7-31-2007

The Internet Evolution: Savvy Research Strategies for Paralegals

Michelle Rigual
University of New Mexico - School of Law

Recommended Citation

Michelle Rigual, *The Internet Evolution: Savvy Research Strategies for Paralegals*, Paralegal Education Seminar (2007).
Available at: https://digitalrepository.unm.edu/law_facultyscholarship/180

The Internet Evolution:

Savvy Research Strategies for Paralegals

Albuquerque, New Mexico

July 31, 2007

UNDERSTANDING SEARCH TOOLS FOR MORE EFFECTIVE QUERIES

A. How do Search Engines Really Work?

1. Comparing Search Engines and Directories

In a nutshell: search engines are fully automated while directories have human editors. Search engines are huge databases (collections of web pages and other files) that have been assembled automatically by machine. Subject directories are usually created and maintained by human editors who review and select sites for inclusion in their directories; they also often annotate their list of resources. When you initiate a keyword search of a directory's contents, the directory attempts to match your keywords and phrases with those in its written descriptions.

Like the yellow pages of a telephone book, subject directories are best for browsing and for searches of a general nature. They are good sources for information on popular topics, organizations, commercial sites and products. When you'd like to see what kind of information is available on the Web in a particular field or area of interest, go to a directory and browse through the subject categories.

Directory editors typically organize directories hierarchically into browseable subject categories and sub-categories. When you're clicking through several subject layers to get to an actual Web page, this kind of organization may appear cumbersome, but it is also the directory's strength. Because of the human oversight, they have the capability of delivering a higher quality of content. They may also provide fewer results out of context than search engines.

Most directories do not compile databases of their own. Instead of storing pages, they point to them. This situation sometimes creates problems because, once accepted for inclusion in a directory, the Web page could change content and the editors might not realize it. The directory might continue to point to a page that has been moved or that no longer exists. Dead links are a real problem for subject directories.

Today, the line between subject directories and search engines is blurred. The most popular search engine is Google. However, Google also has a directory, which you can reach by clicking on the directory link. The directory is simply imported from the Open Directory Project. Similarly, Yahoo uses Google's search engine results but also has its own directory too.

2. Indexing – Spiders and Web Crawlers

Search engine databases are compiled by employing "spiders" or "robots" to crawl through web space from link to link. Once the spiders get to a web site, they typically index most of the words on the publicly available pages at the site. Web page owners also may submit their URLs to search engines for "crawling" and eventual inclusion in their databases.

Whenever you search the web using a search engine, you're asking the engine to scan its index of sites and match your keywords and phrases with those in the texts of documents within the engine's database. You are not searching the entire web as it exists at this moment. You are actually searching a portion of the web, captured in a fixed index created at an earlier date. Spiders regularly return to the web pages they index to look for changes. When changes occur, the index is updated to reflect the new information. The process of updating can take a while, depending upon how often the spiders make their rounds and then, how promptly the information they gather is added to the index. Until a page has been both "spidered" AND "indexed," you won't be able to access the new information.

While most search engine indexes are not "up to the minute" current, they have partnered with specialized news databases that are. To find late breaking news, look for a "news" tab somewhere on the search engine or directory page.

Search engines are best at finding unique keywords, phrases, quotes, and information buried in the full-text of web pages. Because they index word by word, search engines are also useful in retrieving tons of documents. If you want a wide range of responses to specific queries, use a search engine.

Search engines provide access to a fairly large portion of the publicly available pages on the Web, which itself is growing exponentially. They are the best means devised yet for searching the web. Stranded in the middle of this global electronic library of information without any recognizable structure, how else are you going to find what you're looking for?

On the down side, the sheer number of words indexed by search engines increases the likelihood that they will return hundreds of thousands of responses to simple search requests. Remember, they will return lengthy documents in which your keyword appears only once. Additionally, many of these responses will be irrelevant to your search.

Spider-based search engines have three major elements. First is the spider that visits a web page, reads it, and then follows links to other pages within the site. This is what it means when someone refers to a site being "spidered" or "crawled." The spider returns to the site on a regular basis, such as every month or two, to look for changes.

Everything the spider finds goes into the second part of the search engine, the index. The index, sometimes called the catalog, is like a giant book containing a copy of every web page that the spider finds. If a web page changes, then this book is updated with new information.

Sometimes it can take a while for new pages or changes that the spider finds to be added to the index. Thus, a web page may have been "spidered" but not yet "indexed." Until it is indexed -- added to the index -- it is not available to those searching with the search engine.

3. How “Relevant” Information is Retrieved

Search engine software is the third part of a search engine. This is the program that sifts through the millions of pages recorded in the index to find matches to a search and rank them in order of what it believes is most relevant.

All crawler-based search engines have the basic parts described above, but there are differences in how these parts are tuned. That is why the same search on different search engines often produces different results.

Search engines have a variety of ways for you to refine and control your searches. Some of them offer menu systems for this. Others require you to use special commands. The more specific your search is, the more likely you will find what you want. Don't be afraid to tell a search engine exactly what you are looking for.

4. Using Meta-Crawlers to Achieve More Comprehensive Searches

Meta-crawlers, also called meta-search engines, do not send out their own spiders to build an index. Instead, a meta-crawler enables the search to be run in several search engines at the same time. Results are gathered and grouped on the meta-crawler page. Each different meta-crawler features a unique combination of search engines and its own method of reporting results, so similar searches will produce different outcomes across meta-crawlers.

Meta-crawlers work especially well when no searching has been done in a subject area and the goal is to get an overview of what is available where. A meta-search can help determine the best direction for further research. No specialized functions unique to individual search engines work within meta-searches.

Some of the more popular meta-crawlers include:

- Mamma.com
- Search.com
- Dogpile.com.

B. Mastering the Rules of Boolean Searching and Advanced Search Functions

Boolean logic takes its name from 19th century British mathematician George Boole who wrote about a system of logic designed to produce better search results by formulating precise queries. He called it the "calculus of thought." From his writings, we have derived Boolean logic and its operators: AND, OR, NOT, and others which we use to link words and phrases for more precise queries.

AND. The Boolean AND narrows your search by retrieving only documents that contain every one of the keywords you enter. The more terms you enter, the narrower your search becomes.

truth AND justice
truth AND justice AND ethics AND congress

OR. The Boolean OR expands your search by returning documents in which either or both keywords appear. Since the OR operator is usually used for keywords that are similar or synonymous, the more keywords you enter, the more documents you will retrieve.

college OR university
college OR university OR institution OR campus

NOT/AND NOT. The Boolean NOT or AND NOT (sometimes typed as ANDNOT) limits your search by returning only your first keyword but not the second, even if the first word appears in that document, too.

saturn AND NOT car
rico AND NOT puerto

Nesting. Nesting, i.e., using parentheses, is an effective way to combine several search statements into one search statement. Use parentheses to separate keywords when you are using more than one operator and three or more keywords. For best results, always enclose OR statements in parentheses.

(hybrid OR electric) AND (Toyota OR Honda)

Boolean logic is not always simple or easy. Different search engines handle Boolean operators differently. For example, some accept NOT, while one accepts ANDNOT as one word, others AND NOT as two words. Some require the operators to be typed in capital letters while others do not.

Proximity. Proximity, or positional, operators (NEAR and ADJ) are not really part of Boolean logic, but they serve a similar function in formulating search statements.

Not all search engines accept proximity operators, but a few accept NEAR in their advanced search option. The NEAR operator allows you to search for terms situated within a specified distance of each other in any order. The closer they are, the higher the document appears in the results list. Using NEAR, when possible, in place of the Boolean AND usually returns more relevant results.

Advanced Searching. Nearly all search engines and professional databases offer "basic" and "advanced" searching. Advanced search refining options differ from one search engine to another, but some of the possibilities include the ability to use Boolean operators, to search on more than one word, to give more weight to one search term than you give to another, and to exclude words that might be likely to muddy the results. You might also be able to search on proper names, on phrases, and on words that are found within a certain proximity to other search terms. Some search engines also allow you to specify what form you'd like your results to appear in, and whether you wish to restrict your search to certain types of pages such as those from a university or from the government or to specific parts of Web documents such as the title or URL. Look at the home page for links that say “advanced search.”

C. How Ranking, Meta-Data and Search Engine Optimization Affect Your Searches

Search engine optimization is the process web developers use to increase the volume and quality of traffic from search engines to their web site. Usually, the earlier a site is presented in the search results, or the higher it "ranks," the more searchers will visit that site.

Early versions of search algorithms relied on webmaster-provided information such as the keyword meta-tag. Meta-tags provided a guide to each page's content. But using meta-data to index pages was found to be less than reliable, because some webmasters abused meta-tags by including irrelevant keywords to artificially increase page impressions for their website and to increase their ad revenue. Inaccurate, incomplete, and inconsistent meta-data in meta-tags caused pages to rank for irrelevant searches, and fail to rank for relevant searches.

By relying so much on factors exclusively within a webmaster's control, early search engines suffered from abuse and ranking manipulation. To provide better results to their users, search engines had to adapt to ensure their results pages showed the most relevant search results, rather than unrelated pages stuffed with numerous keywords by unscrupulous webmasters. Search engines responded by developing more complex ranking algorithms, taking into account additional factors that were more difficult for webmasters to manipulate.

Using Google as an example, the heart of Google’s software relies on the uniquely democratic nature of the web by using its vast link structure as an indicator of an individual page's value. In essence, Google interprets a link from page A to page B as a vote, by page A, for page B. But, Google looks at considerably more than the sheer volume of votes, or links a page receives; for example, it also analyzes the page that casts the vote. Votes cast by pages that are themselves "important" weigh more heavily and help to make other pages "important." Using these and other factors, Google provides its views on pages' relative importance.

Of course, important pages mean nothing to you if they don't match your query. So, Google combines its ranking system with text-matching techniques to find pages that are both important and relevant to your search. Google goes far beyond the number of times a term appears on a page and examines dozens of aspects of the page's content (and the content of the pages linking to it) to determine if it's a good match for your query.

D. Using Keywords Creatively

Before searching, try to imagine what the ideal page you would like to access would look like. Think about the words its title would contain. Think about what words would be in the first couple of sentences of a webpage that you would consider useful. Use those words, or that phrase, when you enter your query.

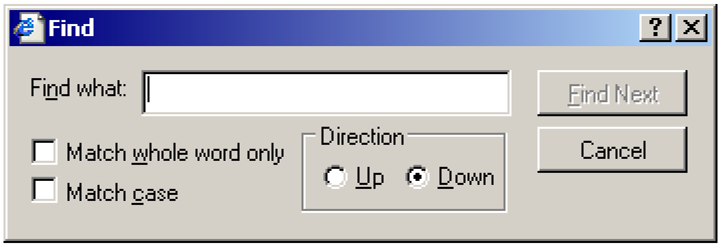
A search engine's ability to understand what you want is very limited. It will obediently look for occurrences of your keywords all over the Web, but it doesn't understand what your keywords mean or why they're important to you. To a search engine, a keyword is just a string of characters. It doesn't know the difference between cancer the crab and cancer the disease and it doesn't care.

Boolean operators and advanced search functions give meaning to your keywords. Here are a few other useful search strategies:

- Put your most important keywords first in the string.
- Use quotation marks around phrases to ensure they are searched exactly as is, with the words side by side in the same order.
- Type keywords and phrases in lower case to find both lower and upper case versions. Typing capital letters will usually return only an exact match.
- Use plus and minus signs in front of words to force their inclusion and/or exclusion in searches.

- Combine phrases with keywords, using the double quotes and the plus and/or minus signs.
- Use truncation and wildcards to look for variations in spelling and word form.
- Know the default settings your search engine uses (OR or AND).
- Know whether or not the search engine you are using maintains a stop word list. If it does, don't use known stop words in your search statement.
- In Boolean searches, always enclose OR statements in parentheses.
- Always use capital letters when typing Boolean operators in your search statements. Some engines require that the operators be capitalized; others will accept either so you're on safe ground if you stick to all CAPS.

E. Maximizing Browser Power – The "Find" Function and Managing Bookmarks



Find what:

☐ Match whole word only

☐ Match case

Direction ○ Up ● Down

Once you have found the Web page you want, it is easy to quickly scan the page.

Look to the top left corner of the webpage, click on the Edit menu, and then click on Find. A handy "find" box will appear in the middle of your screen. This box lets you type in a specific word. After you type in your search term, hit the Find Next button and your browser will look for that word on the page and take you right to it. Once you see the word, click Cancel.

Even easier, is holding down the CTRL and F buttons simultaneously. This shortcut will also make the find box appear.

Traditional Internet Bookmarks. Bookmarks (called "Favorites" in Internet Explorer) record an Internet URL so you can revisit it later without having to remember its address.

Bookmarks are one of the early innovations web browsers. Early on, bookmarks were simple lists of URLs stored in a "Bookmark" menu. They couldn't be sorted, and all new sites went to the bottom of the bookmark list. Today, modern browsers and third party bookmark managers provide a range of advanced management features.

A knowledge of Bookmarks enables you to easily revisit the sites you like without having to remember their exact URLs. You can collect lists of good sites, and build up a library of your favorite and most useful destinations.

Once you add a bookmark to your list, you simply select it from your browser's bookmark list to revisit the page. You add a page to your bookmarks by visiting it and then adding it with your browser's add bookmark feature.

A range of third party tools have been developed to help you manage bookmarks. Many provide import and export capability with the leading browsers. Different bookmark managers provide different capabilities, such as advanced organization and searching features, automatic checking for dead links, upload to the web, and other functions. You can find more information at the following sites:

- Google Bookmark Managers
- Open Directory Bookmark Managers
- Yahoo Bookmark Managers

Managing bookmarks consists of deciding on a set of categories in which to store your bookmarks. Most browsers come with a built-in set of categories and bookmarks, but you should create your own so they aren't cluttered by sites you don't use.

Guidelines for useful bookmarks:

- Create a set of categories for your major subject area interests.
- Move the categories you use frequently to the beginning, and arrange the rest in alphabetical order. - Drag items up and down in right-half of window.
- Add subfolders for subcategories when necessary; a general rule of thumb is to create subfolders when you get more than twenty bookmarks in a category. Keep in mind that you maximize a bookmark's visibility by keeping it at a high level, and you double the difficulty of finding a bookmark when you move it down one level.
- Put in separators to divide different sets of bookmarks.

HOW TO FIND WHAT YOUR SEARCH ENGINE CAN'T:
THE "INVISIBLE" WEB

A. What Information Is Available on the "Invisible" Web?

The "visible web" is what you see in the results pages from general web search engines and subject directories. The "invisible web" is what you cannot retrieve in the search results and other links contained in these types of tools.

Search engines cannot type or think. If access to a web pages requires typing, web crawlers encounter a barrier they cannot go beyond. They cannot search our online catalogs and they cannot enter a password or login.

A database is a collection of data. Most of the invisible or deep web is made up of the contents of thousands of specialized searchable databases made available via the web. When you type a search in a database, the search results are delivered to you in web pages that are generated just in answer to your search. Rarely are such pages stored anywhere: it is easier and cheaper to dynamically generate the answer page for each query than to store all the possible pages containing all the possible answers to all the possible queries people could make to the database.

There are many thousands of public-record, official, and special-purpose databases containing government, financial, logistical, and other types of information that is needed to answer very specific inquiries of interest to very few people. Even if stable links existed to such pages, search engines would not want them because they would because of the clutter they would create.

B. How to Find "Invisible" Documents

Invisible web search tools are slowly being designed. Their crawlers are set to identify and interact with searchable databases, aiming to provide access to deep Web content. Until the day these search tools are perfected, you have two options:

- Find searchable databases in general web directories and search engines by searching a subject term and the word "database."
- Use collections of material licensed by libraries and some industries for their users.

C. Guidelines for Online Database Searching

Every database will have a different interface but most share similar features such as search boxes, advanced search options, the ability to limit your search, site maps, and help screens.

Keyword searching. Just as with search engines, when searching online databases you want to combine keywords with Boolean operators to narrow or broaden your search results. “Help” screens can be particularly useful for discovering which Boolean operators the database recognizes. Sophisticated databases may have controlled vocabulary terms that describe the subject content of the document.

Search Results. The database will usually respond to your search request by displaying a list of hits. These may be links to webpages, documents records, or full text documents.

Organizing your Results. Databases may provide options for organizing your search results, including the ability to select, display, print, email and save your search results.

D. Where Do Old Web Pages Go? - Learn to Locate Them.

The web is constantly changing and web pages disappear or are changed without warning creating dead links and frustration. Even after a page has been modified or taken off the web there are ways to find it, access its content, and even link to it.

The Internet Archive is working to counter the transitory nature of information on the Internet through its Wayback Machine (http://www.archive.org). The Wayback Machine is the largest database in the world, containing multiple copies of the entire publicly available web. The database is created by using the same spiders employed by the commercial search engines. Each time the spiders crawl through the web they identify pages that have changed since the previous visit. Visitors to the site type in a URL and are presented with a list of archived versions of the site going back to 1996. Clicking on a date reveals the archived site. The archived pages can be linked to simply by copying the URL. Many of the websites that are part of the invisible web are also missing from this site. A site may not be included in the database if it is password protected, includes a command that won't allow the spider to search, or if the owner has requested it be removed.

The CyberCemetary (http://govinfo.library.unt.edu) provides permanent public access to the Web sites and publications of defunct U.S. government agencies and commissions. The old sites are listed alphabetically, by subject, and by region. The CyberCemetary was created and is maintained by the University of North Texas Libraries and the U.S. Government Printing Office, as part of the Federal Depository Library Program.

Search Engines also provide access to old web sites. As Google crawls the web it takes a snapshot of each page examined and caches these as a back-up in case the original page is unavailable. If you click on the "Cached" link, you will see the web page as it looked the previous time it was indexed. The cached content is the content Google uses to judge whether this page is a relevant match for your query.

When the cached page is displayed, it will have a header at the top to serve as a reminder that it is not necessarily the most recent version of the page. Terms that match your query are highlighted on the cached version to make it easier for you to see why your page is relevant.

The "Cached" link will be missing for sites that have not been indexed, as well as for sites whose owners have requested their content not be cached. Yahoo provides a similar service, and Gigablast has a link to the Wayback Machine’s archived versions of the pages.