Tutorial

BrightPlanet

Guide to Effective Searching of the Internet – 2005

December 2004

by Michael K. Bergman

About BrightPlanet

BrightPlanet Corporation, Sioux Falls, SD, Washington, DC, and New York City, is a private venture-backed company founded in 1999, though its technology legacy extends to the early 1980s. BrightPlanet's mission is to obtain value from document assets. BrightPlanet is the leader in deep document content and the development of innovative ways to efficiently search, monitor and manage all Internet and internal content. BrightPlanet offers unique technologies for discovery, harvest, management, aggregation, qualification, and classification of this content.

BrightPlanet products include collaborative, high-productivity solutions for professional knowledge workers as well as comprehensive federated portal sites that provide single access to pre-qualified content for individual companies, public agencies and associations. Prominent customers include the intelligence community, state and federal agencies, and Fortune 2000 companies. Three patents are pending with four granted for various aspects of BrightPlanet's automation, discovery and content management technologies.

Executive Summary: *Internet Search in Two Minutes*

To illustrate some of the basic concepts and recommendations covered in this tutorial, let's say we have an interest in recent findings about new planets being discovered outside our solar system. Using the information "contained" in this statement, you can see how an effective query can be built by following these guidelines.

We'll summarize the recommendation, show how the statement is phrased, describe why it's important, and provide a pointer to the specific topic number in the tutorial that covers this recommendation. See the table of contents for relating topic numbers to subject titles.

Recommendation	Example	Why Important?	Topic #
1. Use nouns and objects as query keywords	**planet** or **planets**	Actions (verbs), modifiers (adjectives, adverbs, predicate subjects), and conjunctions are either "thrown away" by the search engines or too variable to be useful	**6**, **7**, **8**
2. Use 6 to 8 keywords in query	**new**, **planet**, **discovery**, **solar**, **system**	More keywords, chosen at the appropriate "level", can reduce the universe of possible documents returned by 99% or more	**8**,**10**
3. Use appropriate word stem variants via the **OR** operator	**planet OR planets**	Word stem variants can increase coverage by up to 50%	**9**
4. Use synonyms via the **OR** operator	**discovery OR find**	Cover the likely different ways a concept can be described; generally avoid **OR** in other cases	**11**
5. Combine keywords into phrases where possible	**"solar system"**	Use quotes to denote phrases. Phrases restrict results to EXACT matches; if combining terms is a natural marriage, narrows and targets results by many times	**12**
6. Combine 2 to 3 "concepts" in query	**"solar system" "new planet" discovery**	Triangulating on multiple query concepts narrows and targets results, generally by more than 100-to-1	**20**
7. Distinguish "concepts" with parentheses	**("solar system") ("new planet*") (discover* OR find)**	Nest single query "concepts" with parentheses. (Overkill for now, but good practice when first learning.) Simple way to ensure the search engines evaluate your query in the way you want, from left to right	**19**
8. Order "concepts" with subject first	**("new planet*") (discover* OR find) ("solar system")**	Put main subject first. Engines tend to rank documents more highly that match first terms or phrases evaluated	**7**, **19**, **20**

Recommendation	Example	Why Important?	Topic #
9. Link "concepts" with the **AND** operator	**(("new planet" OR "new planets")) AND (discover OR discovery OR discovered OR find) AND ("solar system")**	**AND** glues the query together. The resulting query is not overly complicated nor nested, and proper left-to-right evaluation order is ensured	**14**, **20**
10. Issue query to full "Boolean" search engine	As above	Full-Boolean engines give you this control	

By issuing the query in **#9** above to Yahoo!, we are able to restrict results from a baseline of 20,700,000 documents using the query **new AND planet** (actually 22,200,000 if we were to properly include **planets** as well) to a count of 32,600 documents.[1,2,3] Though that number still seems like a lot, we have reduced our possible universe of results by about 600 times, and the first page of results listings is excellent.

Go ahead; try these queries for yourself!

Do you want to be able to get such impressive results for your own queries? Then, welcome. It's now time to start the tutorial.

[1] All citations are footnoted at the bottom of each page. Some footnotes refer to earlier citations on prior pages.
[2] Results counts used in this tutorial were conducted in early December 2004 and are based on the Yahoo! Advanced Search option (http://search.yahoo.com/web/advanced?ei=UTF-8). In the original version of this tutorial first published in April 1998, the AltaVista search engine was used. However, AltaVista has not kept pace with broad indexing of the Web and in early 2004 many of that engines advanced syntax features were removed from the service. Indeed, there has been a general reduction in advanced Boolean search features for all search engines over the past few years. See further the notes on search engine changes in the Introduction for more background.
[3] Here is a measure of document growth on the Web. When we first issued the **new AND planet (OR planets)** query in April 1998, the AltaVista counts were 418,934 and 551, 936, respectively. The recommended #9 query also returned 934 results. When previously updated in May 1999, these three figures were 917,754, 1,139,832 and 2,036, respectively. With the less precise numbers shown above from the December 2004 update, annual results growth have exceeded 210% per year!

Table of Contents

INTRODUCTION

Looking for that perfect condo for your ski trip? Needing specifications for a manufacturer's particular piece of equipment? Want discussion and commentary on your favorite, but obscure, author? Trying to find out what your competitors are up to? Seeking recent studies on planets in other solar systems? Needing information on special scholarships for which you might be qualified?

These, and millions of queries covering every conceivable topic, are now being posed daily to the Internet's search services. The Internet has become a vast, global storehouse of information. The only problem is: how do you find what you're looking for? [4]

Three Classes of Internet Information Look-up

Unfortunately, there is no Dewey decimal system or central "card catalog" for the Internet. You must use one or more search services to find new information. Search services come in one of three main flavors. Each has its place, depending on your information needs.

Directories

'Directories' use trained professionals to classify useful Web sites into a hierarchical, subject-based structure. Yahoo is the best known and most used of these services, though the largest is the Open Directory Project (http://www.dmoz.org). Directories are most useful when looking for information in clear categories, such as makers of yogurt or listings of educational institutions. Each directory uses its own categories and means to screen useful sites and assign them to a single category.

Search Engines

'Search engines' work differently. Google, AlltheWeb, Teoma, MSN Search and Yahoo! (through its search function) are some of the best known engines. They "index" (record by word) each word within all or parts of documents. When you pose a query to a search engine, it matches your query words against the records it has in its databases to present a listing of possible documents meeting your request. Search engines are best for searches in more difficult topic areas or those which fall into the gray areas between the subject classifications used by directories. But, search engines are stupid, and can only give you what you ask for. You can sometimes get thousands (millions!) of documents matching a query. Also, at best, even the biggest search engines only index a small fraction of the Internet's public documents.[5]

[4] All citations are footnoted at the bottom of each page. Some footnotes refer to earlier citations on prior pages.
[5] As of December 2004, Google indexed the largest portion of the surface Web with about 8 billion documents covered. See further, **PART 10: THE 'DEEP' WEB**, for more information.

'Deep' Web Databases

'Deep' Web databases look and act like search engines in that they have a text box for entering queries and serve up a results page with links dynamically. But they differ fundamentally in scope. Unlike search engines that attempt to have a broad reach for content across the variety of sites on the Internet, deep Web databases serve up content specific to the site's owner. Deep Web databases often support less powerful search syntax than broadscale Internet engines and are highly variable in their constructs and operation, though many are massive in content size. For example, the 60 largest deep Web sites alone contain 84 billion pages of content, or ten times more than what is indexed by the largest Internet search engine, Google.[6]

There are perhaps 350,000 individual deep Web content databases on the Internet, about 150,000 of which have unique, valuable content. Because their content can not be crawled by the spiders of the standard Internet search engines, their content can not be found through standard Internet search. **Part 10** of this tutorial addresses specific issues of finding and searching deep Web content.

The Central Role of Your Queries

Your ability to find the information you seek on the Internet is a function of how precise your queries are and how effectively you use search services. Poor queries return poor results; good queries return great results. Contrary to the hype surrounding "intelligent agents" and "artificial intelligence," the fact remains that search results are only as good as the query you pose and how you search. There is no silver bullet.

Most Internet searchers, perhaps including you, tend to use only one or two words in a query. Big mistake! Also, there are very effective ways to "structure" a query and use special operators to target the results you seek. Absent these techniques, you will spend endless hours looking at useless documents that do not contain the information you want. Or you will give up in frustration after search-click-download-reviewing long lists of documents before you find what you want.

General Approaches to Searching

While three quarters of the Web users cite finding information as their most important use of the Internet, that same percentage also cite their inability to find the information they want as their biggest frustration. The purpose of this tutorial is to help you end that frustration.

[6] M.K. Bergman, "The Deep Web: Surfacing Hidden Value," *BrightPlanet Corporation White Paper*, June 2000. The most recent version of the study was published by the University of Michigan's Journal of Electronic Publishing in July 2001. See http://www.press.umich.edu/jep/07-01/bergman.html.

The information professionals at the University of California at Berkeley recommend a graduated approach to Web searching.[7] Here's their stepwise sequence of steps to follow, which we generally endorse for beginning searchers:

1. ANALYZE your topic to decide where to begin
2. Pick the right starting place
3. Learn as you go & VARY your approach with what you learn
4. Don't bog down in any strategy that doesn't work
5. Return to previous strategies better informed.

As you gain experience, you can begin cutting out the middle steps. By the time you're doing real heavy lifting with your queries, you really only need spend some time first getting your query right and then cutting to the bottom line with a full Boolean search using phrases and three or so concepts linked through the **AND** operator and multiple search engines.

Fondren Library at Rice University has also published useful tips on Internet search strategies [8]. For advanced topics, and a resource that is increasingly focusing on Web-related topics, you may want to consult *Searcher: The Magazine for Database Professionals.*[9]

Following these guidelines, here are recommended steps to approaching the Internet search challenge:

- Spend time BEFORE your search to **analyze what** it is **you're looking for**
- **Use nouns** in your queries – the who/what, when, where, how and why; avoid conjunctions, verbs, adverbs and adjectives
- Use **keywords at** the **right "level"** of specificity: precise, but not overly restrictive
- **Use phrases** where natural; they are your most powerful weapon
- **Use** structured ("Boolean") syntax, especially the '**AND**' **operator**
- Constrain your search by using two or three related, but narrowing, concepts in your query
- BUT, generally, keep overall query length limited to six to eight keywords maximum
- **Use advanced** search **options** and specialty features when appropriate
- For **difficult searches**, **use** only search engines that support **Boolean syntax**, or tools or metasearchers that do
- For specific topic searches, consider search engines tailored to those topics

[7] "Recommended Search Strategy: Search With Peripheral Vision," issued by the Teaching Library Internet Workshops from UC Berkeley, found at: http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Strategies.html.
[8] See http://www.rice.edu/Fondren/Netguides/strategies.html.
[9] See http://www.infotoday.com/searcher/default.htm.

- Save time by **learn**ing **your search engines** and advanced, 'power searching' techniques.

Useful tips for how to govern the accuracy and scope of your searches are:

Search Action	Search Scope	Results Likelihood	Comments
Focused Keywords	narrows	higher	yes; but can be too focused
Broad Keywords	broadens	lower	low yield
Use of Synonyms	broadens	higher	watch for long query sizes
Additional Keywords	broadens	higher	good, if related well
More Query 'Concepts'	narrows	higher	should not exceed 3 to 4
Fewer Query 'Concepts'	broadens	lower	single concept or keyword MAJOR search mistake
Use of Phrases	narrows	higher	exact word order critical
Use of Wildcards	broadens	higher	recommend; watch short stems
Multiple Queries	broadens	higher	useful when search uncertain
Simple Text Search	broadens	lower	quick; same as all **OR** operators
Structured (Boolean) Search	narrows	higher	takes time to master
AND Operator	narrows	higher	highly recommended
OR Operator	broadens	lower	only for synonyms; be careful when using with **AND**
NEAR Operator	narrows	higher	excellent alternative to phrases
AND NOT Operator	narrows	higher	useful in limited circumstances
Use of Parentheses	depends	depends	great when done well; tricky to do; keep simple
Redundant Keywords	broadens	lower	use care and remove
Alternate Spellings	broadens	higher	not common; be aware
Filters	narrows	depends	can be useful or too narrow

Other Useful Sources

For general issues relating to search engines, their capabilities, market share and how they work, two excellent resources are Danny Sullivan's SearchEngineWatch[10] and Greg Notess' Search Engine Showdown.[11] You may also enjoy checking out Steve Steinberg's fascinating historical article for Wired on the nature of search services and the general topic of why knowledge organization matters.[12]

Notes on Recent Search Engine Changes

There have been major shifts in the major Internet search engine landscape since the last complete update of this tutorial five years ago. The first shockwave was the emergence and then dominance of the Google search engine. This event, plus general consolidation, have led to further major changes in the past year.

[10] See http://www.searchenginewatch.com.
[11] See http://notess.com/search/.
[12] The Steinberg article may be found at: http://www.wired.com/wired/archive/4.05/indexweb.html..

The company Overture acquired the Alta Vista and AlltheWeb search services in 2003. Yahoo! then acquired Overture after an earlier 2003 acquisition of the Inktomi search service. This consolidation put three of the major prior search services under Yahoo! ownership.

In parallel with these acquisitions, there were also major changes as to the search providers used by other major search sites. Microsoft, for example, is now releasing its own search service after having licensed both Google and Inktomi options. The number of unique indexes is dropping, even though branded sites are still maintained by Alta Vista, AlltheWeb, and Hotbot (all are now based on Yahoo! companies).

There have been two significant benefits, and two significant losses, due to this consolidation:

- **Benefit** – major search index sizes have grown substantially
- **Benefit** – some unique search power has been introduced through options such as Yahoo!'s shortcuts (similar to earlier metadata search options) and Google and Yahoo! localization of search results
- **Loss** – wildcard, or "stemmed" or truncated, searching is now largely absent from major search services, and
- **Loss** – proximity searches (excepting for phrase searches, which all major services maintain) such as **NEAR**, **BEFORE** and the **AFTER** operators are now non-existent.[13]

With the massive increases in scale, results counts from searches have also become approximate. A few years ago specific counts could be obtained from the Alta Vista and Inktomi services; they are now counting approximations.

Beginning the Tutorial

All of us need information. But few of us have studied information or library science, and not everyone has used search services or Internet search engines sufficiently to learn all of the nuances. This tutorial is for those who are learning the ropes about 'power searching.' But, even if you're quite experienced in these areas, you might find some benefit from glancing through these topics.

Simple to follow examples are presented in each topic. We've written it to be a one-stop reference. Don't feel you need to work through all of the topics in one sitting. But, if you do take the time to work through this material, we guarantee you'll reap big dividends in faster and more accurate results. And, you will be on your way to earning the title of an Internet "Power Searcher."

[13] Some of these Boolean features made Alta Vista a favorite of power searchers. For a good overview of these broader changes, see G.R. Notess, "The New Yahoo! Search," *Online* magazine, Vol 28 No. 4, July/August 2004. See http://www.infotoday.com/online/jul04/OnTheNet.shtml.

PART 1: INTERNET SEARCH BASICS

Much is discussed on the Internet regarding its growth and user-driven, decentralized nature. This part overviews the current state of searching and search services on the Internet. The essential arguments are that your time is well spent learning how to issue more effective queries and to understand the basic operations of the search services you employ.

Topic 1: Searcher Frustrations

Many have likened the Internet to a huge, global library. While true in some aspects, it has some unique differences. There is no central "card catalog"; the Internet's growth is outpacing the ability of humans or technology to keep up with it; its sheer size is unknown and perhaps unknowable; and content is (to say the least) of uneven quality. Here's some of what we know (or think we know) about information on the Internet:

- Document growth is, at minimum, doubling each year[14]
- Two-thirds to three-quarters of all users cite finding information as one of their primary uses of the Internet
- Two-thirds to three-quarters of all users cite the inability to find the information they seek as one of their primary frustrations (second only in frustration to slowness of response)
- Use of structured, or 'Boolean' queries, while known to help obtain better search results, can be difficult and frustrating for some users to learn.

One of the challenges of the Internet is to make its value available to the millions of new users who have had no formal training or experience in query formulation or search strategies.

Topic 2: Search Engine and Directory Basics

Search services on the Internet come in two main flavors: 1) 'search engines' that index words or terms in Internet documents, including searchable 'deep' Web databases; and 2) 'directories' that classify Web documents or locations into an arbitrary subject classification scheme or taxonomy. Most of the above are examples of the former; Yahoo, About.com and LookSmart are examples of the latter.

General Internet search engines use 'spiders' or 'robots' to go out and retrieve individual Web pages or documents, either because they've found them themselves, or because the Web site has asked to be listed. Search engines tend to "index" (record by word) all of the terms on a given Web document. Or they may index all of the terms within the first few sentences, the Web site title, or the

[14] US Department of Commerce, "The Emerging Digital Economy," April 15, 1998.

document's metatags. Due to the ever-changing nature of the Internet, the services must re-sample their sites on a periodic basis. Some of these services re-sample their sites on a weekly or less-frequent basis.

Precision, *recall* and *coverage* are limiting factors for most search engines. Precision measures how well the retrieved documents match the query; recall measures what fraction of relevant documents are retrieved.[15] Coverage refers to what percentage of the potential universe of relevant documents is cataloged by the engine. For example, consider a search engine with 10 documents, five of which mention eagles, out of a total universe of 50 potential documents mentioning eagle (45 of which are not indexed by that engine). A query on eagle that returned four documents and two others from this engine would have a precision of 0.66, a recall of 0.80 and coverage of 0.10.

Precision is a problem because of the high incidence of false positives. (That is why you get so many seemingly irrelevant documents in your searches.) This is due to imprecision in the query (searching on eagle and missing the mention of eagles), indexing mistakes by the engine, and keywords entered by the Web document developer that do not actually appear in the document. Coverage is a problem for all engines, particularly with the dominant presence of the deep Web.

Search directories operate on a different principle. They require people to view the individual Web site and determine its placement into a subject classification scheme or taxonomy. Once done, certain keywords associated with those sites can be used for searching the directory's data banks to find Web sites of interest.

For searches that are easily classified, such as vendors of sunglasses, the search directories tend to provide the most consistent and well-clustered results. This advantage is generally limited solely to those classification areas already used in the taxonomy by that service. Yahoo, for example, has about 2,000 classifications (excluding what it calls 'Regional' ones, which are a duplication of the major classification areas by geographic region) in its current taxonomy. The Open Directory Project has nearly 400,000 categories.[16]

When a given classification level reaches 40 to more listings, the staff editors split the category into one or more subcategories. If a given topic area has not been specifically classified by the search directories, finding related information on that topic is made more difficult. Another disadvantage of directories is their lack of coverage because of the cost and time in individually assigning sites to categories.

[15] G. Salton, *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*, Addison Wesley, 1989.
[16] Data as of December 12, 2004; see http://www.dmoz.org/

Most searches of a research or cross-cutting nature tend to be better served by the search engines. That is because there is no classification structure behind the listings; only whether the keywords requested appear in that search engine's index database or not.

The flexibility of indexing every word gives users complete search control. But the growth of standard Internet search engines is now creating a different kind of problem: too many results. In the worst cases, submitting broad query terms to such engines can result in literally millions of potential documents identified. Since the user is limited to viewing potential sites one-by-one, clearly too many results can often be a greater problem than too few.[17]

Increasingly, the growth of the Internet is causing the specialization or balkanization of search services, and is a key driver in the faster growth of the deep Web. Lawyers, astronomers or investors, for examples, may want information specifically focused on their interest topics. By cataloging information in only those areas, users interested in those topics are better able to keep their search results bounded. Such specialization can also lead to more targeted advertising on those search service sites. Again, though, like the directories, such specialization can limit search results to the boundaries chosen by the service, which may or may not conform to the boundaries sought by the user.

The ultimate challenges to any of these centralized search services, therefore, are to: 1) keep pace with explosive document growth; 2) understand the "boundary" needs of their user communities; 3) provide sufficient "intelligence" to infer what users are really asking for even when their queries don't specify it; and 4) ensure sufficient coverage to provide one-stop searching.

Topic 3: How Search Engines Rank Documents

A Web page, or document, can contain various kinds of content (as opposed to display or presentation options like sound, animation or frames), some of which is not shown when you view the document in your browser:[18]

- **Title** – an embedded description provided by the document designer; viewable in the titlebar (it is also used as the description of a newly created bookmark by most browsers)
- **Description** – a type of metatag which provides a short, summary description provided by the document designer; not viewable on the actual page; this is

[17] However, even though most search engines indicate huge numbers of results, such as '1-10 or 876,000 results,' in actuality the depth to which these results can be viewed is limited. Google, for example, limits results viewing to approximately the top 800 or so results.

[18] If you'd like to see these hidden tags on a given Web document, save the document while viewing in your browser using the 'Save As' option on your browser's 'File' menu option. Then, view that document with a text editor or Wordpad. You will see these hidden fields shown in HTML brackets (e.g., <Description= …>).

frequently the description of the document shown on the documents listings by the search engines that use metatags

- **Keywords** – another type of metatag consisting of a listing of keywords that the document designer wants search engines to use to identify the document. These too, are not viewable on the actual page
- **Body** – the actual, viewable content of the document.

Search engines may index all or some of these content fields when storing a document on their databases. (Over time, engines have tended to index fewer words and fields.) Then, using proprietary algorithms that differ substantially from engine to engine, when a search query is evaluated by that engine its listing of document results is presented in order of 'relevance.' Because of these differences in degree of indexing and algorithms used, the same document listed on different search engines can appear at a much higher or lower ranking (order of presentation) than on other engines.

Google significantly altered the Internet search engine landscape when it first went commercial in 1998. Besides evaluation of the readable text in the document, it added the scoring concept of "Page Rank." Page rank is a measure of how many other sites link to the target, a measure of the target's popularity. For standard searches finding common information, the page rank addition has proven to be a very successful ranking adjustment (from the standpoint of content value, however, that addition may be of less value).

Additionally, some search engines began to alter their results presentations with "paid listings," wherein advertisers or through a payment of a fee by the target site, higher rankings are presented to users. This approach, while perhaps of financial benefit to the Internet search engine, have been controversial and not as well received by end users.

Though not hard and fast, and highly variable from engine to engine, these factors thus tend to influence greatly the ranking of a document for a given query and a given search engine:

1. **Order a keyword term appears** – keyword terms that appear sooner in the document's listing or index tend to be ranked higher
2. **Frequency of keyword term** – keywords that appear multiple times in a document's index tend to be ranked higher
3. **Occurrence of keyword in the title** – keywords that appear in the document's title, or perhaps metatag description or keyword description fields, can be given higher weight than terms only in the document body
4. **Rare, or less frequent, keywords** – rare or unusual keywords that do not appear as frequently in the engine's index database are often ranked more highly than common terms or keywords

5. **Page rank** – the popularity of the target document
6. **Paid listing** – an "artificial" boost to document rankings.

The specifics of these algorithms are very important because most users tend to focus only on the first results page of candidate documents presented by a search engine, and then only to the highest listed results.

Some engines attempt to "infer" what you mean in a query based on its context. Thus, the meaning of **heart** can differ if the context of your search is cardiac disease as opposed to Valentine's Day. The methods by which these inferences are made are statistically based on the occurrence of some words in conjunction with others. Though useful for simpler queries, such inference techniques tend to break down when the subject of the query or its modifiers do not fit expected query relationships. For commonly-searched topics, this is generally not a problem; for difficult queries, it is a disadvantage to standard full-text indexing.

Cottage industries have emerged to help Web site developers place themselves higher in the search engines' listings (it is clearly more valuable to be within the first few listings sent to a user than be buried hundreds, or thousands, of documents lower). A constant battle is being waged between the engines and those desiring high listings from jimmying the system to "unfair" advantage.

Crude, early attempts to "spam" search engines to get higher listings included adding hidden terms like "sex" that were searched frequently but not the real subject of the document. Other techniques were to use certain keywords repeatedly, such as "cars cars cars cars cars" to get a higher frequency rating. Another was to cram the page with high-interest terms using the same color as the overall Web page, thus "hiding" the added keywords. The leading search engines have caught on to these and now have automated ways to prevent the worst of these spamming techniques.

More subtle techniques, however, are hard to prevent. For example, a listing for ski resorts in Utah could also add hidden tags for "Caribbean" or "beach resort" knowing that wealthy Caribbean travelers may also be looking to take ski vacations. If you as the searcher asked for Caribbean vacations you may logically wonder why you've gotten a listing for Utah ski resorts. It is because of such techniques (among others) that you can sometimes get document listings from a search that seemingly have nothing to do with your query.

So, differences in how search services rank documents, how developer's themselves choose to characterize their Web documents, and just simple errors in how computers process and index these pages can all lead to highly variable ranking results from different search services.

Topic 4: Characteristics of Searchers and What Takes Search Time

Professional information searchers do not have a single style. There is no "correct" way to search on the Internet. Search styles have been described as ranging from 'ants' – the carefully planned, methodical search hoping to get exact results on the first try – to 'grasshoppers' – intuitively jumping from topic to topic, refining results as more is learned.[19] Only you can determine what your style is.

There is only one meaningful measure for a successful search: getting the results you desire. And within that context, there is only one meaningful basis for judging whether one search strategy or another is superior: whether those results are obtained faster.

Surfing and browsing on the Internet are seductive. One begins with an objective in mind, finds new tidbits of interest, and hours later can wonder where the time has gone. It is often difficult to apply metrics against whether the original search interest was obtained, or whether the whole process was productive or not. So, let's look at some aspects of a typical search. The example assumes a 56.6 KB modem and a relative "fast" time for the Internet.[20] This is perhaps an optimistic mid-range and illustrative example for current users of the Internet:

Search Step	Est. Process Time (sec.)	No. Repeats	Total Time (min.)	Cumulative Time (min.)
Formulate Query	120	3	6.0	6.0
Issue Search	10	3	0.5	6.5
Get Search Listings from Service (30/query)	10	9	1.5	8.0
Review Documents; Select for Download	12	50	10.0	18.0
Download Document	15	50	12.5	30.5
Review Document	18	50	15.0	45.5
Average Time per Document (90 document example)				0.5

These estimates are likely an underestimate. Information professionals using the Web to do searches in comparison with traditional online search services like Dialog found it took on average 2.4 minutes per document to get acceptable results.[21]

[19] These terms were coined by Barbara Quint; see further: http://www.hut.fi/~ipaavola/FLI/Courses/Database/quest01.html.
[20] Unpublished BrightPlanet Corporation testing data.
[21] This article presents results of searching on conventional search databases, such as Dialog, versus the Internet: http://www.infotoday.com/searcher/feb98/story1.htm..

Whatever the actual "average" search time is, it will not apply to your circumstances in any case. However, what ***is*** the case is that certain aspects of searching can add delays to getting desired results and increase frustration:

- No matter how precise or accurate the query, a large percentage of results returned by search services will not be what you're looking for
- Actual search time in getting candidate listings from services is relatively fast; the one-by-one document download and review is the most time consuming part of the process
- Larger listings of candidate documents from the services require more evaluation time
- Often too little time is spent on search and query formulation; any improvements you can make toward more precise and accurate queries will lead to fewer documents to review and faster overall times to the results you want.

The essential conclusion is that time is well-spent in understanding how to pose a proper query and how to take advantage of the way that search services work. These topics are the focus of the rest of this tutorial.

PART 2: KEYWORDS – THE ESSENCE OF THE SEARCH

Despite all the gobbledygook about things like 'Boolean' and query operators, the most difficult – and fundamental – aspect of a search are the keywords used in your query.

A search is inherently looking for information about a ***topic***. This part describes how you can proceed from search concepts to identifying the specific keywords – or terms – that will give you the results you're seeking. We begin by presenting an information problem which will be the basis for progressing through the tutorial's remaining topics.[22],[23]

Topic 5: Sample Information Problem for this Tutorial

Jan is an office worker in downtown Minneapolis. While on lunch break one fine Spring day, Jan's eye is caught by a flash in the sky above. Jan sees a bird about the size of a crow diving at high speed and catching in mid-air what appears to be a pigeon. The bird then swoops out of sight. Jan is captivated by the mostly gray and white bird, with the crooked black and yellow beak. Jan has never seen this bird before, and wonders what it is doing in the city. That night, Jan decides to find out more on the Internet about this mystery bird.

Where does Jan begin?

Topic 6: Query Concepts: What, Where, When, How, Why

Mastering the concepts behind a search is not as complicated as may seem at first. The first few searches are perhaps difficult, but, once done, the nuggets behind your information request start becoming clear. Like riding a bike for the first time, it does take some practice.

One of the bigger mistakes you can make in preparing a query is not providing enough keywords. On average, most users submit 1.5 keywords per query [24]. This number is insufficient to accurately find the information you are seeking. Thus, a central task in query formulation is for you to identify a sufficient number of appropriate keywords.

[22] Results counts used in this tutorial were conducted in early December 2004 and are based on the Yahoo! Advanced Search option (http://search.yahoo.com/web/advanced?ei=UTF-8). In the original version of this tutorial first published in April 1998, the AltaVista search engine was used. However, AltaVista has not kept pace with broad indexing of the Web and in early 2004 many of that engines advanced syntax features were removed from the service. Indeed, there are been a general reduction in advanced Boolean search features for all search engines over the past few years. See further the notes on search engine changes in the introduction for more background.

[23] Search engines only now provide approximateions – not precise – counts of results. Also, since this tutorial has been in the Internet domain for some time, some counts are slightly inflated due to the references to the sample queries used herein.

[24] B. Pinkerton, "Finding What People Want: Experiences with WebCrawler." See http://info.webcrawler.com/bp/WWW94.html.

If you are new to searching, the first task we recommend when formulating a search is writing down what information you are seeking. This is best done – go ahead, use some paper and a pen – in the form of some questions. Before doing a search, it is important to bound your topic as completely yet succinctly as possible. After experience is gained, you can skip writing things down and plunge right into it.

Formulating a query is akin to solving a mystery. Some pieces of information are available, but if sufficient information were available the answer would be known and there would be no need to seek more. This is the essence of a query: missing information. It is up to you, the searcher, to define your snare – the query (quarry? pun intended) – sufficiently to trap that missing information and solve the mystery.

As any good detective would, it is useful to begin by listing what you do know according to these standard categories. Jan lists these for the mystery bird:

TIP:
*Always keep in mind the **who**, **what**, **where**, **when**, **how** and why in formulating your query.*

- **WHO / WHAT?** – gray and white bird, about the size of a crow; yellow and black beak
- **WHERE?** – downtown office buildings in the City of Minneapolis
- **WHEN?** – daylight in the Spring
- **HOW?** – fast flyer, hunting pigeons (?) as prey
- **WHY?** – hunting bird; why never seen before? blown off course? is it migrating?

Of course, not all of these five categories will apply to a given query, and the specifics will obviously vary for your desired topic. But it is useful to keep these five categories in mind – the what, where, when, how and why – when analyzing the major components.

Topic 7: Breaking Down Your Query

Let's take the five responses to the query tests in **Topic 5** apart (yours will differ substantially, but the same ideas apply). First, there are many common words in these responses that are prepositions, conjunctions or common verbs. These include: **and, about, the, of, a, in, as, if, not, why, never, before, is** and **it**. These common words are referred to as "stoplist" words: they are essential to the connecting tissue in language, but they are filler in any search request. **All** search engines ignore them because they have minimal information value and are found commonly in all language. Search services may include on the order of 600 of these common words in their "stoplists"; if you use them in a query they are ignored. Therefore, you should ignore them as well.

Okay, removing such words from our responses leaves these remaining words:

gray	**downtown**	**flyer**
white	**office**	**hunting**
bird	**buildings**	**pigeons**
size	**city**	**blown**
crow	**Minneapolis**	**off**
yellow	**daylight**	**course**
black	**Spring**	**migrating**
beak	**fast**	

TIP:
Never use articles, pronouns, conjunctions or prepositions – the connecting tissue in language – in your queries.

Now, let's further classify these terms into three categories, similar to diagramming a sentence (but made simpler for our purposes). Let's use the classifications of objects/nouns, actions/verbs and modifiers/qualifiers (adjectives, adverbs and predicate subjects). And, let's now re-list these words by these categories:

Objects	Actions	Modifiers
bird	blown	gray
buildings	migrating	white
city	not seen	size
Spring		crow
daylight		yellow
		black
		beak
		downtown
		office
		Minneapolis
		fast
		flyer
		hunting
		pigeons
		off
		course

Not all of these categories are equally useful in a query.

Topic 8: Focus on Nouns and Objects

Almost without exception, the central keywords in your queries will be nouns. Though sometimes adverbs and adjectives can help refine your search, the key pivot point is a noun, or series of nouns. Why is this?

The most precise terms we have in language are for tangible, concrete "things" or objects. Actions and modifiers are very diverse, easily substitutable, and generally not universally applied in any given description. For, example, take the concept of "fast". A thesaurus will give 75 or more different words for fast. Here are some counts from Yahoo! for numbers of Web documents containing these terms:

fast	130,000,000
speed	89,900,000
quick	122,000,000
rapid	23,400,000
fleet	10,800,000
swift	8,530,000
breakneck	384,000

TIP:
The keywords in your queries will most often be nouns – and then likely no more than 6 or 8 of them.

Or, alternatively, take a modifying concept like 'color'. Again, here are the Yahoo! document counts:

color	99,300,000
red	139,000,000
yellow	65,800,000
blue	111,000,000
gray	23,000,000
grey	17,500,000
slate	4,930,000
white	155,000,000

Note three aspects about these lists. First, some modifiers are also act as nouns like truck 'fleet', the bird 'swift', pool table 'slate' or Justice 'White'. Second, a concept like speed or color can be described in lots of ways (most of which are not shown). Third, you generally don't know how others would describe the same thing. In our example of Jan's mystery hunting bird, would someone else describe it as "fast", "quick" or "like a bolt from the sky"? Would someone else describe the bird as "gray", "grey", "slate-gray" or "smoky"?

The same kind of ambiguity and substitutability applies to actions or verbs. Does the bird "fly", "soar", "swoop" or "glide", or any of the other dozens of ways the act of flying can be described?

As a general rule, try to avoid using action terms and mostly try to avoid using modifiers in your queries. Where exceptions to these guidelines may make sense is when a modifier helps to precisely define your object, such as in "Limburger cheese."

We've thus gone through a process that has led us to these possible objects as the focal points for constructing our query terms:

bird
buildings
city
Spring

daylight

The obvious main subject is **bird**. The next few topics will concentrate on it; we'll return to the other objects as we later refine our final query.

Topic 9: Word Root Variants

One of the first mistakes in query formulation is not using word root variants sufficiently. Let's look at this question in regards to our subject, **bird**. Accounting for singular and plural cases of an object is easy to overlook; but, if done, can act to unduly restrict the universe of documents in which you will be conducting your search. Using Yahoo! again, here are the document counts for

24,500,000

18,800,000

the single and plural versions of **bird**:

By using either only **bird** or **birds** as our subject, we would eliminate half or so of the potential documents that we'd like to use as our search basis. We should therefore use both **bird** and **birds** as query variants. [25,26]

By using the query **bird OR birds** we are able to get complete coverage, upping the result count on Yahoo! to 37,500,000.

---

[25] General Internet search engines have unfortunately moved away from "stemming" or "truncating" words using the wildcard asterisk (*), with MSN Search still supporting it on its advanced search form. With the current state of search engines, it is now necessary to use plural and other variants explicitly.

[26] If you DO have a search site that supports wildcards through truncation, that is by the better alternative to specifically entered query variants. Truncation is applying a wildcard character after the first few letters in a term (the "stem"). The asterisk (*) is the almost universally accepted truncation wildcard. Generally, you must also have a minimum of three characters at the beginning of the word as your stem basis. Once marked for truncation, then any matching characters after that will be picked up in the search query. Using the asterisk wildcard will generally be ignored or you'll get a query format error if the search engine doesn't support it.

Remember, ANY words with characters after the stem will be matched to your query term if the search engine supports truncation. If you do use ttruncation, you need to be aware of unintended consequences. In the case of the stem bird* there are relatively few unwanted words (birdbrain) picked up in the search. But let's look at another of the objects, city, in our mystery bird sample problem.

To stem and pick up the plural form of city, cities, we would need to specify cit*. But look at some of the words this stem specification would match:

citadel	citations	cited	citizen	citizenship	citriculture	citronella
citadels	cite	cities	citizenry	citrate	citrine	citrus
citation	cites	citify	citizens	citric	citrone	city

The cit* stem clearly picks up way too many unwanted words.

Stemming tends to work best when the actual stem is longer, when plurals are represented by an added '-s' (as opposed to '-ies' or other forms), and the stem itself is not a root to many other common words. With just a little thought, however, truncation is easy and can pay useful dividends in properly scoping your query with a minimum of keywords. We highly recommend its use where the search form supports it.

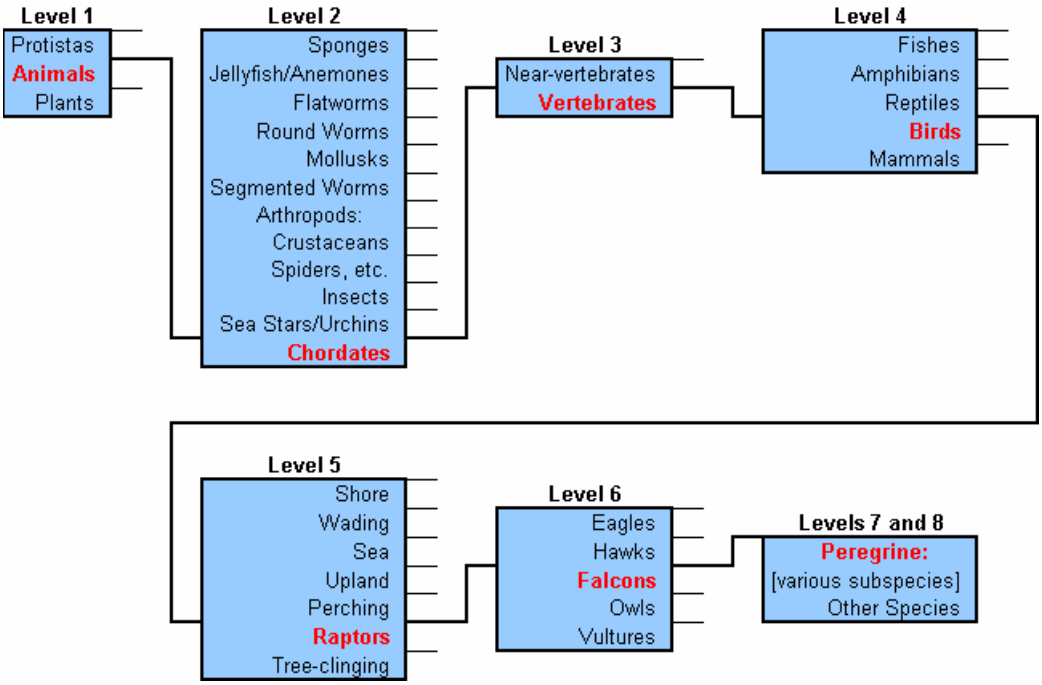
Topic 10: Finding the Right Level

Perhaps you've already noticed, but our query subject **bird*** is contained on more than 37 million documents (in Yahoo! alone). It would be a little difficult to review all of those documents at one sitting.

THE MOST CRITICAL PROBLEM IN ALL QUERIES IS FINDING THE RIGHT LEVEL OF SPECIFICITY FOR THE SUBJECT QUERY TERM(S). Too broad a keyword specification, and too many results are returned; too narrow a specification, and too few are returned.

All information is classifiable and amenable to structure. We are all familiar with dictionaries, which classify words alphabetically. However, an alphabetical structure is not of much use to query formulation. But there are many other classification schemes used for information which CAN help find the right level, or specificity, for your keywords. A few examples appropriate to our mystery bird search are presented in this topic.

Our first example classification presents the structure of the animal kingdom [27]:

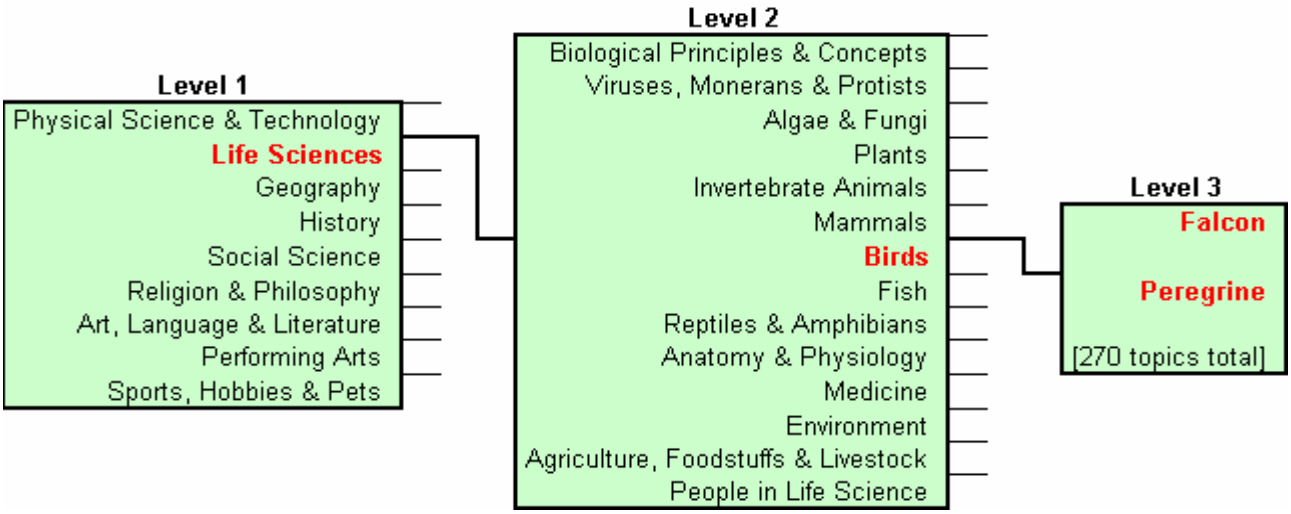


'Level' Example Using the Kingdom of Life

As we will see, our initial keyword term of **bird*** is at least three levels off of where it should be. Using **bird*** as is would lead to massive results sets from the

[27] See http://niko.unl.edu/bs101/notes/lecture19.html.

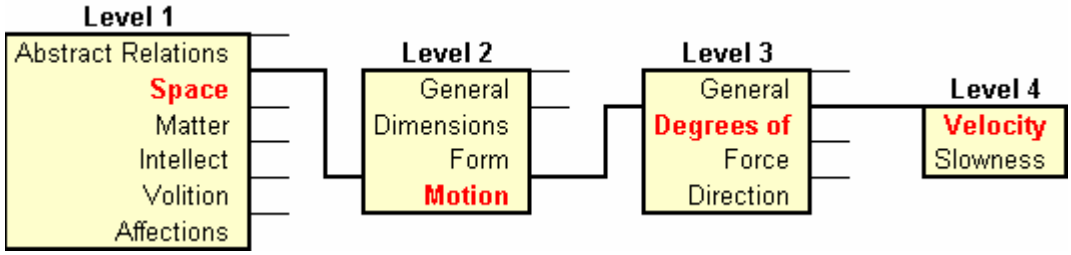
search engines and virtually no likelihood that we will find the information we're looking for.



'Level' Example Using Encarta

Another way to classify information is shown by the encyclopedia, (the above example is from Microsoft's Encarta [28] – the actual encyclopedia doesn't matter; we're only illustrating a point).

As a very different example, the chart below shows how the word "fast" is placed within the structure of a thesaurus [29]:



'Level' Example Using Thesaurus

As noted, search directories also apply a classification structure for how they organize and present Web sites.

Like the first animal phylum example above, **bird*** in the example is about three or four levels off from where our subject keyword should be.

[28] Microsoft Encarta Encyclopedia.
[29] *The Original Roget's Thesaurus*, St. Martin's Press, 1962.

Finding the right level may involve your personal knowledge and experience, doing a preliminary search or consulting other references. In the case of Jan and the mystery bird, looking in a bird book was sufficient to match pictures with the bird seen as a **peregrine falcon**.

The time spent in finding how to characterize your subject at the proper level is definitely well spent, as these document counts from Yahoo! illustrate:[30]

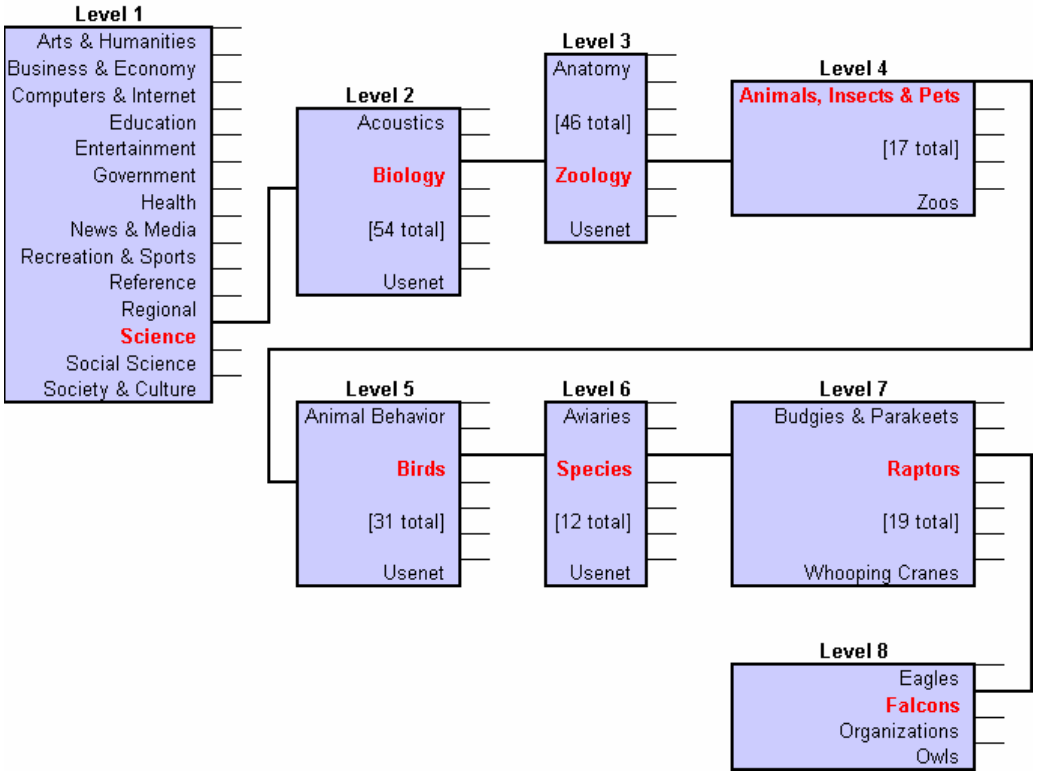
bird*	37,500,000
falcon*	7,220,000
peregrine falcon*	210,000

By identifying our mystery bird as a **peregrine falcon**, we've narrowed the search by 99%! Remember, at 30 seconds to 2.5 minutes per document reviewed, the effort spent in zeroing in on the bird of interest has saved us tremendous overall search time.

The critical point about finding the right "level" in your keywords is that words at levels higher than where you should be return way too many results; those at levels lower than where you should be return too few or no results. This phenomenon is due to the fact that "things" at lower levels tend to "rollup" and sum into "things" at higher levels.

Philosophers, epistemologists, taxonomists, linguists and others can argue for centuries about "proper" ways to classify information. That is not our concern. Rather, the point is that keyword objects can be placed into a structure at various levels. Always keeping forefront whether your query subject is at the right level or not in those structures can bring big benefits in faster, and more accurate, searches.

[30] Note, since our Yahoo! test site does not support wildcarded truncation, our asterisk notation should be understood to be a variant query. That is, **bird*** is issued as **bird OR birds**.



'Level' Example Using Yahoo

Topic 11: Synonyms

Let's assume, however, that Jan was not able to match the bird book pictures with the mystery bird to identify it as a peregrine falcon. How can we use these query concepts to better hone in on what type of bird it is?

One useful place to begin is with synonyms. Jan knows the mystery bird is a hunting bird. Jan lists other synonyms that come to mind for **hunting bird**. We provide Yahoo! document counts for these synonyms:

TIP:
You can use synonyms both to find the right "level" for your query subject and to ensure proper coverage.

hunting bird*	63,600
bird* of prey	767,000

Jan, however, suspects neither of these terms is the "correct" synonym. Attacking this problem from another angle, Jan writes down specific kinds of birds of prey:

hawk
eagle
owl

Using these three keywords, Jan's search immediately turns up a number of sites referring to **raptors**, the technical term for hunting birds. Jan finds a great site on

raptors that also has pictures that positively identifies the mystery bird as a peregrine falcon.[31] Jan also learns that vultures are raptors, too.

The best synonyms provide relatively complete coverage for the subject at hand and are "pitched" for the right informational objective. In Jan's case, it was needing to identify a specific bird, and a more technical term like "raptor" fit the bill. Were Jan's interest more oriented to references in novels, perhaps "hunting bird" or "bird of prey" would have been more appropriate.

An illustration of a good synonym with proper coverage is:

Good Synonyms Provide Good Coverage

Good coverage is not always possible. Where not possible, provide a couple of alternate terms (that is, synonyms). But, remember, always play the numbers game. Your query terms are limited so choose them carefully.

Having determined the mystery bird to be a **peregrine falcon**, Jan considers whether synonyms for this term are also worthwhile. Based on what Jan has learned, these are the possible synonyms and document counts from Yahoo!:

peregrine falcon*	210,000
Falco peregrinus	40,900
duck hawk*	1,260
all three combined (single query)	231,000

Again, note the three synonym counts do not exactly sum due to duplicate query occurrence in individual documents.. This example is a good instance where multiple synonyms do not buy enough increased coverage to be warranted. **peregrine falcon** is the most used description of this bird; adding the other terms increases coverage by only a minor percentage.

You need not get actual document counts from search engines in order to weigh such choices in your own queries. Simply use good judgment of what you're gaining – if anything – by adding more synonyms to your query subjects. Common sense should be a sufficient guide.

TIP:
Always look for natural phrases in your query concepts – they are one of the most powerful

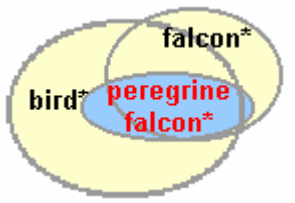
[31] University of Minnesota Raptor Center, http://www.raptor.cvm.umn.edu/.

A thesaurus, a dictionary, personal knowledge or a preliminary Internet search can all be worthwhile places to find synonyms for the major subject(s) in your query. Generally, you should not waste the time thinking about synonyms for other terms in your queries, unless you know them to have very poor coverage.

Topic 12: Use of Phrases

Your most powerful keyword term is the phrase. Phrases are combinations of words that must be found in the search documents in the EXACT order as shown. You denote phrases within closed quotes (**"peregrine falcon*"**). Some search services provide specific options for phrases, some do not allow them at all, but almost all will allow you to enter a phrase in quotes, ignoring the quotations if not supported.

Why phrases are powerful is illustrated below:



Phrases Target Results

Again, using Yahoo! document counts, the ability of phrases to zero in on desired results is clear:

bird*	37,500,000
falcon*	7,220,000
peregrine falcon*	210,000

Phrases should be used where the constituent terms are naturally married. Other examples would be "rain in Spain", "Gettysburg Address", "solar system" or "big bad wolf". Where two or more words are necessary to capture the subject, but may not always be next to one another in the same order, the **AND** or **NEAR** Boolean operators should be used.[32]

In addition to **"peregrine falcon*"**, Jan also uses **"endangered species"** to help focus the search. Jan chose **"endangered species"** because information gained in identifying the mystery bird indicated that peregrine falcons were at risk of

[32] When using phrases, it is important to consider nuances of the phrase that wouldn't normally be of concern. For example, the spaces between words are as important as any other character. If you include a double space between any two words in the query and the phrase typically has only one, the search will fail. Also, sometimes two dashes are used together on Web documents to approximate an en- or em- dash. If you include only one dash, the search engine may miss all those documents that use two. There is variability in the way certain search engines treat spaces, dashes, and the like. If you suspect there may a problem, consider submitting your phrases in different ways to capture these variations.

extinction in the 1970s due to DDT effects. Jan suspects that the answer to the **why** question of the search is the rarity of the bird and not migration or being blown off course. **"endangered species"** is a logical construct for a phrase because the terms are almost always used together to discuss organisms at risk of extinction.

PART 3: BOOLEAN BASICS

Despite its intimidating name, Boolean search techniques are really quite simple to learn and can add tremendous effectiveness to your searching. While working through this part, most of you will recognize constructs that were taught to you in high school math.

"Boolean" searching draws its name from George Boole, a mathematician and logician from the 19th century. He developed Boolean algebra, which is the basis for this form of structured search technique. Boolean algebra is also of prime importance to the design of modern computers.

Most information on the Web is highly unstructured. Boolean search techniques were first applied by information professionals to traditional search services like Dialog or Lexis-Nexis. Boolean techniques, while not supported by all Internet search services, provide a way for you to bring structure to this unstructured environment.

Without Boolean techniques, you are stuck with doing a lot of free-text searching; meaning, looking for documents that contain words you think will be in the document you are seeking. Sheer document volume makes free-text searching difficult and prone to failure. Boolean techniques give you the power to narrow your search to a reasonable number of potentially useful documents thereby increasing your likelihood of success.

Topic 13: Boolean Overview

Boolean logic is used to construct search statements using logical **operators** and specified **syntax**. These are combined into **Boolean expressions**, which always are either true or false when evaluated.

The shopping list of operators and syntax available to Boolean searching (though not supported by all search services) is:

- **AND** – terms on both sides of this operator must be present somewhere in the document in order to be scored as a result
- **OR** – terms on EITHER side of this operator are sufficient to be scored as a result
- **AND NOT** – documents containing the NEXT term to the right of this operator are rejected from the results set
- **NEAR** – similar to **AND**, only both terms have to be within a specified word distance from one another in order to be scored as a result; this is an 'adjacency' operator

- **BEFORE** – similar to **NEAR**, only the first (left-hand) term before this operator has to occur within a specified word distance before the term on the right side of this operator in order for the source document to be scored as a result; this is an 'adjacency' operator
- **AFTER** – similar to **NEAR**, only the first (left-hand) term before this operator has to occur within a specified word distance after the term on the right side of this operator in order for the source document to be scored as a result; this is an 'adjacency' operator
- Phrases – combined words or terms that must appear directly adjacent to one another and in the phrase order for the source document to be scored as a result
- Wildcards (stemming) – beginning characters that must match the same beginning characters in a document's words in order for it to be scored
- Parentheses – nested operators that are evaluated in an inside-out, then left-to-right order of precedence (though some search engines violate this rule).

Example uses of these operators are based on the sample tutorial problem of finding information on the peregrine falcon discussed in the previous topics.

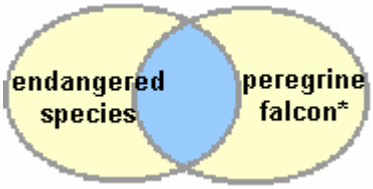
The underlying premise of Boolean logic is set theory. The **AND** operator is equivalent to the set **intersection** operation; the **OR** operator is equivalent to the **union** set operation. To help explain these concepts, specific topics below use so-called Venn diagrams. Don't worry about the fancy name. The diagrams are color-coded to indicate the result of an operation. The universe of possible results is shown in yellow on these diagrams; the accepted results in blue.

One way to decide when to use the **AND** or **OR** operators is to test whether your keywords are different concepts, or a just different ways (synonyms) to say the same thing. For different concepts, use **AND**; for synonyms, use **OR**.

Boolean search syntax needs to follow a precise structure. Queries constructed using Boolean syntax do not look like real sentences. The **AND** and **OR** Boolean operators, in particular, sometimes seem to mean the opposite of what they do in natural language. Searching based on simple sentences and phrases is a different construct known as **natural text searching**.

Topic 14: AND Operator

AND means "I want *only* documents that contain ***both*** words." **AND** logic focuses, coordinates and narrows a search. The connector **AND** narrows a search, retrieving only those records containing at least one term or phrase from each concept. The **AND** operator is a binary one; that is, it operates on the terms or phrases on both sides of it. It is the same concept as **intersection** in set theory.



Example of AND Operator

Using Yahoo! document counts, the results of the query **"endangered species" AND "peregrine falcon*"** is:

endangered species	2,220,000
peregrine falcon*	210,000
endangered species AND peregrine falcon*	33,100

Note the **AND** operator says nothing about where the terms or phrases are located in the document with respect to one another, nor whether their linkage makes sense or not. This operator only requires that the terms or phrases immediately on both sides of the **AND** must both appear in the document.

TIP:
AND *should be your most frequently used Boolean operator.*

The **AND** operator can be used to chain a number of required terms or phrases together, all of which must be present in order for the outcome to be a successful result. For example, the query **London AND "Big Ben" AND "Buckingham Palace" AND Trafalgar** would only return documents that contained all four terms or phrases.

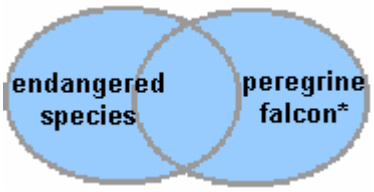
The **AND** operator is also a very useful qualifier. For example, Yahoo! counts for **falcon*** total 7,220,000. Some of these references are to cars, others to various companies, falconry or a sundry of products using the name falcon. To zero in on the falcon bird, a search phrase of **falcon AND bird*** removes these extraneous references. The Yahoo! document count now becomes 738,000.

False "results" can be common using the **AND** operator. For example, let's apply Jan's query of **endangered species AND peregrine falcon*** to a large document discussing unusual birds. In one section it could discuss the 200 mph diving speed of peregrine falcons; in another the extinction of the dodo bird. A positive result would be scored for this document, even though there is no discussion about the endangered status of peregrine falcons. One of the reasons these false positives occur on the Internet is the occurrence of large Web documents that simply list links or references to other documents and contain HUGE numbers of terms. They often produce false results.

Topic 15: OR Operator

OR means "I want documents that contain ***either*** word; I don't care which word." **OR** broadens a search and makes it less focused. It is equivalent to the **union**

operator in set theory. Again, using our peregrine falcon example, the results set for this operator looks like:



Example of OR Operator

TIP:
*Use **OR** to string together synonyms; be careful about mixing it in with **AND** !*

The document counts from Yahoo! using this **OR** operator are:

endangered species	2,220,000
peregrine falcon*	210,000
endangered species OR peregrine falcon*	2,270,000

Note the **OR** operator is NOT strictly equivalent to a sum. Documents which contain both phrases still get counted as a single document. Nonetheless, our document results counts do approximate the total of the two separate queries.

The **OR** operator can be used to chain a number of terms or phrases together, any one of which must be present in order for the outcome to be a successful result. For example, the query **London OR "Big Ben" OR "Buckingham Palace" OR Trafalgar** would return all documents that contained one or more of these four terms or phrases. As with the **AND** operator, there is no assurance that any of these terms or phrases are logically or conceptually linked in any of the results documents.

Unless used in parenthetical clauses (most useful for synonyms) or as a fishing expedition as part of preliminaries to a search, we do not recommend the use of the **OR** operator. Overuse of the **OR** operator can cause results sets to grow too large to be useful.

Nonetheless, the **OR** operator is one of the two main operators within Boolean syntax. It should be used in a controlled way to expand your results set, most often as part of a parenthetical argument dealing with synonyms or closely related concepts.

PART 5: ADVANCED OPERATORS

There are four additional Boolean operators that provide more fine-grained control than the basic **AND** and **OR**. These operators are less frequently used and are generally *NOT* supported by search services with basic Boolean capabilities.

Topic 16: AND NOT Operator

AND NOT removes any documents that contain that term or phrase. **AND NOT** is a unary operator; that is, it only works on the term or phrase that immediately follows the operator. It does not evaluate terms or phrases on both sides of the operator.

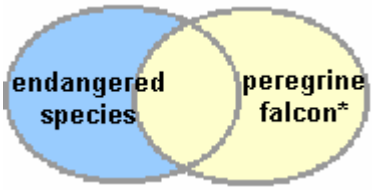
Most of the major search services support the **AND NOT** operator. It is sometimes called **BUT NOT** or **NOT** and sometimes denoted by placing a minus sign (-) before the term or phrase to be removed. NOTE: Technically **NOT** is the unary operator. For example,

NOT falcon

would exclude all documents that use the word **falcon**. The problem arises in the middle of a query. While some search engines allow **NOT** by itself, such as:

falcon NOT car

which would return documents using the word **falcon** but not **car**, the statement is technically ambiguous as to how to treat **falcon**. As a result, most engines require matching **NOT** in the middle of a query with **AND** or **OR** (**OR NOT** is rarely used). This removes the ambiguity and is the form we've adopted herein for use within the middle of a query.



Example of AND NOT Operator

Again using Yahoo! document counts, here are the results for this operator:

endangered species	2,220,000
peregrine falcon*	210,000
endangered species AND NOT peregrine falcon*	2,240,000

TIP:
AND NOT *is a powerful operator, use with care! A single instance will cause a document to be excluded.*

AND NOT is a very powerful command that should be used with care. **AND NOT** works to narrow a search, subtracting all citations that contain the specified term or phrase.

AND NOT is completely non-discriminatory; it only takes one instance of a word or phrase to eliminate a document from your results set. As one source describes it, think of **AND NOT** logic sort of like peeling a potato [33]. A peeled potato is **potato AND NOT peel**. There's only one trouble: some of the good part of the potato goes with the peel. So, use the **AND NOT** operator with as much care as you would a paring knife, and only when you're absolutely sure you want to exclude a term or phrase from your results.

Generally, we do not recommend using **AND NOT** in the beginning iterations of a search. See what results are obtained in the early steps before applying this operator, if at all. Then, apply it incrementally to make sure you're not stripping away too much of the fruit.

A good example of where this might apply is with the **falcon*** search noted for the **AND** operator. The term **falcon*** returns references to cars, products, companies and place names, in addition to birds. Successively applying **AND NOT** to **car***, **product*** and **compan*** is another approximation to the search **bird* AND falcon***. On the other hand, using **AND NOT** with **place*** could be going too far by eliminating references to falcon bird sightings that occur in various places.

Though in this example we have a good **AND** qualifier in **bird*** for our interest in peregrine falcons, a suitably encompassing word such as **bird*** may not apply to other search topics. In these cases, **AND NOT**, judiciously applied, can be an alternate way of getting to the same end.

NOTE: The **AND NOT** operator may use that form, instead using the **NOT** or the "-" sign. Depending on the search engine, it may also be applied through a 'Must not include' dropdown list choice. Depending on the engine, you may also not see proper behavior when phrases are used or other anomalies.

Topic 17: NEAR Operator

Remember for the **AND** operator that the terms or phrases on both sides of the operator can appear *anywhere* in the document in order to get a successful result. One example above described how a successful result for **"peregrine falcon*" AND "endangered species"** could be obtained, even though the falcon reference was to 200 mph diving speeds and the endangered species discussion was many

[33] See http://www.netstrider.com/search/logic.html..

pages later dealing with the dodo bird. The **NEAR** operator is designed specifically to avoid such false results.[34]

The **NEAR** operator requires the two phrases or terms to be within a specified word count of one another to be counted as a successful result. Generally, most search engines that support the **NEAR** operator have a set value of a ten word maximum distance. Some engines also use **ADJ** (for adjacent) as the equivalent operator to **NEAR**.

TIP:
*Use **NEAR** as an alternative to phrases and an improvement to **AND**, but only when you know the concepts are closely linked.*

The **NEAR** operator does not care which of the phrases or terms on either side of the argument comes first or not, just that the two phrases or terms are within the specified distance.

The **NEAR** operator, is supported, is a great way in very specific instances to ensure that your search terms occur within the same sentence or same paragraph. It can remove large, comprehensive Web sites that have a reference to everything under the sun, but not specific information of use to your search.

The **NEAR** operator can have drawbacks, however. It is possible to overlook the definitive document on endangered peregrine falcons, for example, if in one section of the document it uses peregrine falcon but elsewhere when its endangered status is discussed it only uses the word peregrine. It is very difficult in all cases to foretell how document authors will use, repeat or link such terms.

Another drawback is that relatively few search services, and none of the major ones, support this operator. This problem can be overcome when using third-party search tools such as BrightPlanet's Deep Query Manager™ (see **PART 12**) that work on the results of search engines but support this operator themselves.

But, if your terms can pass the test of confidently appearing within a sentence or so of one another, we recommend you consider the use of the **NEAR** operator.

Topic 18: BEFORE and AFTER Operators

The **BEFORE** and **AFTER** operators work in the exact same manner as the **NEAR** operator, only you can now specify which terms or phrases need to come first or second. In the case of the **BEFORE** operator, the first term or phrase MUST occur BEFORE the second term or phrase within the specified word distance. In the case of the **AFTER** operator, the first term or phrase MUST occur AFTER the second term or phrase within the specified word distance.

[34] Even though current major Internet search engines do not support these adjacency operators, third party tools do and the techniques are still useful.

These operators do provide greater control to your searches. But their drawbacks are even more severe than the **NEAR** operator. First, not only must your terms appear within the word distance, but you also must get the order right.

For these reasons we've included these operators here for the sake of completeness, but we do not recommend that you seriously consider using them . If you become an Internet 'power searcher' and you decide you disagree with this recommendation, then your skills have surpassed the purpose of this tutorial anyway.

PART 6: ADVANCED CONSTRUCTION

This part builds on the Boolean operators and basic search concepts previously discussed to show how they can be combined into effective, complete queries. Much of the discussion concerns how to construct proper syntax. This part ends with a reprise of our sample search problem for Jan's mystery bird [see **Topic 5**]. The guidance below, however, should be generally applicable to most engines that support structured, Boolean syntax:

Standard Syntax	Meaning	Alternative Syntax	If Not Supported
AND	both required	**+**	ignored
OR	either required	blank	all support
AND NOT	exclude following	**-, BUT NOT, NOT**	ignored
NEAR	required within set word distance	**ADJ**	ignored
BEFORE	first required before within distance		ignored
AFTER	first required after within distance		ignored
( )			ignored
" "	treat as phrase	checkbox option	treated as **AND**
*****	stem word	checkbox option	ignored

Topic 19: Use of Parentheses

Search services that support structured (Boolean) syntax do not always read from left to right like we do. Instead, they read "inside-out", in order of the nested levels of arguments set off by parentheses. Each bounded argument set off by parentheses is called a **Boolean expression**. (The entire query is also assumed to have parentheses around it, whether you put them in or not.) This is the same concept drummed home in high school math in how to evaluate an algebraic expression.

Learning how to construct this Boolean syntax structure is easy. You only need to remember four things:

1. You define a Boolean expression through use of an open parenthesis ['('] to begin it, and a closed parenthesis [')'] to end it
2. Make sure the first search concept you want evaluated is at the inner-most level of your Boolean expressions; followed by subsequent expressions in your desired order
3. Make sure you have a balanced (equal) number of open and close parentheses in your entire query
4. Expressions at the same "level" are read in order, from left to right.

It is really worth your time to master these simple rules. It adds immensely to your control over your queries and their ability to return the results you desire.

Though some search services support quite a few layers of nested Boolean expressions, in practice the amount of nesting you need or is even desirable is quite low, likely no more than three at most. And, it is also the case that most current search engines do not support parenthetical nesting. To show a three-level example, consider the following dummy query:

THIRD expression (SECOND expression (FIRST expression evaluated) evaluated) evaluated

TIP:
Don't heavily "nest" your parentheses. Remember, keep it simple!

Note, you do not need to put parentheses around the entire query; the outermost layer is evaluated last in any case. But, even when you think the computer is going to do what you want, it is always safer to use parentheses if there is even a chance of confusion. Parentheses will also help you read your own searches.

In the absence of any nesting, or with expressions at equivalent levels, the order of query interpretation is from left to right. For example:

FIRST expression AND SECOND AND THIRD AND FOURTH

or,

(FIRST main subject) AND THIRD expression AND (SECOND expression)

AS A GENERAL RULE, YOU SHOULD ALWAYS PLACE YOUR MAIN SUBJECT TO BE EVALUATED FIRST. This is because many search engines determine the rank order of document results by relevance, with first query terms to be evaluated ranked higher. This rule can be a bit tricky until you get used to it. For example, taking the last query example above, but forgetting the initial set of parentheses shown, produces the following:

TIP:
Don't assume an evaluation order. Specify the order you want by using parentheses.

SECOND main subject AND THIRD expression AND (FIRST expression)

Using the form above, if you placed your main query subject first in your query expecting it to be evaluated first, you would get the unintended consequence of having it evaluated second.

Finally, Boolean operator precedence is enforced by most search engines with **AND** and **AND NOT** being evaluated before **OR**. If you have doubts of operator precedence, consult the help system for the search engine being used. Our recommendation: eliminate ambiguity as to how a given engine treats operator precedence by explicity putting your expressions into parentheses in the evaluation order you desire.

The **OR** operator should generally be used solely within nested expressions, and then mostly to capture synonyms.

For example, you may recall from our sample problem of Jan's mystery bird that Jan wanted the concept of having seen the bird in the city as part of the query. Also recall there is a problem with picking up too many unwanted words when city is truncated as **cit***. A good way to handle this problem is with a nested Boolean expression using **OR**. Thus, to capture both the singular and plural forms of city, Jan would write:

(city OR cities)

This expression now covers the singular and plural without inadvertently adding undesired words (such as 'citizen' or 'citrus') to the query term list.

Whenever you mix Boolean operators in a query you should always use parentheses to force the evaluation order you want. This helps avoid unintended consequences. For example, the following query without parentheses.

hawks AND eagles OR falcons AND owls OR vultures

May actually be evaluated as:

(hawks AND eagles) OR (falcons AND owls) OR vultures

The result of this expression is not very useful. The expression does not require any one term. You could end up with pages containing only vultures or only owls and falcons or only hawks and eagles. This is most likely not the way you intended it.

Lastly, there are times when parentheses are not needed. This is when all operators are either **AND** or **OR** in the query. For example,

hawks AND eagles AND falcons AND owls AND vultures

or,

hawks OR eagles OR falcons OR owls OR vultures

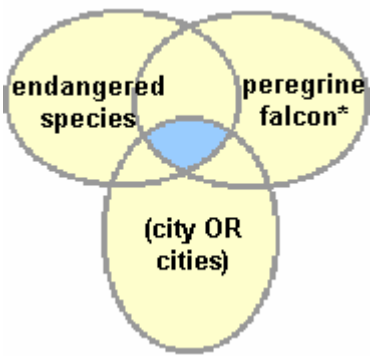
The former requires all five types of bird to be included in a successful document; the latter only one. Additional examples of possible pitfall query syntax is shown in **PART 7**.

Topic 20: Combining Concepts for Power Searching

A good rule of thumb when searching for relatively hard-to-find information on the Internet is to juxtapose three "concepts" in your query (we've also used the term **Boolean expression** to represent a "concept"). The first concept should be your subject, defined at the proper level [**Topic 10**], with synonyms or phrases as appropriate to provide adequate yet accurate subject coverage. The other two concepts should correspond to two of the when, where, how and why concepts discussed in **Topic 6**.

Each of these concepts should be provided as a Boolean expression with the **AND** operator connecting all three. In the case of Jan's mystery bird example, the resulting query can be represented as:

TIP:
*Try to link three concepts together in your queries, joining with the **AND** operator.*



Combine Concepts in Query

Note how this acts to restrict your final results space. Posing this query to Yahoo! in the form:

("peregrine falcon") AND ("endangered species") AND (city OR cities)

produces a results set of 9,740 documents. This number may sound like a lot, but remember we began with millions, and as Jan discovers, the first twenty of which (at least) directly respond to the desired results.

You should generally not need to exceed three concepts in a successfully constructed query; four is unusual. If you find you can't narrow them to two or three, double check to be sure all the concepts are necessary and all are at the right level.

Topic 21: Punctuation and Capitalization

Not all search engines handle punctuation equivalently. When in doubt, you should consult the help file of the search engine you are using. Most search engines are insensitive to whether you use upper, lower or mixed case in your queries. You are generally safest to use lower case. Where the engine does

support upper or mixed case, if you use upper case characters the engine assumes you want an exact match. Most engines also do not care if you use upper or lower case for Boolean operators, but there ARE exceptions (such as MSN Search) that may require all upper case.

For the few engines that do support capitalization, you can use this fact to advantage in finding proper names or place names.

Topic 22: Query Term Refinements

Finding the "right" terms or phrases in formulating a query can be quite tricky. There are sometimes technical terms or terms of art in a given domain that us, as lay people, may not be aware of. For example, a concern over **"bone loss"** or **"calcium deficiency"** may result in documents that more precisely define the condition as **osteoporesis**.

Here are some techniques for finding alternative – and more precise – query terms or phrases:

1. Simply inspect some of the initial results returned by the search engine. If more specific terms or phrases are discovered, substitute them into your new query
2. Use an online site that provides related words based on an initial term submission. One of the most excellent of these is the Lexical Freenet site.[35] In submitting the term **falcon** to this site, here are some additional terms or phrases that were identified: pigeons, nest, hawk, hunt, american kestrel, caracara, falco columbarius, falco peregrinus, falco rusticolus, falco sparverius, falco subbuteo, falco tinnunculus, gerfalcon, gyrfalcon, peregrine falcon, pigeon hawk, sparrow hawk
3. Use an online site to find related nouns in the same set. A very useful site for this is the experimental Google Sets facility.[36] By entering **falcon**, **owl**, **hawk** and **eagle** to this site, these additional prompts were obtained: raven, swan, crow, duck, wren, heron, cardinal, vulture, parrot, penguin, goose, thrush, ostrich, hummingbird, pigeon and peacock. These prompts are not as useful for this query term, but in some cases can be quite helpful
4. Look up your initial query terms in an online or bookshelf dictionary.

Topic 23: Sample Information Problem Revisited

Through successive refinement of the subject, Boolean expressions and query syntax, Jan found a listing of 9,740 Web documents related to the mystery bird, the most highly ranked of which met the desired results. Here's what Jan discovered:

[35] See http://www.lexfn.com/.
[36] See http://labs.google.com/sets.

- The mystery bird was a male, peregrine falcon. Nearly lost to extinction, in at least the Eastern U.S., the bird was making a stunning comeback through a combination of breeding-and-release programs and a cleaner environment free of DDT
- Peregrine falcons had found a natural home in downtown cities, where the building ledges gave them protection as their natural cliff habitats had, and where there were plenty of delectable pigeons to feed on
- Breeding pairs of peregrine falcons were now found in such urban areas as Cincinnati, Dayton, Columbus, New York City, Cleveland, Toledo, Chicago, Milwaukee, Toronto, Montreal, Philadelphia, Wilmington, Baltimore, Washington, DC, Salt Lake City and Pittsburgh
- From a base of zero in the 1970s, there are more than 1,000 breeding pairs now known East of the Rocky Mountains
- Live-cams showing peregrine falcon nests on building ledges are now being beamed 24 hrs per day over the Internet from Toronto, Montreal, Columbus and Pittsburgh
- Jan's sighting in Minneapolis was the first recorded in that city
- Tremendous additional information was gained about great viewing sites for peregrine falcons at nature preserves and general information about the species.

Jan came to understand that the recovery of peregrine falcons was one of the great environmental success stories of the past two decades. Jan is presently setting up Minneapolis' own live-cam to monitor the new breeding pair in that city. Jan is also now a local celebrity and resident authority on peregrine falcons.

PART 7: PITFALLS TO AVOID

This part describes many of the common errors made by Internet searchers. Some are within the control of you, the searcher. Others are due to the rapid growth of the Internet and the inherent limitations to search services on the Internet.

Topic 24: Avoid Misspellings

You know, it's so obvious that it is most often not mentioned: Searchers on the Internet are atrocious spellers. See for yourself. A number of existing search sites such as Word Tracker, Dogpile, MetaCrawler, Excite and Webcrawler, among others, enable you to monitor in real time the queries being issued on the Internet. Observe for yourself bad spelling, not to mention bad or totally lacking query construction.[37] (WARNING: unless you use a filtering option if available, you may see graphical sexual content.)

It is not the purpose of this tutorial to rap people on the knuckles if they misspell words. But, in your query and searching, if you misspell your keywords, you are immediately penalized. Let's do a little exercise to test this with the terms **query** and **searching** used in the previous sentence. Again, our document counts are based on Yahoo!:

query	31,100,000
querry	44,900
qerry	842
kwerrie	37
searching	38,100,000
serching	28,200
searchng	1,380
seerching	35
sherching	40

Clearly, Web developers also misspell words on their own documents (don't we all!). (Note: some of the misspelled instances above refer to this search tutorial on-line.)

Computers and indexing algorithms are inherently stupid. If the Web developer misspells a word, it is entered as such on the database. If the searcher issues a misspelled query term, that is what is searched for. So, recognize that computers are stupid and guard against these mistakes yourself. Sloppy entry of query terms will cost you time and cause you frustration.

[37] See banner at top of http://www.searchspy.com/, or http://www.dogpile.com/info.dogpl/searchspy/, or http://www.metacrawler.com/info.metac/searchspy or http://www.infospace.com/info.xcite/searchspy or http://msxml.webcrawler.com/info.wbcrwl/searchspy/.

Fortunately, most major Internet search engines – while issuing your query as entered – will prompt you for possible spelling variants if they detect possible misspells. Make sure to note where your favorite search engines provide these prompts.

Topic 25: Redundant Terms

Think of constructing a query as being in a card game. You have only so many cards (terms) to play to get a winning hand, or successful results from your query. Using redundant terms "burns" one of your cards, and diminishes your prospects for success.

TIP:
Limit your keywords to six to eight. Check to make sure you're not duplicating "levels" in your terms.

Redundant terms mostly arise from combining terms from multiple "levels" dealing with the same concept [see **Topic 10**]. For example, in Jan's search case, the subject of the query became **peregrine falcon***. Were Jan to also add **bird*** to the query it would repeat information – at the wrong level to boot.

You can generally spot redundant terms by asking the question, "Is this term already covered by another term?" If the answer is yes, pick the term at the appropriate level and discard the other one.

Topic 26: Ignored Terms and Special Characters

Recall from **Topic 7** the class of terms known as "stoplist" terms. These are common conjunctions, prepositions, articles or verbs that are generally ignored or stripped out of your queries. Depending on the Internet search engine, these stoplist terms can be as many as 600 or more common words.

But there is another emerging class of words that are also becoming like stoplist terms – often ignored by the search engines because of their ubiquity on the Internet. Examples include: computer, Internet, Web, sex and software. These words, and others like them, are not always ignored. It appears that at high-demand search times that some of the engines choose to ignore processing them.

Should you experience such behavior, one solution, if you indeed need to use such ignored terms in your query, is to make sure that you place these words in quotes or make them part of a phrase. The ignored behavior appears to be limited to use of such terms as individual words in queries, and then only at some times of the day.

You should also be aware that most of the search services covered in this tutorial handle do not handle special characters such as: **~ ! @ # $ % ^ & ( ) = | { } ‘ “ < > ? / , . _** , or non-English language characters such as the cedilla (ç) or umlaut (Ö) (or many others). Depending on the engine, the special characters

are generally ignored, but are sometimes treated as a space. Generally, too, the characters of - and + have reserved meanings for **NOT** and **AND**, respectively.

Topic 27: Alternate Spellings

English has become the standard language for Internet communications. However, some of the largest user domains on the Internet come from a background of traditional public school (U.K.) English. There are perhaps 50 countries around the world whose English is traditional, and not based on usage and spelling in the United States.

As a searcher, you should be aware that many common terms – colour/color, organise/organize, behaviour/behavior – may differ in spelling between these two forms. If you suspect that a keyword in your queries may have alternate spellings, we advise you to treat these alternates in the same way you handle synonyms: list both forms in an **OR** Boolean expression.

Topic 28: Too Many Terms, Synonyms

We have recommended throughout this tutorial three overall guidelines for the size of your queries:

- Avoid the 1.5 keyword trap for your queries and be more expansive, but
- Limit the key concepts (*e.g.*, Boolean expressions) to three or fewer; under rare occasions this guideline can increase to four
- Keep the actual terms in your queries to no more than six to eight.

These guidelines are not just a goad to refine query construction, content and syntax. They are also driven by experience that indicates that at high numbers of term counts search engine behavior can become erratic and unpredictable.

For example, most major search engines appear to shorten the issued query to a given (and undefined) number of characters. You may be able to enter a long query, have it apparently issue well without warning, but only the first portion of a long query may be evaluated. There are other anomalies that occur as well, such as different query treatments between basic and advanced search forms, undocumented default behavior, and others.

It is difficult to judge these anomalies, since each search service closely guards how it indexes, retrieves and scores queries. Attraction of eyeballs has become a highly-competitive factor of the Internet; many are vying to gain advantage in where they are listed on search engine results; and there are real technical demands to serve all search requesters in real time at peak demand periods.

The fact that search service rules are today opaque is unlikely to change any time soon. As users, we are left with observing engine behavior, reading the public

help documents, and gleaning insights from others on the Web who have been focused on similar questions. This is not really an attractive state of affairs. Absent definitive and public disclosure by the search services of how they handle these matters, room for misinterpretation and misunderstanding looms large.

Topic 29: Improper Boolean or Complicated Construction

PART 6 describes advanced construction of Boolean queries. This topic elaborates on four pitfalls that you may encounter:

- Excessive nesting or terms, which search services may not process in all instances and which may not achieve what you want the query to do
- Unintended results from combining the **AND** and **OR** operators
- Improper (and unintended) use of the **AND NOT** operator
- Unbalanced parentheses.

Let's reprise a complicated form of our standard mystery bird query, only this time focusing on citations in those cities which are known to have Internet live camera shots of falcon breeding pairs. The number shown after the query is the number of documents identified by Yahoo!. Let's say our first query is as follows:

("peregrine falcon") AND ("endangered species" OR extinct) AND ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [4,610 counts]

Whew! That's a complicated query. Let's also say that we are ambivalent about whether the endangered species status or the listing of cities both need to be in our results set. We could thus change the query as follows:

("peregrine falcon") AND ("endangered species" OR extinct) OR ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [1,010,000 counts]

Whoa! Why did the results set zoom to over 1 million? First, because of the precedence order of evaluating nesting, the query above is really being evaluated as follows:

(("peregrine falcon") AND ("endangered species" OR extinct)) OR ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [1,020,000 counts]

This really amounts to both sides of our query being evaluated independently, and then combined:

TIP:
Avoid complicated nesting with too many parentheses; they can sometimes give results you did not intend.

("peregrine falcon") AND ("endangered species" OR extinct) [31,200 counts]
((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding) [987,000 counts]

Clearly, this is not what we intended. We can try to fix the evaluation order by changing the nesting order by now bracketing around the two concepts for which we didn't have a preference, endangered species status or presence in one of the named cities:

("peregrine falcon") AND (("endangered species" OR extinct) OR ((Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) AND breeding)) [35,100 counts]

The main point of these examples is that combining **AND** and **OR** operators in long, complicated queries can lead to undesirable results and some difficulty in figuring out what is being evaluated first.

A more important point is to slim down your query terms and make your construction simpler. Taking the first query above, let's do that. We first get rid of **extinct**; we think it is covered pretty well by **"endangered species"**. We then decide to eliminate the **breeding** term because we deem it to have much lower informational value than the other query concepts. Finally, we will put all of the concepts at the same evaluation level by linking them with **AND** operators and putting each within its own parenthetical listing. Our streamlined query now becomes:

("peregrine falcon") AND ("endangered species") AND (Montreal OR Toronto OR "New York" OR Columbus OR Pittsburgh) [3,680 counts]

Now, our results set has become acceptably low, and the query is easier to read and understand.

So, despite the fact that Boolean queries can become quite complicated with different operators that you can use, the better rule is **Keep it Simple**. As long as you try to combine two or three query concepts at the same level linked via the **AND** operator, you should be just fine in getting meaningful results.

A different kind of unintended consequence can arise from the use of the **AND NOT** operator. To illustrate this, let's take this query as our starting example:

hawk AND eagle AND falcon AND raptor [38,200 counts]

We see, however, that we violated one of the rules of mixing redundant terms at different levels. Hawks, eagles and falcons are all raptors. So to test what happens when we pull the **raptor** term out, we try the **AND NOT** operator:

hawk AND eagle AND falcon AND NOT raptor [190,000 counts]

But, wait, why didn't our document count go down? It went way up! Didn't we remove a term from our query?

This is a good illustration of a common misperception about operators and the universe upon which they operate. In fact, based on the left-to-right evaluation rule (absent nesting), the universe upon which the **AND NOT** operator was working in this query is:

hawk AND eagle AND falcon [227,000 counts]

Thus, some 37,000 of these source documents do not contain the word raptor.

Lastly, unbalanced parentheses can be a common mistake in query formulation. All of the leading search engines that support Boolean queries test for this and give you a bad syntax error should you forget an open or close parenthesis. However, if you keep your nesting simple as we recommend, you should minimize occurrences of this mistake.

PART 8: USING FILTERS

Filters provide a different dimension or perspective by which you can "slice and dice" your search results. They are totally independent of the query. Filters determine the population to which a given query can apply.

Most of the major search engines support filters to greater or lesser degrees. Some also offer filter capabilities unique to themselves. For certain specialty searches or needs, you can use these unique filter capabilities to great advantage. You may want to check out the comparison chart in **Topic 35** to see how the major engines stack up and which unique capabilities they offer.

TIP:
Filters provide a useful complement to queries to target and restrict your results.

Topic 30: Site Filters

Site filters allow you to limit your search universe to specific or partial specifications contained in a site's universal resource locator, or URL. The URL is what you link to when you click on a reference in a Web document or enter a new site address in the location edit box on your browser.

To use site filters effectively, you need to understand what is contained in a URL. Let's take this one as an example, which we'll look at in parts:

http://www.brightplanet.com/deepcontent/tutorials/search/index.asp
------ ---------------------- ------ ---------------------------------- -------------
1 **2** **3** **4** **5**

1 The **http://** is a standard prefix to all Web site addresses. You may not even see it in all cases, because if it is lacking, your browser assigns the prefix to the URL. You should ignore it when using site filters (in other words, DO NOT enter it or use it!)

2 The **www.brightplanet** is the subdomain name. It often has a **www** prefix (for World Wide Web), or it may not. You can generally ignore the **www** in any case with site filtering. The subdomain is all information that appears between the **http://** and the major domain or country name (**3**). It can sometimes appear in multiple parts, especially for larger organizations that may have multiple servers accessing the Internet. For example, for an educational institution you might see **bigserver1.mystateu** shown as the subdomain name. Most often the subdomain contains useful identifying information about the organization (**cornell**, **microsoft**, **ibm**) to search on

3 The generic, major domain name (**com**, in this case) is shown in this field. This is one of the broadest and most useful site restrictions you can apply to

specialty searches. The use of generic domains is heavily oriented to United States sites. The major domain names are:

com – companies and commercial sites
edu – educational institutions
gov – government organizations
mil – military organizations
net – Internet service providers and services
org – non-profit organizations

These major domains are now expanded to include:

biz – businesses
coop – cooperatives
info – all uses
museum – museums
name – for individuals
pro – for professions
aero – telecommunications and aerospace.

Additional top-level domains will be added over time, though the newest ones generally have limited usefulness as filter restrictions. To monitor changes in domain names, see the ICANN Web site[38]

3 You may also see country domains (also known as geographical or ISO3166 domains) are the top-level domains maintained by every country and territory in the world. These domains are organized by locality, and are useful to organizations and business that wish to operate overseas OR want to protect their company or brand identity. Like generic domains, country domains are accessible to any user of the Internet. Country domains have two-letter designators, *e.g.* .**fr** for France, .**uk** for the United Kingdom, .**au** for Australia, .**us** for the United States (not generally used), etc. There are over 230 top-level geographical domains, of which about 190 currently accept domain registrations. You may obtain a complete listing of these abbreviations from[39]

4 This information is for the subdirectory locations where the information resides on the subdomain's Web server. This field can contain useful information, such as **deepcontent** or **tutorials**, but is sometimes quite cryptic and often can be quite long. Note that absent a designation in this field you are generally directed to the home, index or main page of the given site. Also note that some engines that support site filtering do not allow you to search in this field

[38] See http://www.icann.org/tlds/.
[39] A complete listing of country two-letter domain codes can be found at: http://www.thrall.org/domains.htm..

5 All information prior to this point identifies how to get to the given physical location where the Web documents reside. Field **5** now represents the specific Web page for the document (if it is missing, index.html or similar is assumed). Note, however, that this last field contains the .asp file extension. It is also a useful way to filter content for search engines that support it.

Generally, fields **2** and **3** are the most useful to use when restricting sites. **5** is subject to much variation and is not always supported, but the file extension can be helpful. We recommend that you only use it when you have advance information or specification of the given document(s) for which you are looking.

When using site filters, you need to be careful that you don't enter too broad a specification. For example, using '**com**' as a site filter specification would result in including sites with the '**.com**' domain as well as sites such as **commonplace.edu**, **commercial.net** or **markettips.org/commercialization.html**. Attentive use of periods ('.') and slashes ('/') can help narrow your restrictions for those search engines that support the site filtering feature.

Topic 31: Size Filters

Presently, no major search services are known to filter documents by size. There are third-party products, however, that can do so.

Topic 32: Date Filters

Date filters can be especially useful when doing research on time-sensitive information. Depending on the engines that support this feature, you can restrict retrievals to documents modified since a certain date or within a range of dates. Date filtering provides a good argument for keeping a record of your exact query and its date for very important searches. Then, should you want to see what results have been updated or added to the Internet since your last search, you can simply re-submit the initial query and select the appropriate date restriction.

NOTE: this option on most major search engines is not very accurate. The dates shown used by the engines are (generally) the date the page was indexed, not created. (Date created fields are available to Web developers, but not all use them. Also, not all engines read this field, anyway.) Some search engines are running days to weeks behind in indexing pages. To prevent possible gaps in your date searches, you may want to consider moving the start date back from the absolute date you want to filter.

Topic 33: Specialty Filters and Search Options

In the competitive race to provide more features, many search engines are providing specialty filters and search options. For a listing of these features by

major services, see **Topic 35.** Here, however, we describe what options are available. Please note these options are supported by only a limited number of services. Also note that these features may be described slightly differently by different services; consult their specific help files.

- **People's Names** – there is often a 'white pages' option provided by many major Internet search engines. In addition, there are special engines on the Internet specifically for finding people, such as Switchboard.com
- **Business Names** – there is often a 'yellow pages' option provided by many major Internet search engines. In addition, there are special engines on the Internet specifically for finding businesses, such as Switchboard.com or InfoUSA.com
- **Depth** – provides the ability to retrieve additional pages from a given site; 'depth' represents the nested levels to retrieve
- **Applet** – identifies documents with Java applets corresponding to the name provided
- **Domain** – finds documents restricted to the country or generic domain specified
- **Host** – finds documents on the specific computer specified for 'host'
- **Image** – identifies documents with images (graphics) corresponding to the filename specified
- **Link** – finds documents with links to the URL specified as the argument
- **Title** – identifies documents that contain the word or phrase specified in their titles
- **URL** – finds documents whose URLs match the word or phrase specified
- **File/Media Types** – identifies documents which contain the file or media type specified; useful, for example, in finding Microsoft Office® documents or audio or video
- **Language** – searches can be restricted by up to one or more of about 40 different languages.

PART 9: UNDERSTAND YOUR ENGINES

Effective searching requires understanding how best to utilize the features of your search services. But, Internet searching is a highly-competitive, dynamic area. New search engines are cropping up continually, others are folding or being acquired, and feature sets change almost daily in order to keep pace.

Topic 34: Boolean or Not?

For serious searching, perhaps the most important first choice facing you is choice of search engines. Which search engines better cover the topics you are interested in? Which support the search features that will enable you to find what you want?

TIP:
Use search engines with full-text indexing and Boolean support for your most demanding queries.

Not all searches are created equal. The increasing ability of some search engines to take your requests in context, and then enable you to narrow results based on your first attempt, is a promising development. Certainly being able to type in a few words and then begin receiving documents of value bodes well for common-topic searches. We ourselves use this approach when quick searches are needed.

We doubt, however, the ability of search engines in the near term to improve on this process for complicated searches or for hard-to-find information. Not only is coverage of such topics weak for a given engine, but the ability to anticipate refinements is weakened by the need to categorize information into levels insufficiently specific to the difficult query.

Thus, for difficult search topics, we still must recommend the use of search engines with full Boolean support. Only you know what information you are seeking (even though it may be ill-defined or abstract). With full Boolean searching, you have complete control to find what you seek.

This recommendation, however, exacerbates the lack of coverage of any given search engine. By definition, hard-to-find information is not well-indexed, meaning you will likely need to use more than one search engine to get the robust results you desire.

Topic 35: Features of Leading Search Services

Given the growth, consolidation and emergence of new services, it is very difficult to keep pace with the specific search features of various leading search services. We recommend you study closely the help documentation for your favorite service, emphasizing the advanced search form. You may want to monitor such sites as SearchEngineWatch, Search Engine Showdown, Law Library Resource Xchange, and others for their assessment of actual search

functionality.[40] Much of the information from the table below was derived from the services' own help documentation and these sources.

	AlltheWeb	Google	MSN	Teoma	Yahoo
OPERATORS					
Default	and	and	and	and	and
AND	AND, +	+	AND, +	+	AND, +
OR	OR	OR	OR	OR	OR
AND NOT	NOT, -	-	NOT, -	-	AND NOT, NOT, -
ADJACENCY					
Nesting	(yes)		(yes)		(yes)
Phrases	"yes"	yes	yes	yes	Yes
Synonyms		~			synonym
Stemming		in phrases	advanced		
OTHER					
Case	no; lower	no; lower	some mixed		
Stoplist	no	yes; "+ " option	yes; + option		yes; "+ " option
Select no. results	yes	yes		yes	Yes
Notes			operators in uppercase		undocumented search functions
FILTERS/FIELDS					
"Safe search"	yes	yes			yes
File formats	6	8	5		9
Languages	36	35	15	10	38
Diacritics					yes
Geography/region	10		9	9	24
Localized results		yes			yes
Date	yes	yes		yes	
Title	title:	allintitle: intitle:		intitle:	intitle:
URL		allinurl: inurl:		inurl:	url: inurl:
Domain	domain:		form	form	hostname:
Link	link:	link:			link:
Site	site:	site:		site:	site:
Similar sites		related:			
Depth			5		
Numeric ranges		num ..num			
Text occurrence		5 options		2	
News	option	option			news
Products		Froogle			['Other shortcuts']
Cache		cache:			
Site information		info:			
Term definition		define:			define
Stock ticker symbols		stock:			
Stock quotes					quote

[40] See, as examples, these sources that update their charts on periodic bases: http://searchenginewatch.com/facts/article.php/2155981, or http://www.searchengineshowdown.com/features/byfeature.shtml, or http://valencia.cc.fl.us/lrcwest/searchchart.html, or http://www.llrx.com/features/searchenginechart.htm.

	AlltheWeb	Google	MSN	Teoma	Yahoo
Applets			form		
Scripts			form		
Other shortcuts					22 keywords
ESTIMATED SIZE					
Millions docs	3,200	8,000	5,000	1,500	4,200
Index depth	?	101K	150K	?	500+K

The estimated engine sizes and index depth is from a recent SearchEngineWatch posting.[41]

The options shown in the table are often noted by different terms by the services that support them, and usually involve special syntax rules. Sometimes, too, the descriptions of how these features operate is difficult to find from the main pages of the services. Directly consult each service's home page; and, then, try consulting advanced or power searching, the help sections or the frequently asked questions (FAQ) areas to read about the special operators and their rules.

Topic 36: Some Perplexing Behaviors

Search engines may not always perform as indicated on their help pages. These differences are due to constant changes in how they handle their service, strange quirks relating to their scoring and indexing methodologies, errors made by the developers of Web pages, and decisions the service may make to speed performance at high-traffic volume periods. Most telling, however, is the incomplete nature of the posted help documentation. We can illustrate some of these quirks using our standard Yahoo! search source. (It should be mentioned that most help pages have these same limitations and are not limited to Yahoo!)

The first example for Yahoo! is the fact it supports the standard Boolean operators of AND, AND NOT and NOT without noting so. It also supports nested Boolean queries without documenting it.

The second example involves perplexing Boolean treatment. Using our summary example of new planets being discovered, here are comparisons for two Boolean queries that should produce the same result:

new AND (planet OR planets)	18,000,000
(new AND planet) OR (new AND planets)	22,200,000

The third example is order on a seemingly equivalent pair of queries, which can also produce slightly different results. Let's compare these two queries:

[41] See http://blog.searchenginewatch.com/blog/041111-084221.

(bird OR birds) AND falcon 698,000
falcon AND (bird OR birds) 699,000

It is clear that results counts are provided through estimation and are not exact, and the count differences between the two queries are immaterial. But there are differences.

Note you may also get differing results counts when issuing queries at different time of the day. This may indicate that at times of high traffic limits are placed internally on the search. In the worst cases, Yahoo! or other engines may even provide a message that the server is busy, and prompt you to return at a later time to obtain results. Most often when this error occurs, a quick re-issue of the query request will obtain results.

Of course, counts are not what you the searcher wants when you search. The actual results pages for these examples were quite similar. But it's useful to realize that how an engine operates exactly may not be clear or consistent.

These points are not meant to be a criticism of the search engines, or of Yahoo! in particular. Major search engines are indexing millions of pages in very short periods of time and need to provide snappy response in all instances. The fact they do accurately index very high percentages is remarkable. But, you, as a searcher, should be aware results are not foolproof and behavior and functionality may not be exactly as documented by the service.

PART 10: THE 'DEEP' WEB

In the earliest days of the Web, there were relatively few documents and sites. It was a manageable task to post all documents as "static" pages. Because all results were persistent and constantly available, they could easily be crawled by conventional search engines. For example, in July 1994, Lycos went public with a catalog of only 54,000 documents.[42] This older way of posting documents on the Internet is known as the "surface" Web. However, in recent years, the nature of the Internet has changed.

Topic 37: What is the 'Deep' Web and How Does it Differ?

Beginning about 1996, Internet growth was shifting content sites to the "deep" Web. It is now accepted practice that large data producers such as the Census Bureau, Securities and Exchange Commission and Patents and Trademarks Office, among literally hundreds of thousands of other large content sites, use the Web as their preferred medium for information notification and transfer.

Deep Web content resides in searchable databases, the results from which can only be discovered by a direct query. In that regard, deep Web sites are very much like major Internet search engines: Results are only served up as the result of a direct request to the site. Without the directed query, the database does not publish the result. But, unlike search engines, deep Web sites are not indexed by those search engines and they must be queried directly to obtain their dynamic results.

Standard search engines obtain their listings in two ways. Authors may submit their own Web pages for listing, generally acknowledged to be a minor contributor to total listings. Or, search engines "crawl" or "spider" documents by following one hypertext link to another. Simply stated, when indexing a given document or page, if the crawler encounters a hypertext link on that page to another document, it records that incidence and schedules that new page for later crawling. Like ripples propagating across a pond, in this manner search engine crawlers are able to extend their indexes further and further from their starting points.

In contrast, results from searchable databases can only be obtained by a direct query. Without the directed query, the database does not publish the result. And, when published, the results page is dynamic, only existing within the user's browser. Thus, while the content is there, it can not be found by traditional search engine crawlers because there is no "static" URL link to follow.

[42] See http://www.wiley.com/compbooks/sonnenreich/history.html.

Deep Web content is thus dynamic, hosted on a database, and exists only in real time. First, a query is requested from the site, the result of which is a results page listing possible documents of interest, often indicated in the URL by a question mark and the repeating back of the initial query:

http://www.completeplanet.com/scripts/SearchPrime.dll?query=banking&start=1&page=15&x=49&y=7

Second, the site presents results documents links, which then takes the user to the candidate document, often indicated by a database record ID (possibly noted with the query) in the URL:

http://www.10kwizard.com/fil_blurb.asp?iacc=1250543&exp=banking&g=&Kfilter=

In these ways, searching a deep Web content database is very much akin to working with a standard Internet search engine.

Topic 38: Size and Scope of the Deep Web

In 2000, BrightPlanet first documented the importance of the deep Web,[43] or what some others have called the "invisible Web".[44] Some of the findings from updated BrightPlanet studies about the deep Web include:

- Public information on the deep Web is currently 400 to 550 times larger than the surface Web
- More than an estimated 350,000 deep Web sites presently exist, 150,000 of which have unique, valuable content
- The deep Web is the largest growing category of new information on the Internet
- Deep Web content is highly relevant to every information need, market and domain
- More than half of the deep Web content resides in topic specific databases
- A full 95% of the deep Web is publicly accessible information — not subject to fees or subscriptions.

Other independent assessments have estimated deep Web content to range from about 6-8 times to 100 times larger than the surface Web.[45] Nonetheless, no matter how measured, information on the Internet is greater than what can be obtained through standard Internet search engines alone and should be a considered portion of your Internet search strategy.

[43] The most recent version of the study was published by the University of Michjgan's Journal of Electronic Pulishing in July 2001. See http://www.press.umich.edu/jep/07-01/bergman.html.

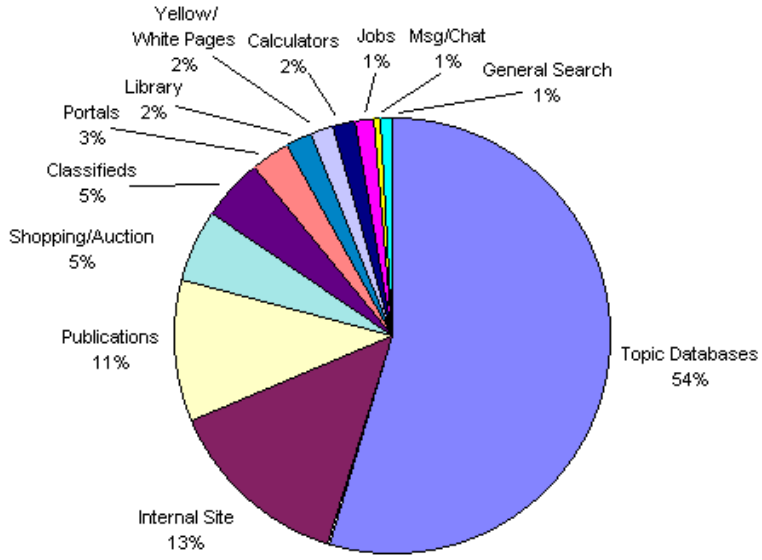
[44] But the term "invisible Web" is inaccurate. The only thing "invisible" about searchable databases is that they are not indexable nor able to be queried by conventional search engines. The real problem is not the "visibility" or "invisibility" of the Web, but the spidering technologies used by conventional search engines to collect their content.

[45] See, for example, C. Sherman and G. Price, *The Invisible Web*, Information Today, Inc., Medford, NJ, 2001, 439 pp., and P.Pedley, *The Invisible Web: Searching the Hidden Parts of the Internet*, Aslib-IMI, London, 2001, 138pp.

Deep Web content is highly relevant to every information need, market and domain, as the table below shows:

Agriculture	2.7%	Employment	4.1%	Law/Politics	3.9%	References	4.5%
Arts	6.6%	Engineering	3.1%	Lifestyles	4.0%	Science, Math	4.0%
Business	5.9%	Government	3.9%	News, Media	12.2%	Shopping	3.2%
Computing/Web	6.9%	Health	5.5%	People, Companies	4.9%	Travel	3.4%
Education	4.3%	Humanities	13.5%	Recreation, Sports	3.5%		

Nearly 80% of all deep Web sites are either topical, serve up large internal site documents, or contain archived publications, as the following figure shows:



Topic 39: Finding Deep Web Sources

Given the estimate of 150,000 valuable content databases not available through standard search engines, how do you find these sources and incorporate them into your Internet content needs?

Though there are third-party tools to assist you in this task, there are also some pretty comprehensive listings of these databases on the Web:

- CompletePlanet.com – an online directory organized by subject area that includes 70,000+ deep Web databases
- InvisibleWeb.com – a similar listing with about 2,000 sources listed
- ResourceShelf.com – Gary Price's various listings and resources.

There are other less comprehensive listings. However, you likely already know and use some deep Web sources and as you encounter new ones you can apply

many of the same search techniques in this tutorial to search them to your advantage.

Topic 40: Deep Web v. Search Engines Search Considerations

Like search engines, deep Web content databases are accessed via a search form. Most of the same tips and techniques in this tutorial will apply to these sites.

However, there is great variability to the search functionality and syntax provided by deep Web sites. Many are homegrown, or rely on third-party search utilities that do not share the feature set of major search engines. Moreover, many lack adequate search help pages (though that is true for the major engines, as well).

Third-party utilities are very useful in combining searches to multiple deep Web sites. However, if you do not have access to one of these, try to learn about your important deep Web sites and apply the same lessons in this tutorial for effective access.

PART 11: SPECIALTY SEARCHES

This part provides a compendium of tips for specialty search topics, which may be directed to standard Internet search engines or deep Web searchable databases. Most of the individual topics below simply offer bulleted suggestions for ways to first approach these searches. Once used, you will likely find for yourself additional 'power searching' tips.

Topic 41: Product Searches

Here are some tips on finding product information:

- Use product-oriented deeb Web searchable databases
- Make sure and use the actual product name in your search; use if possible a search engine that supports mixed upper and lower case
- Join appropriately stemmed search terms using the product and known company name
- Try limiting your searches with the **.com** filter; this will eliminate references from non-business sites (also note that some countries, such as Australia, United Kingdom and Canada, also use the **.com** site domain for commercial sites before the country domain)
- For product-related announcements, use the domain or url search options.

Topic 42: Competitor Intelligence

Here are some tips on finding information about competitors:

- Job listing or employment sites can be a first indicator of whether competitors are growing or not. Try searching at the individual company's site and job listing engines and monitor trends over time
- To identify new competitors, issue a query with three to five existing competitor names; that often surfaces new players
- Alternatively, but less useful, is to search resume posting services to see if many employees are bailing out. Because employees in this position are generally reluctant to announce their intentions, this tactic is generally less useful than company hiring trends. One important exception: When the company itself has internally announced a staff reduction. Sudden blips in resume postings can be a valuable early indicator
- Many of the major search engines contain sections entitled 'Company Profiles' or a similar category. Try restricting searches to these categories
- One useful way to discover partners of competitors is to use the link search option, providing the source company's name in the link text
- Generally, all company sites have a press releases section, where new advances or partnerships are often announced. Use the domain option for these filters and conjoin your specific text query. Since often press releases

are kept by news services for longer periods than on individual company sites, you may also want to make sure your query also includes the company name

- Archive your useful queries and repeat over time. Search engines that contain a 'CGI-bin' name in the query produced can be saved and used again later
- Monitor trade business news sources in your space.

Topic 43: Market Research

Here are some tips on doing market research:

- For comparative market information, first try combining words or phrases that you know appear for the leading market-share companies or products. For example, in cereals, try conjoining "Rice Krispies" and "Captain Crunch"; for computer software information, try conjoining "IBM" with "Microsoft". These should not be the ending of your query, but the narrowing beginning
- Consider using search engines that support the link, host or domain filters.

Topic 44: Finding People

Here are some tips on finding people on the Internet:

- Use specialty engines, expecially those with a white pages feature
- Use search services that support people searches (Yahoo!, Snap, LookSmart)
- Use search services that support mixed upper and lower case
- Be careful, first names are often not reliable; many individuals use initials or diminutive forms for first names ("Mike" v. "Michael" v. "M."), or may be cited by others in different ways.

Topic 45: Finding Places

Here are some tips for finding information about geographical locations:

- Try limiting your searches by country domains
- Use regional Yahoos or the geographical filters of major search engines
- Used mixed case when searching for proper place names.

Topic 46: Finding Recent News

Normal search engines and services are generally poor sources for recent news. Some of them, however, (Google and Yahoo! as two examples) have separate search options for news postings that tend to work in the same ways and with the same features as the standard engines.

There are very useful magazine and daily periodical resources on the Web. See some of the deep Web database resources noted in **Topic 39** for additional ideas. For example, more than 4,000 newspapers post their archives on the Web. Periodicals and specialty journals are also common.

PART 12: USING THIRD-PARTY TOOLS

Standard browsers and standard search engines are not the only means to find valuable information on the Web. Listed here are some third-party tools that may assist your search needs.

Topic 47: Internet Metasearch Services

There is a cottage industry of "metasearch" tools on the Web. The advantages of these sites are that multiple engines can be simultaneously accessed with a single query and only one query syntax needs to be mastered. The disadvantatges are that integrated results are often limited to the results summary only (not the full-text index) and search syntax needs to conform to the lowest common denominator across constituent engines.

Use of metasearch engines is a matter of personal preference given these trade-offs. While there are perhaps 30 capable metasearchers, the two most prominent are Dogpile (http://www.dogpile.com) and Metascrawler (http://www.metascrawler.com).

Topic 48: Desktop Metasearch Tools

Similar functionality can be brought to your desktop. Up until a few years ago, desktop metasearchers were quite popular. In the past couple of years, however, two of the leading options, Bullseye from Intelliseek and LexiBot from BrightPlanet have been pulled from the market. The leader for many years in the desktop metasearch category has been Copernic.

The major advantage of desktop metasearching is its ability to search potentially hundreds of sites simultaneously with one query syntax. Often, this syntax can implement much stronger Boolean and optional support because results are evaluated locally with the full capabilities of the desktop utility. The major disadvantage of desktop metasearching the is latency in getting candidate results due to "last mile" bandwidth limitations.

There are numerous shareware and freeware desktop metasearchers, but they are generally not to be recommended due to limited source coverage and poor performance.

Topic 49: BrightPlanet's Deep Query Manager™

The Deep Query Manager™ is a server-side solution geared to enterprises; it is not generally affordable for individual searchers. However, it has power and flexibilities found in no other Internet search solution:

- It can simultaneously access more than 70,000 search engines and deep Web searchable databases with a powerful query language and advanced filter and qualification conditions
- Its scheduler and differencing engine provides huge productivity benefits
- Its post-harvest content management and collaboration capabilities based on full-text repositories are geared to knowledge worker use and enterprise deployments

More information about the Deep Query Manager can be found at: http://www.brightplanet.com/products/dqm_benefits.asp.

VERSION NOTES AND ACKNOWLEDGEMENTS

Since the release of the first version of this tutorial more than six years ago, it has come to be the most awarded and recognized aid to searching the Internet. It has won scores of "Best of Web" awards, is used in hundreds of university curriculums throughout the world, and has been translated into many languages.

However, it was last updated in late 1999. Given the significant changes in the Internet search landscape, it was getting "long in the tooth."

The first version of this tutorial was published on April 24, 1998 by TheWebTools Company. That company was acquired by VisualMetrics in 1999, which led to the tutorial's next revision on May 6, 1999. When VisualMetrics' Internet search assets were acquired by BrightPlanet Corporation in late 1999, the tutorial was again updated and published on December 22, 1999. That was the last major update until this present revision.

The tutorial was first undertaken because of our own frustrations in finding a central resource having to do with all things "searching." The first version was prepared by Michael K. Bergman with the technical assistance of Carol Lushbough, Tom Tiahrt, Jerry Tardif and Will Bushee. The 1999 updates were supported by Tardif, Bergman and Bushee. This current revision was prepared by Michael Bergman.

Much has changed in the Internet search landscape within the past few years. The current version has been completely revised to deal with these new changes.

Throughout these versions, the authors have attempted to be as accurate and fair as possible; we welcome your suggestions for improvements or informing of us of errors.

Thanks for taking the time to learn more about how to effectively use the Internet. We hope sincerely this tutorial helps speed you along the path to discovering and accessing better information.