



# DIGITAL DISCOVERY & E-EVIDENCE

**VOL. 10, NO. 16****REPORT****SEPTEMBER 16, 2010**

Reproduced with permission from Digital Discovery & e-Evidence, 10 DDEE 16, 09/16/2010. Copyright © 2010 by The Bureau of National Affairs, Inc. (800-372-1033) <http://www.bna.com>

**BNA INSIGHTS**

## Defending the Use of Analytical Software In Civil Discovery



BY CONOR R. CROWLEY, ESQ.

In complex litigation where the population of documents to be reviewed may consist of millions or even tens of millions of unique documents, subjecting each document to individual manual review is neither reasonable nor consistent with Fed. R. Civ. P. 1, which directs that the Rules be “construed and administered to secure the just, speedy, and inexpensive determination of every action and proceeding.” In such circumstances, first-pass review<sup>1</sup> is most effectively performed using analytical software tools that leverage lawyers’ knowledge of the issues in the case to make accurate,

testable relevancy determinations. The use of such tools is also consistent with the proportionality considerations set forth in the Federal Rules of Civil Procedure, allowing for a reduction in costs through the appropriate application of human versus software review in a given case.<sup>2</sup> Indeed, the use of analytical software tools to reduce the costs of discovery may, in complex litigation, be the only way to ensure that parties fulfill their production obligations in a manner that satisfies the

<sup>1</sup> “First pass review” is often the first step in the document review process where attorneys make initial determinations with respect to the relevance of documents to the issues in a particular matter.

<sup>2</sup> In deciding whether to limit discovery, courts consider whether “the burden or expense of the proposed discovery outweighs its likely benefit, considering the needs of the case, the amount in controversy, the parties’ resources, the importance of the issues at stake in the action, and the importance of the discovery in resolving the issues.” Fed. R. Civ. P. 26(b)(2)(C)(iii).

proportionality considerations set forth in the Federal Rules.

## I. KEYWORDS VERSUS ANALYTICAL SOFTWARE

The use of keywords to filter large volumes of documents into relevant and non-relevant sets is well established. This use of keywords has been widely endorsed by the judiciary as an efficient means of reducing the tremendous volumes of documents that must now be reviewed in complex litigation.<sup>3</sup> However, the problems with using keywords are well-recognized. A seminal 1985 study found that attorney review teams using keywords and a full-text search and retrieval system were relatively ineffective at finding responsive documents.<sup>4</sup> Although they believed that they had retrieved 75 per cent or more of the responsive documents, they had actually found only about 20 per cent of them.<sup>5</sup>

More recently, the National Institute of Standards and Technology has sponsored a Text Retrieval Conference, the Legal Track of which focuses on the evaluation of search technology for discovery of electronically stored information in litigation and regulatory settings.<sup>6</sup> The TREC 2008 results show that teams of reviewers using keywords and complex Boolean queries achieved an average recall rate of 22 per cent.<sup>7</sup> Additionally, as The Sedona Conference has noted,

... 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).<sup>8</sup>

<sup>3</sup> See, e.g., *In re Lorazepam & Clorazepate*, 300 F. Supp. 2d 43, 46 (D.D.C. 2004); *In re CV Therapeutics, Inc.*, 2006 WL 2458720 (N.D. Cal. Aug. 22, 2006) (court endorses employment of search terms as reasonable means of narrowing production); *J.C. Associates v. Fidelity & Guaranty Ins. Co.*, 2006 WL 1445173 (D.D.C. 2006) (requiring search of files using four specified keywords); *FTC v. Ameridebt, Inc.*, 2006 WL 6188563 (N.D. Cal. Mar. 13, 2006) (“e-mail could likely be screened efficiently through the use of electronic search terms that the parties agree upon.”); *Windy City Innovations, LLC v. America Online, Inc.*, 2006 WL 2224057 (N.D. Ill. July 31, 2006) (“[k]eyword searching permits a party to search a document for a specific word more efficiently.”); *Reino de Espana v. Am. Bureau of Shipping*, 2006 WL 3208579 (S.D.N.Y. Nov. 3, 2006) (court approves of keyword search for names and email addresses as a “targeted and focused discovery search”); *U.S. ex rel. Tyson v. Amerigroup Ill., Inc.*, 2005 WL 3111972 (N.D. Ill. Oct. 21, 2005) (referencing agreement by parties to search terms); *Medtronic Sofamor Danek, Inc., v. Michelson*, 229 F.R.D. 550 (W.D. Tenn. 2003) (court orders defendant to conduct searches using the keyword search terms provided by plaintiff).

<sup>4</sup> Blair, D.C., & Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full-text document retrieval system. *Communications of the ACM* 28, 289-29.

<sup>5</sup> Id.

<sup>6</sup> Overview of the TREC 2008 Legal Track available for download at <http://trec.nist.gov/pubs/trec17/papers/LEGAL.OVERVIEW08.pdf>.

<sup>7</sup> Id.

<sup>8</sup> The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery.

Courts have also recognized the potential for incompleteness or over-inclusiveness of keywords.<sup>9</sup> The problems with keyword searching arise not simply from a poor choice of keywords but also from the limitations inherent in keyword searching. Keyword searches retrieve documents that contain words matching those in the search string, and fail to retrieve documents containing antonyms, synonyms, different words with the same meaning (e.g., bid, offer, proposal), different forms of the same word and slang terms.

Additionally, although the use of proximity operators (to find instances where two or more words occur within a given number of words of each other in a document) can retrieve documents deemed relevant because of the occurrence of multiple relevant terms together, keyword searches cannot consider the context in which such terms arise.

Finally, keyword searches result in binary determinations of relevance, resulting in a set of documents that are neither weighed nor ranked based on their degree of relevance. Thus, keywords are a useful tool for filtering large document volumes for relevant documents but have known limitations.

Unlike keyword searches, which determine the relevance of particular documents based simply on the occurrence of specific words, analytical software makes relevance determinations based on the placement and frequency of terms within a document; the proximity of these terms; and implicit or explicit feedback on relevance.<sup>10</sup>

More advanced analytical software, such as dimension reduction systems, also consider the correlation of terms. For example, a document that mentions “plaintiff” is also likely to mention “defendant”, “court”, “complaint”, etc. Although these words are not synonyms, they do have similar or related meanings. Documents that mention any of these words would likely be about law. These systems learn the meanings of the words in documents from the documents themselves. This allows dimension reduction systems to find documents with similar meanings even where those documents contain different words.

## II. THE USE OF ANALYTICAL SOFTWARE IN CIVIL DISCOVERY

Although the use of analytical software tools in civil discovery is relatively recent, courts have already recognized the potential application of these tools for reducing the cost and increasing the accuracy of discovery. For example, in *Victor Stanley, Inc. v. Creative Pipe, Inc.*, 250 F.R.D. 251 (D. Md., 2008), the court acknowledged the limitations of keyword searching and described more advanced search and retrieval techniques that could be used for discovery purposes. Similarly, in *Disability Rights Council of Greater Washing-*

<sup>9</sup> See, e.g., *Alexander v. FBI*, 194 F.R.D. 316 (D.D.C. 2000) (court limits scope of plaintiffs’ proposed keywords to be used to search White House email); *Quinby v. WestLB, AG*, 2006 WL 2597900 (S.D.N.Y. Sept. 5, 2006) (court narrows party’s demand for 170 proposed search terms in part due to the inclusion of commonly used words).

<sup>10</sup> For a more detailed explanation of how the various types of analytical software tools operate, see The Sedona Conference Best Practices Commentary on the Use of Search and Information Retrieval Methods in E-Discovery.

*ton v. Washington Metropolitan Transit Authority*, 2007 WL 1585452 (D.D.C. June 1, 2007), the court suggested that the parties use concept searching instead of keyword searching.

The use of technological tools is also supported by, *inter alia*, The Sedona Conference. Principle 11 of The Sedona Principles provides that:

A responding party may satisfy its good faith obligation to preserve and produce relevant electronically stored information by using electronic tools and processes, such as data sampling, searching, or the use of selection criteria, to identify data reasonably likely to contain relevant information. *The Sedona Principles*, Second Edition (2007).

Additional support for use of analytical tools to select responsive documents from a larger set is found in the Explanatory Note on Evidence Rule 502 Prepared by the Judicial Conference Advisory Committee on Evidence Rules. The Judicial Conference Advisory Committee stated that one of the two major purposes of FRE 502 was to respond to the “widespread complaint that litigation costs necessary to protect against waiver of attorney-client privilege or work product have become prohibitive,” and noted that its concern is particularly acute in cases involving electronic discovery due to the volume of information involved.

As noted by the court in *Hopson v. City of Baltimore*, 232 F.R.D. 228, 244 (D. Md. 2005), “record-by-record pre-production privilege review, on pain of subject matter waiver, would impose upon parties costs of production that bear no proportionality to what is at stake in the litigation.”

The Notes to FRE 502(b) provide that:

A disclosure of a communication or information covered by the attorney-client privilege or work product protection does not operate as a waiver in a state or federal proceeding if the disclosure is inadvertent and is made in connection with federal litigation or federal administrative proceedings—and if the holder of the privilege or work product protection took *reasonable precautions* to prevent disclosure and took reasonably prompt measures, once the holder knew or should have known of the disclosure, to rectify the error, including (if applicable) following the procedures in Fed. R. Civ. P. 26(b)(5)(B). (Emphasis added.)

The Notes to Subdivision (b) explicitly provide that “a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken ‘reasonable steps’ to prevent inadvertent disclosure.” Accordingly, the Federal Rules of Evidence recognize the reasonableness of the use of analytical software applications to identify responsive but privileged documents in discovery.

Thus, both federal case law and the rules governing civil litigation currently recognize the appropriateness and the reasonableness of the use of analytical software tools to identify responsive documents for production, with relevant case law noting the importance of quality assurance testing to ensure completeness of production.

### III. ACHIEVING DEFENSIBILITY

To be defensible, any methodology employed to review documents for responsiveness to discovery re-

quests must be reasonable under the circumstances.<sup>11</sup> Drawing from the holdings in *Victor Stanley* and *Gross Construction*, a recent law review article sets forth a multi-factor analysis that litigants should undertake to determine the reasonableness of a selected search and review process to meet the reasonable inquiry standard of Rule 26(g):

1. Explain how what was done was sufficient;
2. Show that it was reasonable and why it was reasonable;
3. Explain the qualifications of the persons selected to design the search;
4. Craft appropriate keywords with input from the ESI’s custodians to ensure an understanding of the terminology and abbreviations used;
5. Perform quality assurance testing on the methodology to assure accuracy in retrieval.<sup>12</sup>

**Sufficiency.** In order to demonstrate that what was done was sufficient, a statistically significant sample of the document population should be reviewed by attorneys. Relevancy determinations made by analytical software are based on the relevancy designations made by reviewing attorneys; accordingly the sample of documents reviewed by attorneys should be large enough to allow the software to make accurate relevancy determinations.

**Reasonableness.** The Federal Rules of Civil Procedure require that, in response to discovery requests, a party produce non-privileged documents that are relevant to a party’s claims or defenses.<sup>13</sup> This requirement is subject to a reasonableness standard, as evidenced by the requirements of Fed. R. Civ. P. 26(g)(1) which provides that the production of such ESI must be accompanied by a signed certification from the producing attorney or party stating that “to the best of the person’s knowledge, information, and belief formed after a reasonable inquiry . . . the disclosure is complete and correct as of the time it is made.”<sup>14</sup>

As explained above, the reasonableness of the use of analytical software tools to retrieve relevant documents has found acceptance in both case law and the federal rules.<sup>15</sup> To demonstrate reasonableness the levels of recall and precision achieved by the analytical software should be at least as good as the levels of recall and precision that would have been achieved by keywords. Additionally, the reasonableness of the use of analytical software that is based on standard, thoroughly researched technologies will be easier to demonstrate than the use of analytical software that is based on novel, untested, or unrevealed technology.

<sup>11</sup> Fed.R.Civ.P. 26(g)(1), Advisory Committee’s Notes to 1937 adoption. See also, *Mancia v. Mayflower Textile Servs. Co.*, 253 F.R.D. 354, 357 (D. Md. 2008) (“The duty to make a ‘reasonable inquiry’ is satisfied if the investigation undertaken by the attorney and the conclusions drawn therefrom are reasonable under the circumstances.”).

<sup>12</sup> See Oot, P., Kershaw, A., Roitblat, H., Mandating Reasonableness in a Reasonableness Inquiry, 87 DEN. U. L. REV. 533 (2010).

<sup>13</sup> Fed.R.Civ.P. 26(b)(1), 34(a).

<sup>14</sup> Fed.R.Civ.P. 26(g).

<sup>15</sup> See, e.g., the Notes to Fed. R. Evid. 502(b) which provide that “a party that uses advanced analytical software applications and linguistic tools in screening for privilege and work product may be found to have taken ‘reasonable steps’ to prevent inadvertent disclosure.”



**Qualifications.** The review attorneys who make the relevancy designations that are used by the analytical software to determine relevancy should be well-versed in the issues in the case at hand. This ensures that the attorney designations on which the system bases its relevancy determinations are based on the best available information.<sup>16</sup>

**Terminology.** With keywords, recall and precision levels are dependent on the reviewing attorney's investigation of the jargon, lingo and abbreviations used by the parties in a particular case, a linguistic task which, courts have noted, may be beyond the scope of an attorney's skills and training.<sup>17</sup> With analytical software, an attorney makes relevancy determinations that are not dependent on finding exact word matches, but rely instead on the concepts found in each document, resulting in higher levels of recall.

**Quality Assurance Testing.** Quality assurance testing is essential to demonstrating the accuracy and reliability, and thus the reasonableness, of any review methodology.<sup>18</sup> The levels of recall and precision achieved by the analytical software should be readily apparent to the reviewing attorney. Additionally, quality assurance testing whereby the relevancy determinations made by both the analytical software and the reviewing attorneys are reviewed for accuracy should be conducted by an independent reviewer who is at least as familiar with the issues as the reviewing attorneys. This will help to confirm that the relevancy determinations made by the system are at least as accurate as the relevancy determinations made by human review, or any other culling or review method used. Additionally, those documents deemed not relevant by the analytical software should be sampled to ensure that the system has not systematically overlooked relevant documents.

<sup>16</sup> Conversely, leveraging the knowledge of less experienced counsel would likely result in less accurate relevancy determinations, as is often the case in large scale human document review.

<sup>17</sup> See, e.g., *U.S. v. O'Keefe*, 537 F. Supp. 2d at 24 (D.D.C. 2008).

<sup>18</sup> See, *In re: Seroquel*, 244 F.R.D. 650 (M.D. Fla. Aug. 21, 2007) ("sampling and other quality assurance techniques must be employed to meet requirements of completeness."); *Victor Stanley*, 250 F.R.D. 251, 256 (D. Md., 2008) (court noted the importance of showing what quality controls were employed to assess the reliability and accuracy of the search methodology).

## IV. CONCLUSION

With the tremendous volumes of information that must be reviewed in complex litigation, it is no longer practical or cost-effective to conduct page-by-page relevancy review through the exclusive use of human reviewers. Given the limitations inherent in keyword searching, the use of appropriate analytical software tools provides a more defensible method for effectively filtering large-scale document productions. However, it is important that the selected tool employ recognized, accepted technologies, and essential that the human portion of the review is conducted by knowledgeable attorneys. The tool should be capable of quantifying and verifying the levels of precision and recall achieved. This will allow for quality assurance testing to ensure that the results of the review process are demonstrably accurate and reliable, and thus reasonable. If these criteria are satisfied, the use of analytical software for culling and first-pass review should provide accurate, cost-effective relevancy determinations and should be considered no less defensible than the exclusive use of human reviewers.

*Conor R. Crowley is an internationally recognized expert in e-discovery and e-compliance. Mr. Crowley is a member of the Steering Committee for The Sedona Conference Working Group on Best Practices for Electronic Document Retention and Production (WG1), and a member of The Sedona Conference Working Group on International Electronic Information Management, Discovery and Disclosure (WG6), the International Association of Privacy Professionals, the Advisory Board for Georgetown University Law Center's Advanced E-Discovery Institute and the Board of Advisors for BNA's Digital Discovery & e-Evidence. He is a Senior Editor of a number of The Sedona Conference's publications including The Sedona Conference Commentary on Legal Holds and The Sedona Principles (Second Edition): Best Practices Recommendations & Principles for Addressing Electronic Document Production. He can be reached at [ccrowley@crowleylawoffice.com](mailto:ccrowley@crowleylawoffice.com).*