Breaking Through Babel Without Breaking The Bank

Gary Wiener

Liquid Litigation Management

Those who are familiar with the Biblical story of the Tower of Babel cannot help but marvel at the way the Internet has erased the ancient boundaries among most nations. Indeed, the current global economy requires most businesses – and certainly all large ones – to participate daily with markets in other nations. And, thanks to the dirt-cheap cost of computer storage space (as I write this, one-terabyte hard drives are selling for as little as $139, or less than 14 cents per gigabyte), the volume of digital data that businesses must deal with continues to grow.

Unfortunately, the Sedona Conference estimates that a single 14-cent gigabyte of storage space can easily take 150 hours to review.[1] Multiply that times several hard drives, and the numbers can put even the biggest companies' legal budgets on tilt. Can costs of reviewing and producing ESI be kept to a reasonable level?

The preferred method of keeping costs manageable is to run keyword search queries across the documents, to cull the mountain of data down to a much smaller size. Much has been written about concept searching as a preferred alternative to keyword Boolean and fuzzy searching. It's true that keyword searching is prone to return quite a few false "noise" hits. However, a closer look at the pro-concept search articles reveals that they tend to be white papers written by – yep – vendors of concept searching solutions.[2]

The fact remains that "the available evidence suggests that keyword and Boolean searches remain the state of the art and the most appropriate search technology for most cases."[3] Concept searching has simply not evolved to the point that it is a reliable replacement for keyword searching.

The ultimate search query strategy, however, is to consider querying to be an evolving process, every step of which must be thoroughly documented. With careful planning, keyword search queries can be made as effective as possible.

The Search Management Process

In order to get keyword search queries "right," their creation should be thought of as a process, not as a one-time exercise. Processes are fluid and adaptable to changing circumstances, whereas a single attempt at culling, a single time, is not likely to provide anything but frustration. There are several steps to creating an effective search process.

Under Federal Rule of Civil Procedure (FRCP) 26(g)(1), an attorney must certify that to the best of her "knowledge, information and belief, formed after a reasonable inquiry," disclosures are "complete and correct" at the time they were made. For the most part, judges have not issued sanctions because disclosures were imperfect; they have sanctioned because the attorneys could not demonstrate good faith in producing those disclosures.[4] Having a thorough, documented and (above all) good-faith process for culling responsive documents is always going to be the best defense to a motion for sanctions.

***Gary Wiener** is the Director of Litigation Services for Liquid Litigation Management. A former trial lawyer, he advises clients in the effective use of technology to streamline their litigation management practices.*

The Importance Of Testing

In *Victor Stanley v. Creative Pipe*,[5] Magistrate Judge Grimm threw out attorney-client privilege claims as to 165 inadvertently-produced documents. Even though counsel had applied 70 separate search queries to the document universe to filter out privileged documents, the court ruled that they had failed to conduct "quality assurance testing." Clearly, making up queries off the top of one's head – effective though they may be – is not an acceptable method of building a search methodology.

Begin by isolating the key custodians in the matter, and before running any search queries at all, isolate a statistically-significant sample (say, five to 10 percent) of their documents. Reviewing every single page of every single document is not cessary. As one court has observed,

> [T]here is no obligation on the part of the responding party to examine every scrap of paper in its potentially voluminous files in order to comply with its discovery obligations. Rather, it must conduct a diligent search, which involves developing a reasonably comprehensive search strategy. Such a strategy might, for example, include identifying key employees and reviewing any of their files that are likely to be relevant to the claims in the litigation.[6]

Review those documents for relevance and privilege, but note recurring keywords or themes in the responsive documents. Not only will this provide a useful benchmark against which to compare search results against the entire document corpus, but it will educate the legal team as to what to expect from the rest of the documents.

Once the sample documents have been isolated, begin testing search queries based on the results of the manual review. How do the results of the search queries compare to the number of times those results were found manually in the sample set? If there are too many "noise" hits, determine what might be an effective way to limit the noise. (Boolean searching allows for proximity searches, in which one word of the query must appear within a set number of words of a second word in the query in order to be returned as a search hit.) Test a number of different keyword searches, and rank their efficiency (or deficiency).

Also, the resulting familiarity with different types of data sources will allow counsel to predict the costs of retrieval, backup and review of such data.

For example, the sampling process may reveal that certain data sources (such as local hard drives) or file types (such as Microsoft Access files) have such low yield that the collection and review effort is not worthwhile. While opposing counsel may not agree to such decision, the data provided by the sample review will provide the evidentiary support needed to defend the reasonableness of such steps to the court.[7]

After the most effective queries have been identified through testing and ranking, and you are confident that the search plan is defensible in court, you can apply the results to the entire corpus of documents. You can do so with confidence that what you've done is consistent with the due diligence and good faith requirements of the discovery process and the FRCP.

However, the single most important part of this process is the most frequently overlooked. Quite simply, you ***must*** document every single step of this process.[8] By doing so, you can successfully defend your good-faith search process to the court. Without doing so, you may have no defense to sanctions at all.

Searching The Tower Of Babel

The techniques used for creating effective English-language keyword queries apply to foreign languages as well, although languages such as Chinese, Japanese and Korean (the "CJK" languages) require special considerations. When a computer indexes an English-language document, it "sees" everything between spaces or punctuation as a "token", which is then catalogued in an index file. The index file is what a search engine actually reads to produce search results. CJK languages typically do not use spaces between words, which means that tokenization software must be customized to each language using pattern matching. Without the proper tokenization software for the language, keyword searching is impossible.

Also, consider that frequently, CJK documents will contain more than one language or alphabet. Japanese writing, in particular, frequently contains Chinese characters, and those characters do not necessarily mean the same in Japanese as they do in Chinese. Most languages use Arabic numerals. It's not uncommon to see all three types of characters in the same sentence. It's also not uncommon, when dealing with right-to-left languages such as Hebrew or Arabic, to see left-to-right strings of Arabic numerals in the middle of the sentence.

Once the foreign-language documents have been properly tokenized and indexed, the computer's search engine works much as it does for English searches. The computer looks for a matching string of characters, regardless of the language in which those characters are written. Without the correct tokenization software, however, this function is useless. The tokenizer must be able to identify in which language each token is written, and in which direction it should be read; and it must be able to change its methodology on the fly. Ideally, the system you choose would then be able to accept a search query containing keyword search terms in multiple languages.

Without treating the creation of effective keyword search queries as an ongoing process, the expense of review and production may well pale compared to the cost of the sanctions that the court may impose.

Conclusion

Don't lose sight of the fact that the goal of managing all of these aspects of search is to save time and money – and a lot of both. Without treating the creation of effective keyword search queries as an ongoing process, the expense of review and production may well pale compared to the cost of the sanctions that the court may impose.

Please email the author at gwiener@llminc.com with questions about this article.

[1] *Jason R. Baron, Ed., Search and Information Retrieval Methods*, The Sedona Conference Journal *191, 192 (2007).*

[2] *An apparent exception would be Judge Facciola's opinion in* Equity Analytics LLC v. Lundin, *248 F.R.D. 331 (D.D.C. 2008), in which he wrote that "determining whether a particular search methodology, such as keywords, will or will not be effective certainly requires knowledge beyond the ken of a lay person (and a lay lawyer) and requires expert testimony that meets the requirements of Rule 702 of the Federal Rules of Evidence." 248 F.R.D. at 333. However, the matter at issue in* Equity Analytics *was illegal access to electronic documents by a fired employee, and therefore the context of Judge Facciola's observation was most likely in the context of forensic examination of the questionable data, which requires specific expertise.* See *Debra R. Bernard and Mary Rose Hughes, "Much Ado About Text Searching",* National Law Journal *(Aug. 28, 2008).*

[3] See *H. Christopher Boehning and Daniel J. Toal, "Assessing Alternative Search Methodologies,"* New York Law Journal *(April 22, 2008).*

[4] See, e.g., *Fed. R. Civ. P. 26(f), Fed. R. Civ. P. 37(f). Keep in mind, though, that the court is not likely to take a claim of "good faith" strictly at face value.* See, e.g., Qualcomm Inc. v. Broadcom Corp. *(S.D. Cal. 05-CV-1968-B, Aug. 6, 2007) at 38, 51 (although Defendant claimed good faith and a reasonable search for responsive documents, they failed to produce over 200,000 pages of responsive ESI found with what the Court called "such an obvious search" query. The Court later sanctioned Defendant over $8.5 million in legal expenses, and referred six attorneys to the California State Bar for disciplinary action).*

[5] *250 F.R.D. 251 (D. Md. 2008).*

[6] Treppel v. Biovail Corp., *233 F.R.D. 363, 374 (S.D.N.Y.)*

[7] *Boehning and Toal*, supra *n. 3, at n. 8.*

[8] *Even though a party's search methodologies are arguably protected from disclosure by the attorney work-product privilege, in three recent cases, the court required attorneys to explain and defend to the court its methods.* See Victor Stanley, supra *n. 5, at 256;* U.S. v. O'Keefe, *537 F. Supp. 2d 14, 23-24 (D.D.C. 2008); and* Equity Analytics, LLC v. Lundin, *248 F.R.D. 331 (D.D.C. 2008). In* Victor Stanley, *the court found that by failing to disclose the keywords used to search for privileged documents, defense counsel had waived its privilege.* Victor Stanley, supra.