

WHERE ANGELS FEAR TO TREAD: ELECTRONIC DATA SEARCH PROTOCOLS

By Sharon D. Nelson, Esq. and John W. Simek
© 2008 Sensei Enterprises, Inc.

The story is almost always the same. A computer forensics or EDD company receives computers, thumb drives, cell phones, etc. along with search instructions from an attorney. The search instructions are almost always a list, written in plain English without any “search protocols” to refine the search. Here are some examples:

1. System. Now you’ve to be kidding. Do you know how many times that word will appear on ANY computer.
2. Pot. Which will get you “potential,” “potting soil” or even “Spot” unless constructed properly.
3. Ted. Which will get you exhausted, depleted, and well, you get the picture. Can you get a list of Ted’s e-mail addresses, which would be ever so much more helpful?

The silly lists EDD experts receive have been a great source of private merriment to them. The industry sometimes jests that “we’ll do our best to give the client what they wants, not what they ask for.” Fellow computer forensic technologist Craig Ball recently said in an interview “The days of ‘let’s try these search terms and see what happens’ are numbered. Queries that will be run across mushrooming collections must pass muster in terms of noisiness, ambiguity, potential for misspelling affinity to stemming, synonyms, slang, acronyms, IM-speak and other criteria unfamiliar to a profession that prides itself on precise expression.” We couldn’t agree more that lawyers – and judges – are ill-prepared for this very complicated task.

Searching complexity is so great that Magistrate Judge John Facciola, in *U.S. v. O’Keefe* (D.D.C., Feb. 18, 2008), calls keyword search analysis an area of e-discovery “where angels fear to tread.”

Defendant O’Keefe was a government official charged with taking bribes from co-defendant Agrawal for expediting visas for Agrawal’s employees. The court had previously ordered the government to conduct a thorough and complete search of its hard copy and electronic files in “a good faith effort to uncover all responsive information in its custody or control.” After receiving the government’s submissions, the defendants complained that the search was anything but thorough, incomplete and not a good faith effort.

It seems likely that the government’s Visa Unit Chief, who performed the computer searches along with her five member staff, probably was unqualified to construct those searches, though her qualifications were never specified. There was no mention of attorney involvement. Key players were not interviewed. The government did not disclose the software it used or the search terms (and their construction) that were used. However, Judge Facciola declined to rule for the defendants, pending expert testimony in

a motion to compel. It was his view that this indeed is an area where judges and attorneys should “fear to tread” and without expert testimony, he was uncertain he could categorically say that the government’s search terms were inadequate without listening to experts.

If that case doesn’t argue sufficient for expert involvement in the beginning of the EDD effort, try reading *The Sedona Conference, Best Practices Commentary on the Use of Search and Information Retrieval* (2007). Not only will you head hurt, but you will probably begin to regard search protocols as a tar pit that any sensible dinosaur would scrupulously avoid.

An Introduction to The Tar Pit

Veterans of electronic data discovery know the jargon and all the tricks of the trade to rapidly deal with the massive amounts of information in electronic form. What’s the key to efficient handling of electronic evidence? One of the first steps is to reduce the amount of information that has to be reviewed. The most costly part of any case dealing with electronic evidence is the attorney review time. Reducing the amount of information reduces the time needed for review. It’s a simple equation: Less data=less money.

Mind you, there are some vendors that won’t steer you that way, especially if they charge by the volume of data that searches return. Ethical EDD vendors (be sure you check out references and get referrals!) will always be looking to save you money by skillfully crafting search methodologies to trim the volume of data that must be reviewed, without, of course, losing relevant data in the process.

There are so many ways to search for relevant data, but which method should you use? Are there benefits to a particular search method? You may be wondering what search methods there are, especially if you are very new to the handling of electronic evidence. Judge Facciola’s opinion in *O’Keefe* suggesting that lawyers and judges are not capable of properly searching without using an expert woke a lot of attorneys up, making them resolve to “get smarter” on the subject of searching. Do you really need an expert? Sorry, yes. But it helps if you can ask probing questions to make sure you’re getting good advice and it also helps to have a basic understanding of search methodologies. So here’s your guide to the nuts and bolts of searching, with some pros and cons thrown in for good measure. Strong coffee may be required here – this stuff isn’t for the faint of heart – or those who regard themselves as technophobes.

Constraints

What are some of the restrictions and requirements for searching? No matter what data you are analyzing, the first limitation is the capabilities of your search software. What file types can the search software handle? Can it search the internals of an Outlook PST file? What if your e-mail messages are stored in a Notes NSF file? What about attachments? Can the software search for particular terms in a Word document, even if it is an attachment to a message? Can you search the information contained in compound files

such as ZIP files? These are the types of questions that must be answered to understand what limitations may exist in the search software. Not all search software is created equal – and the use of inferior software may make deep inroads on your client’s wallet.

What about foreign languages? Is all of the information in English or do you have to do searching with some other language? This may present a particular challenge, especially if you have to use a foreign alphabet to search the data. Foreign language searching is just beginning to come into its own – and can create considerable expense.

The handling of e-mail is a great concern since that is usually where a lot of the relevant information resides. Many search software applications will not search the attachments as part of the native file. You may have to extract the attachments from the messages prior to searching. No matter what method you are faced with, make sure you know if all the data is being searched.

Basic Types of Searches

So let’s get to the meat of searching. What are the various ways that the data can be searched? Certainly the most common way to search is through the use keywords. What exactly is a keyword? Generally speaking, keywords are words or sections of words that would be contained within the data. As an example, searching for the keyword ‘pot’ could return the words *spot* and *potential*. You may be able to narrow in on the results by using operators of the search engine. Operators tell the search engine how to handle the keyword and any variants. Most of us are familiar with the wildcard symbol (*) or the single character symbol (?). As an example, selecting the keyword as pot* would return the word *potential* but not *spot*. This is because we placed the wildcard at the end of the word and not the beginning. Make sure you know and understand what operators are available for the search engine.

Searching with phrases can also help reduce the amount of “noise” results. Noise results are unintended search results. Obviously, reducing the noise hits means less review time, therefore less money expended in attorney review. Another type of search method is called Boolean searching. Boolean operators are terms such as AND, OR, NOT and NEAR. As an example, you may construct a Boolean search to be ‘Everett AND Washington.’ The NEAR operator means that the words are in the exact order as listed. Some search engines will use quotation marks around the words to indicate that an exact order is required.

Proximity searching is another common method. The operator WITHIN or the plus symbol is used by many search engines to denote a proximity search. An example of this would be ‘apple +5 pie.’ This means the word pie has to appear within 5 words or less of the word apple.

Another advanced ability of searching is stemming. Essentially, stemming is finding variations in the endings of a selected search word. An example of stemming using the

word metal would return *metalized* and *metallic*. This is the same as putting a wildcard operator at the end of the keyword.

GREP Searching

There may be an occasion when you hear the term GREP used when referring to searching. GREP came from the UNIX world meaning Global Regular Expression Print. When using GREP, you'll search through the data for a specific character string pattern. The syntax for GREP can be a little complicated and daunting. Wildcard and placeholder values exist in GREP along with several other operators. As an example the brackets mean to include any character within them. So [a-f] means to include lower case a through f in the character position. Normally you will see GREP used in forensic analysis searches and not in typical electronic data processing. After reading this paragraph, you're probably glad you won't have to deal with it often.

Fuzzy Searching

Some search engines also support fuzzy searching. Fuzzy searches find misspellings in words. This can be particularly effective to find results where words are typically misspelled. Spelling errors are pretty common in cases involving technical terms. The degree of "fuzziness" is normally adjusted via a numeric value. Many search engines will use the number to determine the amount of letters that can be wrong in the misspelling. As an example, a fuzzy value of 1 would mean that only one letter can be wrong in the word and a value of 3 means three letters can be wrong.

Fuzzy searching is a very valuable ability when dealing with sophisticated terms, where the proper spelling is not widely known. How many people spell *pseudonymous* correctly? While fuzzy searching can find data that would normally be missed, it can also generate a lot of "noise" results. Some vendors charge based on the number of search hits so the addition of "noise" results adds to the invoice. Fuzzy searching is like casting a wide net to gather much more information than a direct keyword search.

Conceptual Searching

An extremely powerful (and expensive) capability is the usage of conceptual searching. Concept searching takes the input term and returns results that are related in meaning. This is best explained in this example. If the term 'car' was used, then results of 'automobile' and 'vehicle' would be in the returned hits.

One of the most well known of the concept search engines is Attenex. The Attenex engine is very powerful, but not for the faint of heart. The cost is very high and is typically used in large, high profile cases where there is a large volume of electronic information. The key issue to be aware of with conceptual search engines is that they are only as good as the programming logic. There is some element of artificial intelligence while doing conceptual searches. You are dependent upon the programmer's view of

what term has a related meaning. Some engines allow you to modify your own thesaurus file so that there is some element of control in the results.

Keeping the Tar Pit at Bay

An exhaustive look at searching would require a small book. But keep this “nuts and bolts” guide handy next time you represent someone in a case involving ESI. And run, do not walk, to retain an expert. The lawyer who is a knowledgeable consumer of EDD services or has an expert to assist her is going to save her client money – and grateful clients usually return!

*The authors are the President and Vice President of Sensei Enterprises, Inc., a legal technology and computer forensics firm based in Fairfax, VA. 703-359-0700 (phone)
www.senseient.com*