Defending the Accuracy of Phonetic Audio Search in Civil Discovery

David Fishel, Esq.
Senior Director/ Technology Counsel
Nexidia Inc.

INTRODUCTION

The explosion of digital information – including audio recordings such as voicemail, call center conversations, and web conferences – shows no sign of abating; the data sets are only getting bigger. In this era of electronic discovery, the amount of data has grown so large that unassisted human review is impossible. Electronic tools and new methodologies are required to find, retrieve, review, and track potential evidence.

Discovery of audio recordings requires new technologies to meet discovery requirements, just like other forms of "electronically stored information" (ESI). However, there are no standards by which litigants or courts measure how well any search and retrieval methods – old or new – actually perform. Users of new search and retrieval technologies are concerned that they will be called upon to demonstrate that any technology for discovery is "defensible."[1]

Courts Have Not Assessed the Accuracy of Discovery Search Technologies

There is no agreed-upon standard for measuring different information retrieval methodologies (let alone individual software products) with regard to how they actually perform in civil discovery. There are also no federal benchmarking guidelines for evaluating information retrieval products and methods used in discovery. It is therefore unsurprising that there are no reported cases addressing search methods for audio evidence, and only a very small number that mention even as commonplace a practice as keyword searching of text-based ESI. [2]

Case law applying technology to discovery lags far behind the technology itself. Even common text keyword search technology is rarely mentioned by a court, and no reported cases assess a particular technology or product – whether text search, de-duplication, or concept search – for accuracy.

The ongoing explosion of ESI guarantees that automated discovery tools will be both more necessary and more widely used in the future.[3] The volume of ESI is so much larger than conventional paper discovery that it is impossible to litigate effectively without using search

---

[1] "Defensibility" is a term that means different things to different people, and in different contexts. The issues surrounding "defensibility" for a search technology are different from the issues associated with, for example, a forensic data recovery tool used to extract data from a computer drive. The data recovery tool directly affects the preservation or admissibility of evidence, while the search tool may affect the completeness of a discovery response. Properly used, search tools do not have the ability to alter evidence and do not affect the chain of custody.

[2] Jason R. Baron, *Toward a Federal Benchmarking Standard for Evaluating Information Retrieval Products Used in E-Discovery*, 6 SEDONA CONF. J. 237, 239 (2005).

[3] *See, e.g.*, Principle Number 11, Sedona Conference, *The Sedona Principles for Electronic Document Production*, July 2005 ("a responding party may satisfy its good faith obligation to … produce potentially relevant [ESI] by using electronic tools and processes, such as … searching …")

technology of some kind.[4] Despite the growth of ESI and the constant development of new search and retrieval technologies, as of early 2007 there are no reported cases on Westlaw addressing the use of "concept search" or any other alternative search methods beyond keyword searching in connection with an adjudicated issue in civil discovery.[5] However, it appears that litigants and law firms are actively using such systems in discovery – as evidenced by the burgeoning litigation support vendors creating and touting new technologies.

The growth of ESI encompasses more than email and other forms of text. Although there are no reliable estimates of the total amount of audio recordings stored in corporate systems, the volume is certainly enormous. It is now commonplace to find corporate voicemail systems with 10,000 mailboxes, corporate customer call centers that receive 20-30 million calls per year, and financial service trading desks where 50 phone lines are recorded around the clock every day of the year. Not all audio collections are so large, but audio review quickly becomes costly and difficult with even a few hundred hours of recordings. The ability to search sound recordings is rapidly becoming necessary for conducting effective discovery.

Despite the present lack of standards for evaluating search and retrieval tools "a new jurisprudence on the quality of search methods employed, governing how one must conduct more advanced searches, is likely to evolve."[6] Standards will become even more important than they are today as litigants facing ever-growing mountains of ESI look to innovative technologies as the only way to meet their discovery burdens, and disputes arise over the effectiveness of competing approaches.

Search and retrieval technologies are crucial for any kind of large-scale discovery because they are the only hope for being both thorough and cost effective. Broadly stated, modern discovery operates on the principles that parties should litigate in full possession of the relevant information, and that the justice system is best served when parties economize the time and expense of trying lawsuits. Parties are expected to conduct "reasonable" searches of their discovery materials, not "perfect" ones. Thus parties must find and produce "all" relevant materials, while also balancing issues of completeness and efficiency.

In cases that pre-dated ESI, courts did not typically look into the discovery review process. Instead, courts relied on the representations of counsel that they had made a thorough search and produced all responsive materials. Disputes tended to focus on whether all responsive documents were identified, rather than the specific criteria for selecting them. Now, even though attorneys must rely on search tools, the assumption that counsel has accurately assessed all potential evidence is still prevalent. Courts sometimes examine the discovery process more deeply if

---

[4] *See* MANUAL FOR COMPLEX LITIGATION (4TH) § 11.446 (2004). It is no longer unusual for large civil litigation or government investigations to involve a terabyte (i.e., 1,000 gigabytes, 1 million megabytes, or about 500 billion typewritten pages) of ESI.

[5] George L. Paul and Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, [29] (2007).

[6] *Id*. at 30. *See also*, the forthcoming Sedona Conference paper on search and retrieval, due to be published in the 2007 Sedona Conference Journal.

opponents allege willful or bad faith failures to find and produce discoverable information, but they don't typically investigate search and review methodologies.[7]

Courts understand that it is not reasonable to do a review of e-discovery materials without employing search technology. Although courts expect data reduction techniques such as de-duplication and keyword searching to be used when appropriate, they do not yet closely examine the effectiveness of such tools. Instead, courts rely on counsel to satisfy their professional obligation to provide complete, accurate discovery. It appears that counsel, in turn, relies on technical experts. In most cases, it is not clear what the experts rely on to assert that search technologies and strategies are thorough and accurate, as they are not called upon to present evidence of their effectiveness.[8]

When search technologies have been considered by courts, it has generally been in the context of a party that failed to do *any* automated search. Courts have instructed parties to conduct "keyword" text searches as a means of winnowing large collections and identifying potentially responsive documents. [9] A few courts have directed the parties to confer and agree upon sets of keywords, search terms, or other protocols, but they do not opine on the technologies used in those efforts.[10]

Lawyers and judges seem to take for granted the accuracy and cost effectiveness of most current search products. It may surprise many litigants and courts that the "assumption on the part of lawyers that any form of present-day search methodology will fully find "all" or "nearly all" available documents in a large, heterogeneous collection of data is wrong in the extreme."[11] The courts have not yet addressed (or even acknowledged) the fact that widely accepted search and retrieval technologies and methodologies are far from perfect. Studies have shown that human searchers are not very effective using familiar search tools that use either Boolean logic or more advanced search functionality. Optical Character Recognition (OCR), widely used for many years, has a well-documented error rate that significantly affects search results. Even spelling errors and abbreviations affect accuracy of search results, although the impact of those factors is not known.

---

[7] For example, the court in *Zubulake v. UBS Warburg, LLC*, 229 F.R.D. 422 (S.D.N.Y. 2004) ("Zubulake V"), enumerated several things that the defendant could have done to meet its discovery obligations, but did not.

[8] Again, this discussion is directed at search technologies. Forensic data recovery experts are often required to testify as to their tools, methods, and effectiveness.

[9] At present, no matter how advanced a search technology may be, litigants rely on human judgment to determine whether information is relevant or privileged. Human review costs are almost always the largest cost in a discovery project. Thus, anything that can reduce the amount of material to be reviewed or expedite the review process can greatly reduce the downstream costs.

[10] *See Balboa Threadworks, Inc. v. Stucky*, 2006 WL 763668 (D. Kan. Mar. 24, 2006)(instructing the parties to design a search protocol and to "lean heavily on their respective computer experts" in doing so). *See also Rowe Entertainment v. William Morris Agency*, 205 F.R.D. 421, 432-22 (S.D.N.Y.2002); *Antioch Co. v. Scrapbook Borders, Inc.*, 210 F.R.D. 645, 653-54 (D.Minn.2002); *Johnson v. Kraft Foods N. Am., Inc.*, 2006 WL 3302684 (D. Kan. Nov. 14, 2006); *Simon Property Group L.P. v. MySimon, Inc.*, 194 F.R.D. 639, 641-44 (S.D.Ind.2000).

[11] George L. Paul and Jason R. Baron, *Information Inflation: Can the Legal System Adapt?*, 13 RICH. J.L. & TECH. 10, [40](2007).

In light of the fact that technologies are competing against each other, and against a utopian (and unsupported) "gold standard" of human reviewer competence, there will certainly be more empirical studies conducted. As the process of identifying discovery materials is increasingly turned over to computer hardware and software, litigants should expect to see deeper examination of discovery search and retrieval, especially if judicial benchmarking guidelines are established. In the meantime, how can litigants who use advanced technologies ensure that they are fully meeting their discovery obligations?

The Roles of the Parties in Selecting Discovery Methods

Parties are held to a standard of reasonable professional behavior in conducting discovery. The "reasonableness" standard does not provide much guidance for litigants evaluating technology, and there is no method of responding to e-discovery that is presumptively reasonable.

Choices about discovery technology are left to the parties' discretion. Responding parties have the primary responsibility to select methods of discovery that produce satisfactory results. The Sedona Conference's Electronic Discovery Principle Number 6 asserts that "Responding parties are best situated to evaluate the procedures, methodologies and technologies appropriate for preserving and producing their own electronic data and documents."[12] The corollary to that principle is that the responding party bears the consequences of its policies and actions. If the responding party's methodologies are inept, ineffective, or result in the destruction of discoverable evidence, it may face sanctions. However, there are currently no reported cases where a responding party acting in good faith was sanctioned for using a search technology or method that produced less than perfect results.[13] The primary criteria for determining whether a technology is "reasonable" seem to be 1) the responding party's judgment that it can meet its professional obligations using the technology, and 2) whether it can explain its actions to other parties and the court should the need arise.

As with the courts, if parties and technology vendors are doing empirical accuracy tests, few are publishing the results. Performance evaluation protocols and verifiable results of performance testing are not yet a regular part of the discovery process. At present, as long as the responding party acts in good faith, a requesting party faces an uphill battle if it challenges the responding party's decision to use a "reasonable" technology selection.[14]

---

[12] Sedona Conference, *The Sedona Principles for Electronic Document Production,* July 2005.

[13] One reason for this may be that search technologies, unlike forensic recovery technologies, aren't likely to lead to spoliation of evidence, although they may still prejudice an opponent. If a search strategy proves to be ineffective, the evidence is still available for review. Technologies blunders that lead to loss of evidence are more obvious than a poor result with search and retrieval technology.

[14] Sedona Conference, *The Sedona Principles for Electronic Document Production,* July 2005. Principle Number 7 asserts that the requesting party has the burden of showing that the responding parties steps to produce relevant data were inadequate. See the references in that document with regard to case law under FRCP 37.

First, there is little or no case law to support such a challenge. Second, there are no judicial benchmarking guidelines or standards to invoke should the requester assert that the search technology or method was flawed. Third, courts do not typically impose a requesting party's selected method on the responding party, especially if the responding party makes a colorable showing that its methods are reasonable.[15]

Nevertheless, responding parties will not be well served by failing to cooperate on search and retrieval issues. A bilateral process does not just aid the requesting party. The longer the producing party waits to discuss options, the less credibility it will have in its future dealings with the other side and with the court. In addition, a joint plan puts fair pressure on the requestor to be more focused and particular from the outset.[16] If a requesting party is passive or fails to negotiate reasonably, a responding party may be justified in making unilateral discovery decisions, but such cases should become increasingly rare as the collaborative process becomes the norm.[17]

Requesting parties should fully participate in discovery technology issues. The new paradigm requires more from requesting parties than simply propounding requests for production, or passively awaiting mandatory disclosures. If requesting parties do not want to passively accept whatever the responding party provides, they need to take an active role in the process. By fully participating, getting educated about the data, systems, and technology options, and being reasonable in assessing cost, time, and accuracy concerns, requesters can get the discovery to which they are entitled.

Rule 26 provides a framework for collaboration on technology decision-making by requiring the parties to meet and confer on e-discovery issues very early in the case. Prudent parties will use these conferences to hammer out agreements on technologies, search terms, sampling methods and other issues. Requesting parties that do not avail themselves of the opportunity to participate in the selection of technology, methodology, search terms, etc. are missing an opportunity and are in a poor position to complain about the results. Requesting parties should make every effort to understand all they can about the opponents systems, or they will not be able to fully participate in the process. While the requesting party may not expect to substitute its preferred discovery

---

[15] If the responding party is extremely inept or recalcitrant, there is some precedent for the court to turn the search tasks (and the costs) over to the requestor, along with the obligation to provide search results to the responding party. *See Tulip Computers Int'l B.V. v. Dell Computer Corp.*, 2002 WL 818061, 52 Fed.R.Serv.3d 1420 (D.Del. 2002) (Ordering defendant to provide email of selected executives in electronic form to plaintiff's computer expert for search using the parties' agreed-upon search terms. Plaintiff provided defendant a list of the emails retrieved, and defendant produced the email subject to its own review for privilege and confidentiality designations.)

[16] Robert D. Brownstone, *Collaborative Navigation of the Stormy e-Discovery Seas*, 10 RICH. J.L. & TECH. 53 (2004) at [47]. This article also contains an enlightening section entitled "Case Studies in Lack of Cooperation (*McPeek* and *Tulip*)."

[17] *Treppel v. Biovail Corp.*, 233 F.R.D. 363, 368-69 (S.D.N.Y. 2006) (Stating that plaintiff's refusal to specify search terms was a "missed opportunity," but that defendant nevertheless had the burden of responding, the court mandated the use of certain search terms as an "interim step" in discovery.)

technology, it can still make sure that the responding party's chosen approach is as thorough and accurate as possible.

It seems inevitable that soon, a requesting party will file a motion to compel challenging the responding party's choice of search software or methodology. Consider a scenario where the requesting party is dissatisfied with the results of the responding party's search and retrieval efforts. If the Rule 26 conference process does not produce satisfactory changes to disputed techniques, the requesting party may be able to prevail in a motion to compel by showing that the responding party's actions were not reasonable – e.g., it used a demonstrably flawed or inferior search technology, or set search parameters that unreasonably eliminated potentially relevant discovery materials. Given the balancing of cost, time and accuracy factors, the moving party should be prepared to propose a better approach.

If a responding party is required to defend the accuracy or effectiveness of its search technology, the party must show that it acted reasonably and that its responses were as full and complete as possible. If the responding party lacks test data to support its claims of completeness, documentation for tests, or it cannot explain to the court the workings of the technology, it will be hard-pressed to show that its actions were reasonable. Litigants will find that it is much better to have conducted benchmark tests and not need them, than to need them and not have conducted them.

Evaluating Discovery Search Technologies

In this new discovery environment, benchmark tests will become more common as technologies compete to show that they are the fastest, most accurate, and most cost effective, while at the same time litigants and courts require empirical proof of search and retrieval effectiveness. Since there is no agreed-upon standard for measuring different information retrieval methodologies (let alone particular software products) in the context of how they actually perform in discovery, parties will have to develop their own. The first test results may be surprising -- parties that conduct empirical benchmarking efforts of new technologies will almost certainly find that they have been overly optimistic in their assumptions about the effectiveness of their previous strategies, and that search and retrieval methods used in the past may have been far less accurate than they believed.[18]

A benchmark for any large discovery project should incorporate three variables: cost, time to completion, and accuracy. A selected approach should be defensible on all three factors, although courts provide little guidance on balancing the factors. Accuracy is probably the most critical component in a benchmarking exercise, and the one least often measured by litigants.

[18] The few published studies of legal review methods show precision and recall both under 50%. *See* Paul and Baron, *supra* n. 9, for a discussion of accuracy that includes a description of the 1985 Blair & Maron study in which retrieval effectiveness was measured for 40,000 documents captured in a large corporate litigation. That study showed that attorneys and paralegals who were convinced that they were retrieving over seventy-five percent of the desired documents were actually retrieving only twenty percent.

Accuracy is more precisely defined in search and retrieval theory as "recall" and "precision." "Recall" is a measure of completeness, that is, how well a search has done in retrieving all of the items containing the search terms. Recall is critical in litigation because parties are obligated to produce all responsive information. In addition, a more complete result set enables one to do a more thorough review of evidence and provides a better assessment of the case. "Precision" is a measure of efficiency, that is, how well a search has done in retrieving true hits as a percentage of the total number of search hits retrieved. A low precision rate means that a large number of irrelevant items have to be reviewed. This affects cost and timeliness.

Precision and recall have an inverse relationship; that is, the higher the recall, the lower the precision. Thus, a system may get a very high percentage of hits (say 95%) but at the cost of also retrieving a very large number of irrelevant items mixed in with the relevant ones. Generally, as the recall (total correct hits) rises, the precision (percentage of true hits vs. false alarms) drops. This relationship is plotted on a curve. The inverse relationship between recall and precision is true of virtually every search technology, whether for text or audio recordings. It is up to the parties and courts to decide on acceptable tradeoffs between recall and precision, accuracy and cost. They can best make those decisions based on testing.

Establishing the Accuracy of Nexidia Phonetic Search

Nexidia Phonetic Search is a technology that hunts for sounds – in particular "phonemes," the fundamental sounds that make up any spoken language – rather than searching for words. Spoken words on recordings are simply combinations of phonemes. Phonetic search emphasizes how things sound, and does not require the computer to infer what the sounds mean. This approach completely redefines and simplifies audio search. While the English language consists of hundreds of thousands of words, the English spoken by North Americans is made up of only about 40 phonemes, even accounting for regional accents and dialects of native speakers. Searching for strings of this limited set of phonemes, rather than having a computer try to match a spoken sound to a written word in a lexicon, greatly simplifies and speeds up the search process and, among other things, also means that Nexidia has a completely open vocabulary, and is not limited to a base lexicon or dictionary.

Phonetic search results are returned as a list of putative hit locations, in descending likelihood order. As a user progresses further down the list, they will find more true hits on their search terms, and will also encounter an increasing percentage of false positives.

This performance characteristic is best shown in a curve common in detection theory: the Receiver Operating Characteristic Curve (ROC) (*See Figure 1*). To generate this curve, one needs experimental results from the search engine (the ordered list of putative hits) and the ideal results for the test set (acquired by manual review and documentation of the test data). The test process that generates the data to plot a curve that shows how recall and precision are related.

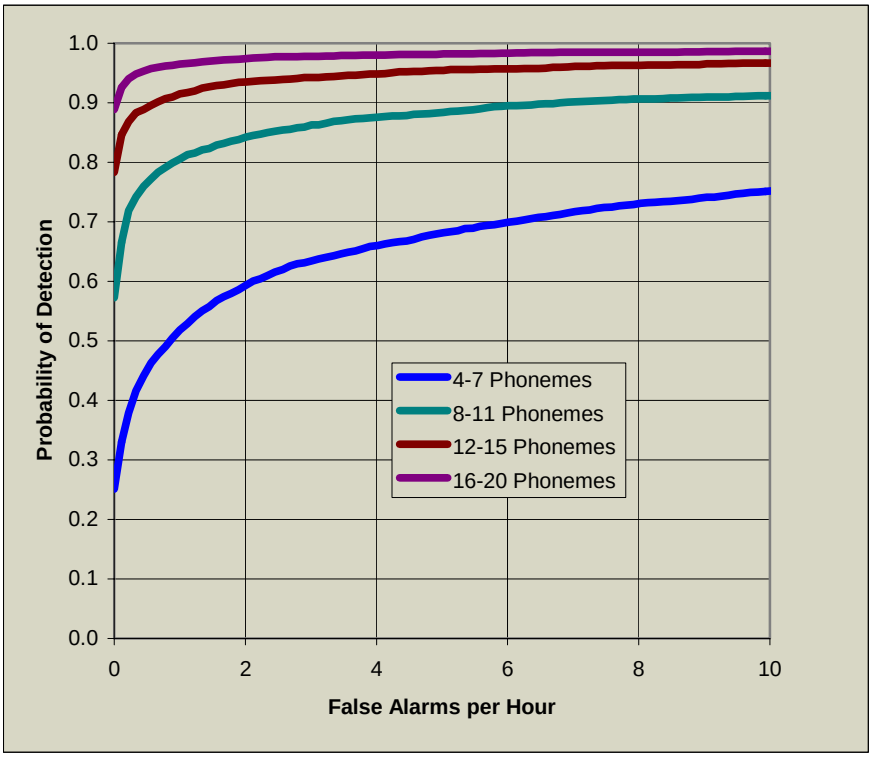
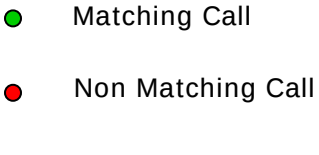
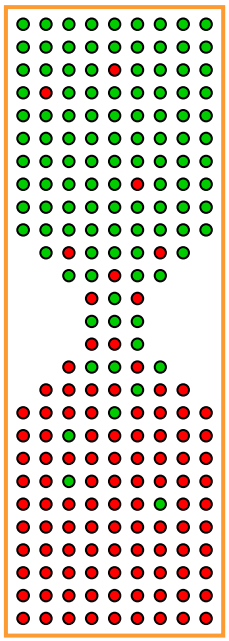


Figure 1

Since Nexidia phonetic search is probabilistic ("how much does this sound like the search term?") it *will* return results that are not exact matches, thus improving recall (total hits). This provides a margin of error for noise in the recordings, unusual pronunciations, and other variations of speech. Of course, it is possible to limit searches very tightly so they don't turn up anything but very highly ranked exact matches, but for legal applications, most users prefer to err on the side of over-inclusiveness – that is to emphasize recall over precision. Based on empirical data, parties are able to negotiate search terms, and to set recall and precision levels based on the needs of the parties and taking cost and time into consideration.

This visualization of the distribution of hits illustrates that near the top of the returned hits are the greatest number of true hits in the collection, with a few "false positives" scattered in. Below a certain level, very few true hits exist, but there are some. With testing, you can establish the threshold below which true hits become so scarce that they most likely are not worth the cost and effort of finding. Using test data, parties are able to intelligently negotiate with regard to search strategies, keywords, and a "reasonableness" standard for search and review thresholds.





Nexidia Methodology for Testing Recall and Precision

Nexidia does extensive testing to assure that the phonetic search has high levels of both precision and recall under real world conditions. [19] The test results, which are derived in a manner similar to the case-specific methodology described below, are available to Nexidia customers. This documentation provides a valid, documented benchmark against a known data set, and shows that Nexidia phonetic search has levels of recall and precision consistent with text search tools that are routinely used in litigation today. It also establishes that Nexidia's phonetic search is significantly more accurate than other automated audio search technologies.

The following is a simplified description of the method used to test the accuracy rate of the Nexidia phonetic search engine against a case-specific data set. First, a statistically significant set of audio is extracted from the full set. Care is taken to ensure that the recordings are representative of both the content and audio quality of the overall collection. Next, a verbatim transcript is created of the selected recordings. The transcript is of the highest quality, professionally transcribed and truthed, with a very high degree of quality assurance.

[19] Separate testing is done for each Nexidia "language pack." A language pack includes the phonetic content of the target language, and also takes into account regional accents, dialects, gender, and other differences in speech. In addition, each language includes different acoustic models to account for the differences in sound quality between, for example, broadcast studio versus landline telephony versus wireless (cellular) telephony. Nexidia currently has phonetic search "language packs" for over 30 languages. The primary language pack used for legal applications is North American English.

This step is the most time-consuming and expensive part of the process. Using automated text search tools, the transcript is time-coded, and a large body of search terms (over a thousand) is randomly selected from the transcript based on simple criteria.20 The time-coded location in the transcript of each search term is noted.

The audio recordings are then pre-processed and indexed into the Nexidia system. The thousands of text search terms are entered as queries for the phonetic search engine, and the phonetic searches are run. The phonetic search results are then compared to the time-coded transcript. Results are automatically classified against the truthed transcripts and the results graphed.

While basic phonetic searches yield good results, there are ways to increase both precision and recall. One way is to carefully select search terms. Regardless of audio recording quality, precision and recall increase with an optimal number of phonemes in a search string. Precision increases even more if searchers are able to identify search strings that are phonetically unique – that is, they don't sound too similar to commonly used terms. Increasing the length of the phonemes in a search string increases the uniqueness of the search term, even if one of the words in the string is commonly used.

Nexidia analysts also use tools such as a Pronunciation Optimizer, which analyzes the phonetic content of selected spoken words and optimizes the phonetic representation of the term in order to use it as a search term. This tool is very effective when speakers pronounce the same word in different ways, or it is not clear to the searcher how to spell a search term in a way that yields the best results. Again, the results can be empirically tested, the results graphed to illustrate precision and recall, and the tests documented so that the users can demonstrate that they got the best results possible from an audio collection.

Using this methodology for assessing the accuracy of phonetic search on a given data collection, the parties and court can test the search tools, negotiate search parameters, and trust the results of audio search. A party that disagrees with the approach would be required to produce empirical proof that its alternative methodology or technology provides better results, taking into consideration the balancing factors of cost, time, and accuracy.

[20] The criteria may include, for example, requirements that all query terms contain three or more letters, occur at least twice in the transcript, and phrases that include more than one word do not span sentence boundaries.

CONCLUSION

The new paradigm of "mega-discovery" requires heavy use of search and retrieval technologies. As more discovery is done by computerized systems, those products will be under growing pressure – from competitors, customers, and courts – to prove their effectiveness. Empirical data should serve as the basis for selecting search and retrieval technology, but until courts develop guidelines for benchmarking performance, parties and vendors will have to develop defensible test protocols. It will be up to the parties to ensure that they are meeting their discovery obligations, and if necessary, demonstrating that they have done so by providing empirical test data and results.

In future disputes over search and retrieval methods (and absent case law or benchmark standards), the courts will continue to make decisions balancing factors of time, cost, and accuracy, just as they always have done with discovery disputes. As disputes arise, courts may be called on to look more deeply into the effectiveness of underlying search technologies, even in the absence of agreed upon standards. A party that can produce defensible test results stands in the best position to prevail in such a dispute.

A responding party that proposes to meet its discovery obligations by using an advanced search technology like Nexidia would be well-positioned to defend against a challenge to its use, since the objecting party would bear the burden of both refuting test results and coming up with an alternative that provides a better balance of cost, time, and accuracy.

A requesting party that is dissatisfied with its opponent's proposed approach to audio discovery (e.g., a human-listening method that would be extremely slow, expensive, and lacking test data establishing the accuracy of the reviewers) would be well-positioned to challenge that approach with a faster, less expensive, and demonstrably accurate alternative.

Nexidia provides audio search tools whose overall performance is tested and documented, and whose accuracy can be demonstrated if necessary on a case-specific basis. Both requesting and responding parties can benefit from good benchmarking to test the efficacy of their automated litigation search and retrieval systems. Better search means a more cost effective way to a more complete view of the evidence.