n. 4, at 42-43.*

[15] *537 F.Supp.2d 14, 24 (D.D.C. 2008).*

[16] *537 F. Supp. 2d at 16 (quoting U.S. v. O' Keefe, 2007 WL 1239204, at *3 (D.D.C. April 27, 2007)) (internal quotations omitted).*

[17] *537 F. Supp. 2d at 16.*

[18] *Based only on what is known from the opinion, it is admittedly somewhat difficult to parse the syntax used in this search string. One is left to surmise that the ambiguity present on the face of the search protocol may have contributed to the court finding the matter of adjudicating a proper search strong to be too difficult a task.*

> 702 of the Federal Rules of Evidence. Accordingly, if defendants are going to contend that the search terms used by the government were insufficient, they will have to specifically so contend in a motion to compel and their contention must be based on evidence that meets the requirements of Rule 702 of the Federal Rules of Evidence.[19]

Whether it is the view of Facciola J that expert opinion testimony must be introduced in all cases on the subject of the reasonableness of the search method or protocol employed immediately generated discussion in subsequent case law and commentary.[20]

However, another remarkable aspect of the opinion has not been widely commented upon, namely, that the court chose to look to federal civil litigation practice as guidance as to how to conduct discovery in a criminal case, by essentially importing novel and emerging e-discovery best practices into the criminal law arena. As stated in *O' Keefe*:

> In criminal cases, there is unfortunately no rule to which the courts can look for guidance in determining whether the production of documents by the government has been in a form or format that is appropriate. This may be because the 'big paper' case is the exception rather than the rule in criminal cases. Be that as it may, Rule 34 of the Federal Rules of Civil Procedure speak specifically to the form of production. The Federal Rules of Civil Procedure in their present form are the product of nearly 70 years of use and have been consistently amended by advisory committees consisting of judges, practitioners, and distinguished academics to meet perceived deficiencies. It is foolish to disregard them merely because this is a criminal case, particularly where, as in the case here, it is far better to use these rules than to reinvent the wheel when the production of documents in criminal and civil cases raises the same problems.
>
> This line of analysis opens up the possibility that there may emerge a jurisprudence of best practice on the subject of keyword searching, and its alternatives may yet find fertile ground in connection with the U.S. criminal docket.[21]

Most recently, U.S. Magistrate Judge Paul Grimm has substantially contributed to the development of a jurisprudence of information retrieval, through issuance of a comprehensive opinion on the subject of privilege review in *Victor Stanley, Inc. v. Creative Pipe, Inc.*[22] At issue was whether the manner in which privileged documents were selected from a larger universe of relevant evidence was sufficient to protect a party from waiver of attorney-client privilege, where 165 privileged documents were provided to the opposing counsel as the result of a keyword search. At the outset, Judge Grimm reported that 'he ordered the parties' computer forensic experts to meet and confer in an effort to identify a joint protocol to search and retrieve relevant ESI' in response to the plaintiff's document requests. The protocol 'contained detailed search and information retrieval instructions, including nearly five pages of keyword/phrase search terms.'[23]

The defendants' counsel subsequently informed the court that they would be conducting a separate review to filter privileged documents from the larger [universe] of 4.9 gigabytes of text-searchable files and 33.7 gigabytes of non-searchable files. In doing so, they claimed to use seventy keywords to distinguish privileged from non-privileged documents; however, Judge Grimm, applying a form of heightened scrutiny to the assertions of counsel, found that their

[19] *537 F. Supp. 2d at 24.*

[20] *Equity Analytics v. Lundin, 248 F.R.D. 331 (D.D.C. 2008) (stating that in O' Keefe '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,' Judge Facciola requires an expert to describe scope of proposed search); See the later discussion of Victor Stanley, Inc. v. Creative Pipe, Inc. 2008 WL 2221841 (D. Md. May 29, 2008).*

[21] *The digital evidence burden of production on the government, as outlined in O' Keefe, arose under Rule 16(a)(1)(E) of the Federal Rules of Criminal Procedure, which states in relevant part that the government 'must permit the defendant to inspect and to copy . . . papers, documents, data, . . . or copies or portions of any of these items, if the item is within the government's possession, custody, or control and (i) the item is material to preparing the defense; [or] (ii) the government intends to use the item in its case-in-chief at trial . . .*
The propriety of keyword searching has also arisen in the very different context of 'search and seizure' criminal law under the Fourth Amendment to the U.S. Constitution, which provides for the right of the people to be secure in their 'papers, and effects, against unreasonable searches and seizures.' The Fourth Amendment has a 'specificity requirement' that 'prevents officers from engaging in general, exploratory searches by limiting their discretion and providing specific guidance as to what can and cannot be searched and seized.' U.S. v. Adjani, 452 F.3d 1140, 1147 (9th Cir. 2006). In Adjani, the Court of Appeals held that seizure of an entire computer, as opposed to conducting a targeted search of its contents on the premises, not to be unreasonable, stating that '[t]o require such a pinpointed computer search restricting the search to an email program or to specific search terms, would likely have failed to cast a sufficiently wide net to capture the evidence sought'). The court went on to note that '[c]omputer files are easy to disguise or rename, and were we to limit the warrant to such a specific search protocol, much evidence could escape discovery because of [defendants'] labeling of the files' at 1150. A subsequent line of authority has generally reached a similar result. See U.S. v. Comprehensive Drug Testing, Inc., 473 F.3d 915 (9th Cir. 2006); U.S. v. Hill, 459 F.3d 966 (9th Cir, 2006). It remains to be determined in future case law whether further importation of 'best practices' standards from civil practice on the subject of keyword searching and its alternatives may yet serve to influence Fourth Amendment practice.

[22] *2008 WL 2221841 (D. Md. May 29, 2008).*

[23] *2008 WL 2221841 (D. Md. May 29, 2008) at *1.*

representations fell short of being sufficient for purposes of explaining why mistakes took place in the production of the documents and in so doing, avoiding waiver. In the court's words:

> [T]he Defendants are regrettably vague in their description of the seventy keywords used for the text-searchable ESI privilege review, how they were developed, how the search was conducted, and what quality controls were employed to assess their reliability and accuracy. . . . [N]othing is known from the affidavits provided to the court regarding their [the parties' and counsel's] qualifications for designing a search and information retrieval strategy that could be expected to produce an effective and reliable privilege review....
>
> [W]hile it is universally acknowledged that keyword searches are useful tools for search and retrieval of ESI, all keyword searches are not created equal; and there is a growing body of literature that highlights the risks associated with conducting an unreliable or inadequate keyword search of relying exclusively on such searches for privilege review.[24]

The opinion goes on to set out at length the limitations of keyword searching, and the need for sampling of the results of such searches, finding that there was no evidence that the defendant did anything but turn over all documents to plaintiff that were identified as the result of the keywords used as non-privileged. Later in the opinion, in several lengthy footnotes, Judge Grimm first goes on to describe what alternatives exist to keyword searching (including fuzzy search models, Bayesian classifiers, clustering, and concept and categorization tools), citing the Sedona Search Commentary,[25] and second, provides a mini-law review essay on the subject of whether Judge Facciola's recent opinions in *O' Keefe and Equity Analytics* should be read to require expert testimony under Federal Rule of Evidence 702 be presented to the finder of fact in every case involving the use of search methodologies. In Judge Grimm's view:

> Viewed in its proper context, all that *O' Keefe and Equity Analytics* required was that the parties be prepared to back up their positions with respect to a dispute involving the appropriateness of ESI search and information retrieval methodology – obviously an area of science or technology – with reliable information from someone with the qualifications to provide helpful opinions, not conclusory argument by counsel. . . . The message to be taken from *O' Keefe* and *Equity Analytics*, and this opinion is that when parties decide to use a particular ESI search and retrieval methodology, they need to be aware of literature describing the strengths and weaknesses of various methodologies, such as [the Sedona Search Commentary] and select the one that they believe is most appropriate for its intended task. Should their selection be challenged by their adversary, and the court be called upon to make a ruling, then they should expect to support their position with affidavits or other equivalent information from persons with the requisite qualifications and experience, based on sufficient facts or data and using reliable principles or methodology.[26]

The new case law on search and information retrieval thus amounts to a change in the way things were before, for both the bar and the bench: counsel has a duty to fairly articulate how they have gone about the task of finding relevant digital evidence, rather than assume that there is only one way to go about doing so with respect to ESI (for example, using keywords), even if the task appears to be a trivial or uninteresting one to perform. Arguably, the 'reasonableness' of one's actions in this area will be judged in large part on how well counsel, on behalf of his or her client, has documented and explained the search process and the methods employed. In an increasing number of cases, courts can be expected not to shirk from applying some degree of searching scrutiny to counsel's actions with respect to information retrieval. This may be greeted as an unwelcome development by some, but comes as an inevitable consequence of the heightened scrutiny being applied to all aspects of e-discovery in the wake of the newly revised Federal Rules of Civil Procedure.

Given the decisions in *Disability Rights, O' Keefe, and Creative Pipe*, it seems certain that in a few years' time there will be large and increasing jurisprudence discussing the efficacy of various search methodologies as employed in litigation. The Sedona Search Commentary has aimed to serve as a guide, and includes within it eight practice pointers for the legal

[24] *2008 WL 2221841 (D. Md. May 29, 2008) at *3.*
[25] *2008 WL 2221841 (D. Md. May 29, 2008) at n. 9.*
[26] *2008 WL 2221841 (D. Md. May 29, 2008) at n. 10.*

community to consider. In addition, the Sedona Search Commentary asks the question, 'What prospects exist for improving present day search and retrieval methodologies?'[27] This author's sense is that lawyers can improve upon their present-day search techniques through a number of strategies that do not necessarily rely in turn on improvements in search technology; these would include, but not be limited to, improving lawyers' knowledge of and ability to construct meaningful Boolean strings, and following Sedona's practice point guidance.

Beyond these observations, at least one continuing research project with an international dimension has shown promise in making advances in this general area. The Text Retrieval Conference (TREC) Legal Track, sponsored by the U.S. National Institute of Standards and Technology, has as its aim the evaluation of search methodologies used in a legal context.[28] Preliminary results from the first two years of the track have shown that a large gap exists between the number of relevant documents retrieved as the result of Boolean searches, and the number of relevant documents cumulatively found by other types of search methods used by information scientists participating in the track.[29] In its third year, an open call to the legal service provider community is expected to yield greater participation in TREC.[30]

As this and other research projects continue, the Sedona Search Commentary has gone on record as expressing two recommendations:

1. The legal community should support collaborative research with the scientific and academic sectors aimed at establishing the efficacy of a range of automated search and information retrieval methods.
2. The legal community should encourage the establishment of objective benchmarking criteria, for use in assisting lawyers in evaluating the competitive legal and regulatory search and retrieval services market.

It is certainly this author's hope that engaging in interdisciplinary research and establishing evaluative criteria may yet go a long way towards advancing the aim of the legal profession in seeking ways to ensure the 'just, speedy, and inexpensive' determination of every action, as provided for in Federal Rule of Civil Procedure 1.

Jason R. Baron is Director of Litigation at the U.S. National Archives and Records Administration, an Adjunct Professor, School of Information Studies, University of Maryland, and serves as Editor-in-Chief of The Sedona Conference® Best Practices Commentary on the Use of Search and Information Retrieval, and a founding coordinator of the TREC Legal Track.

jason.baron@nara.gov

[27] *Sedona Search Commentary, 8 Sedona Conf. J. at 212.*

[28] *http://trec-legal.umiacs.umd.edu/.*

[29] *Stephen Tomlinson, Douglas W. Oard, Jason R. Baron, and Paul Thompson, ' TREC-2007 Legal Track Overview,' 2007 Fifteenth Text Retrieval Conference (TREC 2007) Proceedings, available at http://trec-legal.umiacs.umd.edu/ (78 per cent of relevant documents identified in year 2 were found by search methods other than using Boolean and keyword searching). See generally, Jason R. Baron, ' The TREC Legal Track: Origins and Reflections on the First Year,' 8 Sedona Conf. J. 251 (2007).*

[30] *See 'An Open Letter To Law Firms and Companies In The Legal Tech Sector,' available at http://trec-legal.umiacs.umd.edu/, and referenced at n. 10 of Victor Stanley v. Creative Pipe.*

