

## When E-Discovery Is Put to the Test

Will federal rules on expert testimony govern admission of search engine results?

By Leonard Deutchman  
[Pennsylvania Law Weekly](#)  
May 14, 2008

An influential federal district judge whose opinions on e-discovery are well respected may have set e-discovery on a path toward its most searching scrutiny yet.

In *Disability Rights Council v. Washington Metropolitan Transit Authority*, 242 F.R.D. 139 (D.D.C. 2007), Judge John M. Facciola recommended "concept searching," -- the use of complex search engines that make use of linguistic or statistical patterning to locate responsive e-mails and electronic -documents, in order for a tardy producer of discovery to wade through voluminous electronically stored information quickly. Interestingly, Facciola made no mention of whether the use of concept searching tools should be subject to Federal Rule of Evidence 702, which governs the admission of scientific or expert testimony.

Recently, however, in *United States v. O'Keefe*, 537 F. Supp. 2d 14 (D.D.C. 2008) and *Equity Analytics v. Lundin*, 2008 U.S. Dist. LEXIS 17407 (D.D.C. Mar. 7, 2008), Facciola held that any challenges to or defenses of search methodology in producing e-discovery must be scrutinized under Rule 702, and so ordered hearings under *Daubert v. Merrill Dow Pharmaceuticals*, 509 U.S. 579 (1993).

These rulings give rise to the question of what a *Daubert* hearing for an e-discovery search engine would look like.

### HOW AND WHERE TO SEARCH

The first issue the court would address is how search engines search. The most direct approach is keyword searching, which take three basic forms:

- Direct searching for keywords, e.g., "Locate all files with 'Jones.'"
- Boolean searching, e.g., "Locate 'Jones' or 'Smith'," "Locate 'Jones' but not 'Smith,'" and other combinations.
- Proximity searching, e.g., "Locate 'Jones' within 25 words of 'Smith.'" Often such searching is restricted by date range, e.g., "Locate all e-mails with 'Jones' created after January 1 but before July 1, 2007 only."

Concept searching, as has already been briefly discussed, takes a different approach. It targets information relating to a concept even if specific keywords are not present (e.g., a series of e-mails mentioning the words "Clinton," "McCain" and "Obama" would likely concern the 2008 U.S. presidential election, even if the phrase "presidential election" does not appear).

Some concept searching tools use "taxonomies" or "ontologies," that is, compilations of both commercially available data and data supplied by the client pertinent to the case collected from the lawyers and key players. Some concept searching uses linguistic analysis examining how the communicants discuss matters, while other approaches, such as "clustering" and "latent semantic indexing," use mathematical probabilities to determine whether a given file is related to a given concept. For an excellent discussion of concept searching, see "[The Sedona Conference Best Practices Commentary on the Use of Search and Informational Retrieval Methods in E-Discovery.](#)"

## **A DIFFICULT HEARING**

Regardless of which approach the search engine takes, the actual *Daubert* hearing will prove difficult for two practical reasons, both stemming from the fact that search engine applications are proprietary. First, it simply will be hard to get the designer to appear at the hearing to testify as to how the engine works. Second, the designer will fight giving the best evidence of the efficacy of the engine, that is, the engine's source code, because that code is proprietary. Should the code be revealed, the design would lose its value, as anyone could use that code without having to obtain a license (i.e., a copy of the application) from the designer.

Proprietary applications can be validated without their source code revealed, but only under certain circumstances. The easiest is where a specific positive finding needs to be corroborated. For example, if a proprietary forensic search tool such as Guidance Software's [EnCase](#) reports that a file is found at a particular location on a hard drive, an examiner can use an "open source" tool, i.e., a tool whose source code is known and which has been validated, to confirm the finding. Such corroboration, however, does not validate the search tool, only the result of the use of that tool at a particular time. To validate a proprietary tool generally using open source tools requires months of work, thousands of hours by highly experienced analysts, such as the FBI put in when validating EnCase. Of course, each time a new version of a tool comes out, more hours of validation are needed. Thus, while this means may be reliable, it is hardly practical.

A second means would be to use another proprietary tool -- say, Access Data's [FTK](#) -- to run the same search as EnCase performed and compare results. This method, however, is not truly scientific, since identical search results are just as likely to confirm that the two engines are identically flawed as they are reliable.

A third means to validate a proprietary search engine without revealing its code would be to search test data sets with known test results and which contain the types of data that the engine would search when regularly deployed. Comparing the results of the searches by the proprietary search engine to the known results should validate or invalidate the search engine. Again, however, such testing is extremely time-consuming and expensive. The designer would have to engage in such testing and publish its results; one could hardly expect the typical user of the search engine to engage in such studies.

As previously stated, the second *Daubert* hearing issue is where the searching was done. Specifically, the issue would be whether only the files actively stored on a hard drive, for example, were searched, or whether deleted files, temporary files or file fragments in the

"unallocated space" of a hard drive were also searched. When ESI is gathered, unless bit stream, forensic images (i.e., exact copies of every 1 and 0 on a piece of media) are made, the deleted files, etc., will not even be present to search. To search for such ESI, forensic tools must be used. Thus, in *United States v. O'Keefe*, for example, the defendant challenged the government's search results for potentially exculpatory evidence in its possession by arguing that by not looking "everywhere" on the drive for deleted files or file fragments, the government had not fully discharged its duty to search everywhere.

The problem with searching "everywhere," however, is not so much a Rule 702 problem as a practical one: forensic searches of every possible file fragment take impossibly long, and if many hard drives and servers are involved, the impossible becomes unthinkable. *O'Keefe*, however, raises another issue, one far more interesting and conceptually difficult: for search engines, passing the *Daubert* test may depend upon whether one is trying to prove that something is there or that something is not there.

## **EVIDENCE AND ITS ABSENCE**

Anyone who remembers examining scientific method when taking high school or college science classes will recall the question whether the absence of evidence that "x" is present means that "x" truly is not present or whether the test for finding "x" was simply insufficient. For example, while a PET Scan's positive finding for cancer is conclusive, a failure to detect cancer may mean the absence of cancer or that the PET Scan failed to detect cancer that was present.

Thus, the acceptance of a test as scientific proof under *Daubert* and Rule 702 is more likely when the test is to prove that something is present than absent. In *Sanders v. Texas*, 191 S.W.3d 272 (Ct. App. 2006), for example, the Texas Court of Appeals had no trouble affirming the trial court's findings that the expert's use of EnCase to create a bit stream, forensic image of the defendant's hard drive and his search of the drive to uncover child pornography -- both positive findings -- was scientifically valid. Since Encase's findings can be corroborated by a tool other than the proprietary one used, the validity of the imaging and search is much easier to establish.

In *O'Keefe, Equity Analytics v. Lundin* and the prototypical e-discovery matter, the typical challenge is the opposite of the typical challenge in a criminal matter: the requesting party's typical challenge to e-discovery production is not that it is inauthentic but that it is incomplete. The *Daubert* challenge in e-discovery cases is to prove that the search results yielded "everything."

If the search engine in question were an open-source tool, the challenge could be more easily met: the search engine's methodology would be open for all to test, and it would either work when searching test sets with known results or produce anomalies or mistakes. However, if the search engine is proprietary, proving the negative (it did not miss anything) by proving the positive (this is how it searches) is not available to the tool's proponent. The "third means" discussed above -- subjecting the search tool to known test data to see whether it missed any "hits" -- could work, but that means is extremely time-consuming, expensive and beyond the capability of the typical user.

## THE MYTH OF PERFECTION

The Sedona Conference commentary provides an interesting method of "corroboration." It cites a study in which review attorneys, doing a "manual" review of discovery, were asked how much responsive data they were able to find. The attorneys guessed 75 percent, but a detailed analysis revealed that they had found only 20 percent. Using that study to illustrate what the Sedona Conference's commentary refers to as the "myth of perfection," i.e. that review attorneys slogging through e-documents and e-mails will catch responsive ESI that concept search engines will miss, the commentary makes the scientifically questionable but legally valid point that the validity of concept search tools must be determined by measuring concept search results against the actual results of review attorneys, not against results of a "perfect" search. If concept searching improves upon review practice as it now stands, it is a valid litigation tool.

In making its point, the Sedona Conference commentary returns to a touchstone of discovery practice: that when producing discovery, a "perfect review ... of information is not possible ... . The governing legal principles and best practices do not require perfection in making disclosures or in responding to discovery requests."

The *Daubert* challenge raised by Facciola, then, may be met not by judging the scientific validity of a search engine in an absolute way, but by judging how valid it is to suit the purposes of e-discovery production, an undertaking which involves many factors, such as the costs in time, money and energy to the producing party and their marginal benefit to the requesting party and the litigation, that have no bearing on the scientific validity of the search engine. In other words, the ultimate acceptance of an e-discovery search tool may be informed by its relative perfection but will ultimately depend, like so many other things in the law, upon the totality of circumstances.

Reprinted with permission from the 05/14/2008 edition of Pennsylvania Law Weekly © 2008 ALM Properties, Inc. All right reserved. Further duplication without permission is prohibited.