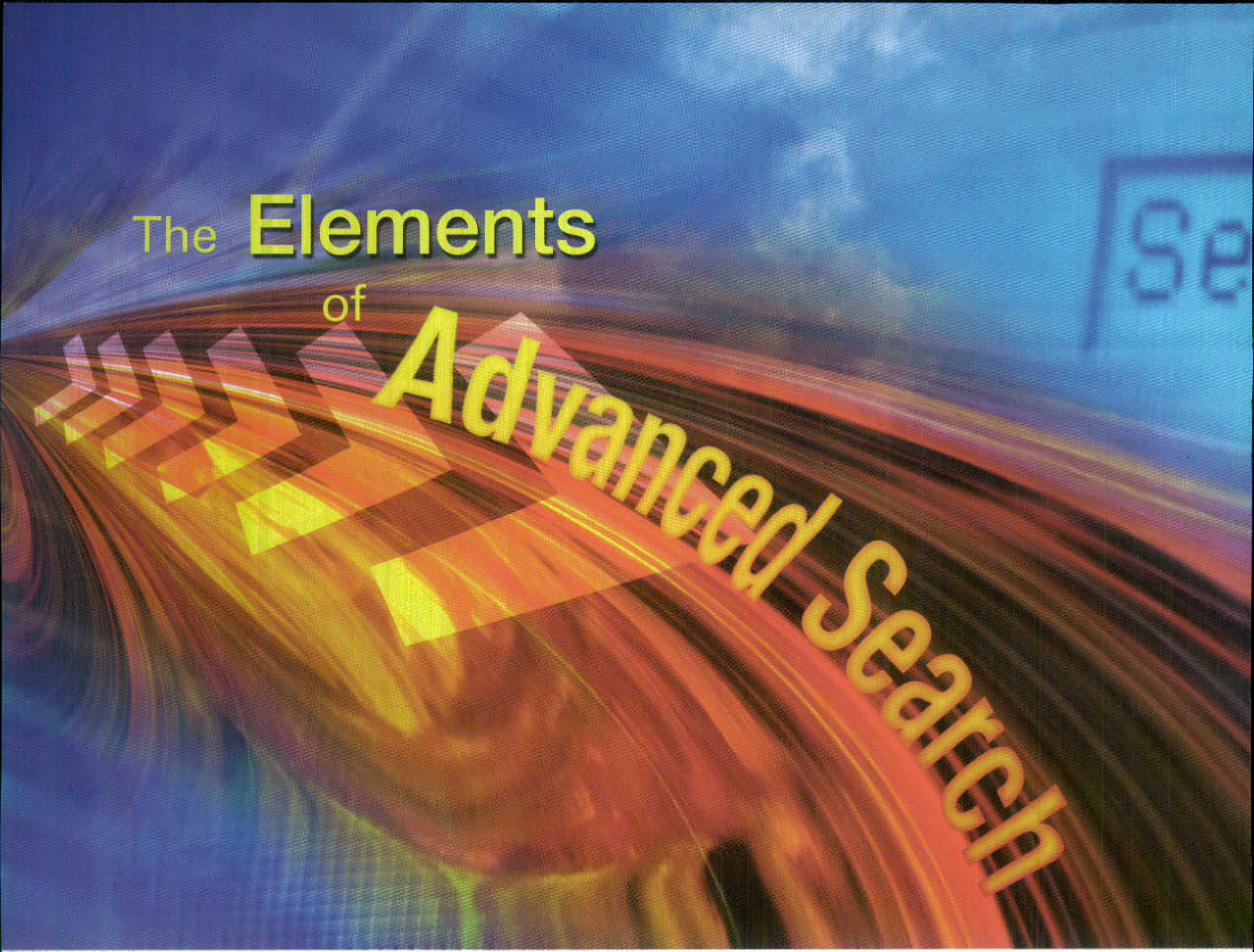
The Elements of Advanced Search



I'D GUESS that most *ONLINE* readers consider themselves advanced searchers and take advantage of advanced search options and techniques. However, even (especially?) for those well-versed in a topic, it is useful to step back and reexamine the elements of that topic to make sure that nothing is being missed, that we are fitting together the pieces optimally, and, perhaps, get some different perspectives.

Searching can mean anything from the mystical to the mundane. ("What is the true path to enlightenment?" "Where did I leave my keys?") I've narrowed that to what information professionals do—searching online databases and web search engines for serious research. When does searching enter the advanced realm? Is it anything beyond merely throwing a few words into a search box and accepting any answers returned? I'd go further and define advanced searching as using options offered by search engines to efficiently, effectively, and directly control the quality of search results. Direct control and implied conscious action by the user are critical to advanced searching.

The concept of quality of results is also critical and widely discussed. Advanced search goes beyond a specific "advanced search" page. It involves additional techniques, not just at the stage of the query entry, but in follow-up steps based on search results. It takes recall and precision into account—which in nontechnical terms can be thought of, respectively, as "Of the relevant material that is in the database, how much we found" and "Of what we found, how much is relevant"—and adjusts search behavior to improve both.

ADVANCED SEARCHERS

Advanced searching requires the advanced search capabilities of a search engine, but perhaps even more it requires an "advanced searcher," someone who brings to the search a particular combination of knowledge, skills, attitudes, and perspectives. Some of those attributes may be fairly specific, such as knowledge of a particular database, but can be more general, including personal qualities of curiosity and alertness to clues. (Numerous studies have been done to try to scientifically assess the importance of the attributes that make a good searcher, but even if you just examine your own searching habits, you will recognize that those attributes play a vital part in successful searching.) It is impossible to talk about advanced search without

by Ran Hock

talking about the advanced searcher. To paraphrase a well-known slogan, "Search engines don't search, searchers do."

The process and potential of advanced searching starts with the user, depends upon the search engine and its database, and has six key elements:

1. Searcher need
2. Searcher awareness/knowledge
3. Search strategy
4. Search engine intent
5. Structure of the database
6. Search engine features/options

SEARCHER'S NEED FOR ADVANCED SEARCHING

Clicking on an advanced search page requires a perceived need to do so or a significant "curiosity" factor. (Based on casual, nonscientific polling in courses that I have taught in a dozen countries, it continues to surprise me that so many users lack sufficient curiosity to cause them even to click on the "images" link on the main page of their favorite search engine.) Many searchers never use an advanced search link.

Casual searchers have modest perceived needs. They receive acceptable results with little effort. This is largely due to significant improvements search engines have made in content and ranking algorithms. For example, enclosing a phrase search in double quotes is still important, yet a search for a common phrase without using quotation marks usually retrieves the phrase in the first 10 records. Likewise, if there are records that have your terms in their title, those records automatically appear near the top of the list in search results, without you asking. Search engine developers have emphatically incorporated an "expert systems" approach to addressing the ranking problem; they have made their programs mimic the way an intelligent human searcher would approach and solve the problem. Searching has become more and more advanced without the user necessarily having to take the reins.

On the level of specific instances, how often does a user actually need advanced search options? For many searches, often not. When a very simple search provides the answers needed, simplicity is the best strategy. The most efficient advanced technique often is to take advantage of built-in relevance ranking. However, whether it is 5%, 10%, or 50%, there are times when users must be more explicit to get what they need.

>

SEARCHER AWARENESS/KNOWLEDGE

Awareness of search options—and that options might exist—is key to successfully finding the best, most relevant information. (Think of the many people who are still not aware of the riches that lie within their local public library and the ease with which they could find information just by asking a reference librarian.) Those of us who teach about the internet, or about other information resources, know that simply making someone "aware" of what exists is often more valuable than teaching specific techniques. How to find it comes next.

Casual internet users are probably not aware of easy ways to find a PowerPoint presentation on a specific topic, to find out who is linking to their website, to translate a webpage from another language into their own, or to do a "site search" on just about any site. "Not knowing what you don't know" is a major problem for casual searchers.

On a broad level, obtaining high-quality search results often depends upon a user having knowledge of the wider realm of information resources. Over-reliance on web search is a problem that is, unfortunately, much too familiar. Advanced searchers are aware of the databases that are part of the hidden web—the websites you must access directly, since that content is not indexed by search engines. The advanced searcher must also be aware of other databases that are in-house or available elsewhere—yes, even print resources.

SEARCH STRATEGY

All internet searchers have at least some strategy, even if it consists of just throwing a few words into a search box. Advanced searchers must take a truly strategic approach, applying a more careful analysis of the topic they are searching and what tools and techniques can best be applied.

SEARCH ENGINE INTENT

Think about search from the search engine companies' perspective: Advanced search options will not be provided unless there is an indication that a significant portion of searchers need them. Providing advanced search features requires effort, cost, and displacement of resources. If the perceived audience gets by nicely without an advanced search page, companies feel no need to provide one. If a feature such as a NEAR connector is not perceived by a search engine producer as something that would be understood and used, it won't be provided.

Professional searchers are not the main focus or the main audience of interest for most web search engine companies (Exalead may be a welcome exception). This factor, and the issue of unstructured data, discussed in what follows, are the primary reasons why web search engine features are not as robust. On the other hand, we advanced searchers are part of the long tail, a subset of users who, though small in percentage, is still significant in absolute numbers.

STRUCTURED AND UNSTRUCTURED DATA

How data is structured determines many of the possible advanced search capabilities. Take, for example, the number 1815. It can have just about any meaning (and consequently no meaning) unless that meaning is somehow identified. Sitting in a sea of text, to a computer it might mean a year, number of employees, population of a town, a classification code, or just about anything else. On the other hand, if 1815 is sitting in a financial report and that report is divided into parts that are specifically identified, the 1815 has meaning.

Context and meaning are provided by identifiers associated with a term. In database parlance, we ordinarily think of these parts of a record as its fields (the basic units of data within a record), with each field being labeled in such a way that a computer or user can know what each piece means in the context of the record in which it is found. If data is in a field, that field can be indexed and made searchable. It is the existence of fields that make specific items of data identifiable and searchable.

Whether the data in a database is actually sitting in discrete records, such as bibliographic records, or sitting in rows in a table, as in a spreadsheet, the concept of a field as a part of a record still applies. Advanced search options can provide the user with a way to associate an otherwise meaningless string of characters with an identifier that gives it meaning. Data that is well-fielded is referred to as highly structured. A good example is USPTO's Patent Database (www.uspto.gov/patft), which has more than 30 searchable fields, each with a very specific meaning. The more structure, the more searchable the data, and the more advanced search options that can be offered to the user.

In contrast to the highly structured data in the USPTO database, most webpages are unstructured. There are relatively few identified parts in a webpage that are consistent throughout all, or even most, webpages. When a computer examines a typical webpage, it can identify (and treat as fields) the title, URL, parts of a URL, headings, links, format, and maybe a few other parts. (In some cases it can also derive other fields, such as language, by using algorithms to figure out the main language of a page.) Some pages and websites may have other identifiable fields, especially if they employ XML (Xtensible Markup Language) instead of plain vanilla HTML. At the moment, however, such pages represent a very small a portion of all webpages. For general web search engines, the availability of searchable fields is currently rather limited and so are our field searching options.

SEARCH ENGINE FEATURES/OPTIONS

Whether it's general web search engines, search engines tied to a specific website, or specialized web search engines (such as news search engines), advanced searching does not necessarily require a separate advanced search page. You can often do an advanced search from within a single search box on a site's main page.

You can either use search syntax ("commands") from a main search box (or prompt) or menus (boxes, pull-down menus, checkboxes, etc.) as found on advanced search pages. The two most common categories of advanced search options are to use Boolean operators or field searching—searching specifically within designated fields.

BOOLEAN AND PROXIMITY SEARCHING

Boolean operators are essential to advanced searching. Advanced search pages spell it out even if they don't use the traditional Boolean AND, OR, NOT designations. If you enter multiple terms in a search box, web search engines assume the AND operator; the plus sign is now reserved for special cases, particularly to force a search for stop words.

All of the largest web search engines allow use of the OR. Live, Yahoo!, and Exalead require the use of parentheses around OR'd terms. Google and Ask don't care. The NOT is accomplished by use of the minus sign directly in front of a term. For other types of search engines, you will probably find that Boolean is dealt with similarly, but you may find other variations. Check the documentation for the site.

Proximity searching adds a specification as to how close together your search terms have to be. The most common instance of this is phrase searching, which insists that two or more terms have to be next to each other in the indicated order. Across the board, for all types of search engines, probably the most common and consistent search option is the use of quotation marks around a phrase to indicate that you want that specific phrase, rather than just the occurrence of those terms somewhere in the record.

Users of commercial database search services, such as LexisNexis, Factiva, and Dialog, know that for those services, proximity can go much further, with the capability of specifying how many words apart two or more terms can be, and whether or not they are in a specific order. Proximity for web search is far more limited. At present, only Exalead provides any other form of proximity: Use the NEAR connector. Its default is 16 words apart. For example, `aspirin NEAR heart` would retrieve pages where both words occur with up to 16 intervening words, while `aspirin NEAR/3 heart` would allow only three words between the terms.

MENU SEARCHING

Any advanced search page will provide you with a quick overview of the major search options that are available. A minute or two glancing at the options can be the most productive time spent on a search site. Don't underestimate the value of using the advanced search page as a tool for learning what options are available. Even for very experienced searchers, advanced search pages are a good way of finding out about new options that have become available.

Boolean is usually dealt with on advanced search pages by the use of boxes with labels such as "all of the words," "one or more of these words," and "none of these words," representing, respectively, AND, OR, and NOT.

However, the advanced search page may not allow you to combine options as you'd like. For example, just using the menus on either Google or Yahoo!'s advanced search pages, you cannot do two different sets of ORs—you cannot specify `forecasts OR predictions` and at the same time `China OR Chinese`.

SYNTAX (COMMAND) SEARCHING

Though a search statement such as `(EU OR "European Union") (accession OR admission) intitle:turkey site:news.bbc.co.uk` may seem cryptic to the uninitiated, for an experienced searcher, the use of search syntax can often be the easiest and most efficient way to accomplish an advanced search. The above query above combines Boolean with the searching of two specific fields (title and site). Not only can this syntax approach be faster than the fill-in-the-blanks approach on an advanced search page, it sometimes can also accomplish a combination of criteria that an advanced search page doesn't allow.

To use syntax effectively, you must know how it changes from one search engine to another. If you are searching a database on a website that you seldom use, it may not be worth the effort to learn the syntax. For general web search engines, it is worth the effort. You can determine the syntax by checking the search engine's help screens, or by searching on an advanced search page and then looking in the search box that appears on the results pages. Depending upon the search engine, there is a good chance that you will see your query "translated" into the syntax there.

FIELD SEARCHING

The real power and flexibility of advanced search is demonstrated when you combine Boolean and field searching, using either menus or syntax. Boolean can be applied to text terms, but also to various combinations of fields. The most common fields available for web search are discussed here. For other kinds of search (bibliographic, people, companies, patents, events, etc.) the searchable fields will depend very much on the specific content of the database.

Most field searching is accomplished by using a prefix directly in front of your search term (`intitle:Turkey`). (For a chart showing the prefixes for typical fields and other comparisons for major search engines, see www.extremesearcher.com/sechart.pdf.)

The prefixes used by various web search engines are similar, but not necessarily identical. Fortunately, even without any apparent cooperation among services, spontaneous standardization has determined which prefixes are used, for example, `intitle:` for title searching. For web search engines, expect some combination of the following searchable fields.

Title—Narrowing to those records that have a specific word or words in their title is by far one of the easiest tools for getting high-quality results. If an author chose to put a word in the title of an article (or book, patent, or webpage), there is usually no better indicator that the article book or webpage is very much about that topic. Consequently, searching by title can yield very high precision. All major web search engines offer the `intitle:` prefix for searching for a single word in the title and some, such as Google and Yahoo!, offer the `allintitle:` prefix for searching for multiple words.

Site—Many websites employ a search box to let you search that site. A similar "site search" can be done for virtually any website by using either a web search engine's advanced search page or by use of the `site:` prefix. Searching `cybertour site:extremesearcher.com` searches the Extreme Searcher site for the word "cybertour." You will frequently find that a search engine retrieves more pages than the site's search box. All major web search engines provide the `site:` option.

URLs—Similar to site search is URL search, searching for a specific term within a URL. Use the `inurl:` prefix (`inurl:aljazeera`). This can be particularly useful when you do not know whether a site has a .com, .net, or .org domain, or more than one (as does Al-Jazeera). Some engines also provide an `allinurl:` prefix search that allows you to specify two or more parts of a URL, not necessarily in a specific order.

Link—Among the parts of records that search engines can treat as searchable fields are the links found on a page. Using a link search to find what pages are linked to a specific site or page can be very useful for finding organizations (companies, terrorist sites, trade associations) with similar interests and that cover similar topics. You can also find out who is linking to your site.

File Type—The ability to search specifically for files not in HTML format—Adobe Acrobat, PowerPoint, Excel, and other formats—appeared only a few years ago. Though probably underused, it is a very powerful advanced search technique. There are situations when limiting your search to a particular file format can more directly retrieve the best answer. For a tutorial on how to use Photoshop's clone tool that will print out nicely, you are better off with a PDF, since printouts of webpages can often be very lengthy and somewhat messy.

Narrowing a search to PowerPoint presentations can be useful for finding a quick summary of a topic or just the key points, since that is the nature of many presentations. Searching for PowerPoint presentations can also be useful simply to avoid "re-inventing the wheel." If you want to review Boolean logic at the next staff meeting, there are already lots of such presentations out there on the web. Why start from scratch and spend a lot of unnecessary, valuable time if someone else has already done it for you?

If you are looking for statistics or other numeric data, consider limiting your search to Excel files. Your recall will be lower, but the precision of your search, and the speed with which you accomplish your goal, will be very high. Other types of formats can quickly provide specific types of data. For example, with Google you can limit your searches to Google Earth files (.kml or .kmz formats).

In Exalead, Google, and Live Search, you can narrow your search to a specific file type either on advanced search pages or by making use of the `filetype:` prefix, followed by the file extension that is associated with that format. For example: `photoshop clone filetype:pdf`. Less user friendly, Yahoo! uses the `originalextension:` prefix. Expect the filetype search to work with the most common file types, .pdf, .xls, .ppt, and so on, but don't hesitate to try other extensions. Though its documentation doesn't say so, in Google almost any type of file that might have a link on a webpage can be searched. (Do a search on `site:gov filetype:csv` to find more than 9,000 government databases that are online and downloadable.

Language—Language searching has some interesting nuances. To search "for" a specific language, you must search "in" that language. If you want a recipe in French for snails, searching on `snails recipe` won't do you much good. On the other hand, if you search for `escargot recette`, for the most part you will automatically get things in French and not even have to bother using the language search option found on advanced search pages.

Language searching highlights how advanced search options differ from engine to engine. Among the major engines, the number of languages that can be specifically searched ranges from 6 to almost 60. In most web search engines, you can search for the default "all languages" or for one specific language. On Yahoo!'s advanced search page, the use of checkboxes means that you have a choice of searching several specific languages at the same time. Language is also one field where you should use the advanced search page instead of a prefix. Some web search engines don't provide a prefix for this and if they do, they may use an abbreviation for the language. For less common languages you will need to see the list in order to know if that language is searchable in that engine.

Language searching reinforces the importance of user awareness and curiosity. Google's 2007 introduction of its enhanced language capabilities was, in my opinion, one of the most important advanced search options introduced in several years. Relatively little was written about the introduction and unless you were curious enough to click on the Language Tools link on Google's main page, you would be unaware of it. This tool lets searchers enter terms in their own language, translates the terms automatically into the target language, and shows results in both languages. (The usual caveat about machine translations: They will not be perfect and probably not pretty, but they usually give a good idea of what a webpage is about.)

Date—When searching for webpages, place date searching high on your list of options to forget about. Remember the significance of structured versus unstructured data and its effect on searchability? Most ordinary webpages do not have a section that can clearly be identified by web search engines as indicative of the date on which the content of the page was created. Even seeing the "modified" date for a webpage file does not reveal much. On the other hand, date searching is a highly effective and reliable technique when searching such things as news search sites and certainly for many other databases found on the web and elsewhere. Those resources will usually clearly identify how to go about limiting to a particular date or date range.

Other Fields—Web search engines often have fields additional to the ones above, sometimes clearly documented in the help screens, other times not. Microsoft's Live Search offers more than 15 search prefixes.

FIELDS WITHIN DATABASES

The above fields are the ones most commonly found in web search engines. For other types of searching, the nature of the database leads to its providing a variety of relevant fields to search. Many databases provide a menu and some provide syntax options. This applies to both free databases on the web and commercial databases. Take PubMed (www.ncbi.nlm.nih.gov/pubmed): It has a single search box on its main page, and the Limits tab takes you to an advanced search page with menus for searching by author, journal, text, abstracts, date, research pertaining to human or animals, gender, languages, topic subsets, type of articles, ages, and tags. Other links on the site lead to more specialized databases. Syntax is not emphasized.

For commercial databases, you are more likely to find a greater number of searchable fields and a greater emphasis on the use of syntax, due largely to the assumed user audience—professional, frequent searchers with a greater depth of background in searching and higher expectations and assumptions regarding searchability. Commercial services often provide options of very different search interfaces, for either the entire service or for portions of the service, with simpler interfaces for more casual searchers but with extensive advanced search options for the services provided to professional searchers. For example, in Dialog's command-driven (syntax-based) Dialog Classic, the MEDLINE database has more than 30 searchable fields and Extel Financial Cards has more than 70.

UTILIZING MULTIPLE SEARCH ENGINES

The advanced searcher recognizes that because of the differing content of search engine indexes and databases and the differences in ranking algorithms, using more than one web search engine or service is often required. Between different web search engines, the results that show up at the beginning of the results lists often differ significantly, and the advanced searcher recognizes that it is often important to go to a second or even third search engine to get the desired level of exhaustivity (recall). Looking at the first couple pages from two or three engines can actually improve precision.

Metasearch engines, such as Dogpile and mamma, combine results from multiple search engines. Some searchers find these useful, but do not be tempted to rely too much on them since many of them do not search the biggest engines (including Google). Even if they do, the first results you see may be ads. Most will return only the first few results from any engine, and advanced features cannot be used. (I hold the strong opinion that, in most cases, a searcher is better off going directly to the search engines themselves and looking beyond the first page of results or perhaps using a comparison search site such as Zuula [zuula.com], which returns results from six search engines).

ADVANCING FURTHER

An advanced search does not stop with the list of search results, or even with subsequent iterations, modifications, and improvements on the search query. If the task of the advanced searcher is to obtain the optimally precise and adequately exhaustive answer, there are often additional steps involved. This includes further processing and modification of search results using special features provided by the search service, such as sorting and translating. Look very closely at search results pages and individual records on those pages. The observant searcher will find, depending upon the search and the search engine, cached copies of a page, translations, similar pages, definitions, maps, suggested related searches, headlines, and links to matching results from other formats and media such as images, news, blogs, forums, and file types.

Even within web search engines, search doesn't stop with the web portion of the search engine. Advanced searchers also need to proceed to the additional databases, such as those for images maps, news, shopping, video, images, blogs, forums, and so on. Within each of these, there are usually special options. Each has the potential to contribute to the breath and quality of a search. A video of a lecture or interview on the current topic can often add a different dimension to the quality of results.

THE FUTURE OF ADVANCED SEARCH

The basic elements involved in advanced searching are not likely to change in the near future. The details will. More databases will come online and increased use of XML and other technologies, such as XBRL, may increase what is possible within databases and with web search engines, especially in alternate or specialized engines that search multimedia content such as video. More and more casual searchers will become advanced searchers, but the degree to which that occurs may depend on how effective information professionals are in increasing the casual user's awareness of what is possible.

Randolph (Ran) Hock *(ran@onstrat.com) is the principal of Online Strategies, which specializes in customized seminars for effectively using the internet and is the author of The Extreme Searcher book series.*

Comments? Send email to the editor (marydee@xmission.com).