D4

e-Discovery

Vantage Point

About the Authors:

As Vice President of D4, Peter Coons is focused on implementing innovative e-Discovery solutions for D4 and its clients. Peter has nearly fourteen years of experience in the legal technology field and holds multiple industry certifications related to computer forensics. He has testified as a computer forensic expert witness and has been appointed by a District Court as a neutral third party examiner. He has vast experience orchestrating large scale electronic data collections relating to litigation, audits, or regulatory requests.

Tom Groom is Director of Discovery Engineering at D4 and has over 15 years experience with litigation support, document review and e-Discovery strategies. Tom advises both corporate and law firm clients on issues involving ESI, including defensible collection methodologies, optimized ESI processing and review workflows, and litigation contingency and readiness planning. He has supported cases involving mergers & acquisitions, antitrust, securities, environmental, patent infringement, products liability, as well as other general litigation.

Uncovering ESI: 20 Search Term Tips

By Peter R. Coons & Tom Groom

Over the years we have keyword searched thousands of hard drives, e-mail stores, thumb drives, CD's and servers. Using keywords to identify potentially relevant documents is a well established practice in the e-Discovery world. When combined with other methods search terms can be a powerful means for culling down a large dataset. When implemented improperly they can cause major headaches. To help on the path to keyword nirvana we have outlined 20 helpful (we hope) tips below.
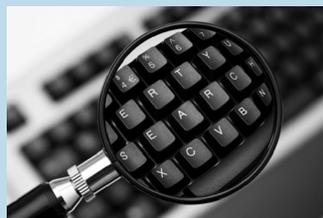
False positive defined: A search term "hits" within a document but not for the meaning that was intended. For example, the term "comput*" would return "computer" (the intended term) but would also return "computational" (not intended). Computational would be the "false positive" term.

Below is a list of tips and tricks

1. Any term less than four (4) characters may result in a lot of false positives. Clients have asked to search for "IT" (information technology) and then wonder why they are getting thousands of false positives.

2. Be aware of "noise word" lists that are being used during the searches. Some software applications don't index the word "it" or "up" for example, so your attempt to find the key phrase "pick up" may fall down. Most noise word lists can be customized.

3. Be aware that *searching numbers* can sometimes return unwanted results. Often we are asked to search for patent numbers such as 1,234,567. If this term is not quoted properly the result may be way off. Try using the word "patent" in conjunction with the number. Be aware that searching for 1,000 will also return 1,000,000 or it could return 2.10,1.000,85697..021. Make sense? Exactly.

4. Don't use wildcards unless it's absolutely necessary. If you want to find DOG or DOGS then don't use DOG* as a search term. Simply provide both variations of the word. If you must use a wildcard then please refrain from leading with a wildcard character. You

may get the result you are looking for but you will bring a lot of unwanted garbage with it.

5. Searching for names of custodians will return a lot of hits if that custodian is part of collection. Usually, all of the documents for that custodian. Same thing with company names or subsidiaries.

6. Before deciding on search terms with the opposing party please try to actually sample documents with the proposed terms. This may seem obvious but this advice is followed about 5% of the time.

7. Searching dates – There are lots of dates associated with ESI. There are created, modified, accessed, sent, received, etc. If the ESI was not forensically collected and instead was collected by the custodians and "dropped on a server" don't be surprised when you find ZERO documents prior to 1/1/2009. The metadata has been obliterated.

8. What are the expectations? Do you expect a 10% return rate and you are getting 90% or vice versa? If so, there may be an issue.

9. Don't request "fuzzy" searching unless you understand what exactly is being requested.

10. DeNISTing does not get rid of all EXE, DLL, and system files. Not exactly related

to searching but we had to throw it in here.

11. Not all logic is the same for all search engines. For example, some may use the "w/" proximity operator and others use "near". Ask the provider or operator to explain the logic and syntax that is required for the software being used.

12. Many characters are traditionally indexed as spaces (e.g. *!@"#$&'()*+,./:;<=>?[\5c]^`{|}~*). This means that "pcoons@d4discovery.com" is indexed as three separate terms: "pcoons" "d4discovery" and "com". The "@" and the " ." are considered spaces. If the characters listed above are all indexed as spaces then my e-mail address would be the same as searching for"pcoons! d4discovery=com". Searching for "D4-Discovery" or "D4 Discovery" will yield the same results if the "-" is indexed as a space.

13. "1,000″ is same as search "1 000" and the "word" 1,000,000 is three separate items in the index (1) (000) and (000) so two words and three entries/items. If we indexed "," as a comma and not a space then we could search for numbers like 5,195,508 but that would cause even greater issues with searching for other words.

14. When searching personal names, use the "w/2" proximity search between the first and last names. (Tom w/2 Groom) will pull back Tom Groom; Groom, Tom; Tom S Groom; Groom, Tom S.

15. Suggest expanding first names with known nicknames. "Bill Johnson" could be searched with ((Bill OR William OR Will) w/2 Johnson). You will obviously need to gather any special nicknames from the customer (only people in our office would know "Mr. Squeeze" would be David Lapresi).

16. It is a good idea to use all caps for connectors like OR and AND. It makes it easier to read and some engines require the connectors to be in all caps.

17. Many search applications like the use of parenthesis to separate unique terms or sets of terms. It also makes it easier to read and correct. Use quotes when you need to search a literal or a phrase. Sometimes the quotes will override the stop or noise words but not always. Here is an example of the use of parenthesis and quotes.((Pete OR Peter) w/2 Coons) OR ((Dave OR David w/2 Lapresi) OR (terminator) OR ("our leaders")

18. Suggest domain names for potentially privileged queries. The term ("lawfirm.com") for example would pick up all email addresses from that domain. This works well to identify communication with outside counsel. (Note the @ is treated like a space so you don't need an * at the beginning of the domain name.)

19. Avoid redundancy. The search ((Dog) OR (Dog w/2 Collar)) is redundant…the second term would already be picked up by the first term. However, the second term would be more limiting than the first term.

20. As shown in the previous example, you can use proximity searches to limit the returns if one of the words is common and returning too many false positives.