

---

---

# **Web Searching – Know Your Tools**

*A Dialog Graduate Education Program White Paper from  
The Professional Topics Series*

---

---

*Written by:*  
Heidi Porth, M.L.I.S. Consultant  
December 2003



As Web search engines are continually changing, it may be difficult to know which search engines and what types of features are available on the Web. "Popping" onto the Web, bringing up a search engine, typing a few words into the entry box, and clicking the Search button is standard procedure today. But exactly what does a search engine do with those few words?

What are the differences between search engines? How are the retrieved "hits" ranked? What data universe is being searched on the Web? No search engine crawls, searches, or indexes everything on the Web, and what is searched or, more importantly, not searched, can adversely affect the results.

Some elements to consider when planning even a "simple" search on the Web are the data universe, simple search defaults, ranking, search engine characteristics, the invisible or deep web, and authority or bias.

### **DATA UNIVERSE**

Although many Web search engines crawl millions of sites, they don't necessarily pick up the same information or index the same sites. Differences can include whether or not the spider has the ability to find the Web page, as well as the type of data that can be retrieved from accessible Web pages. All of these factors contribute to what is known as the data universe for a particular web search engine. For instance, many search engines have limited or no capabilities to handle different document formats, such as PDF, images, audio, and other kinds of electronic media, while others have been specifically developed to find such document types. Web pages that cannot be readily accessed or indexed by Web crawlers are part of what is often referred to as the *invisible* or *deep* Web. Therefore, it is important to determine the characteristics of different Web search engines, which will, in turn, define the data universe in which the search will be performed.

### **SIMPLE SEARCH DEFAULTS**

The next factor to consider is the search defaults of the engine. The main page of a Web search engine generally offers a simple or basic search form with only one input box. Many people type their words into the box and hit *Search* without another thought. But what are the defaults for the basic search box? Are words with spaces between them treated as separate keywords or as a phrase? Also, what sections of the Web site or Web page are examined? Is it the URL, the title, metadata, or page text, for example?

If the search terms are treated as separate keywords, is the default implied connector between them a Boolean *AND* or *OR*? This is important information to know as there can be a huge difference between the kinds of "hits" retrieved with the different types of searches.

**Example:** Find information about Nutro pet food for dogs and cats

Which of the following searches is being executed using the words **Nutro pet food** in the search entry box?

1. **"Nutro pet food"** [as a phrase]
2. **Nutro AND pet AND food**
3. **Nutro OR pet OR food**

The first instance will retrieve the fewest number of hits since it uses the most exact search parameters, while #2 will retrieve more and possibly less relevant results. Finally, the third search string will retrieve the most results, probably with the least relevance, as the presence of all keywords (or concepts) is not required in the results. Therefore, using exactly the same search string, the results can vary widely.

Often special symbols inserted in the entry box can force the execution of the search process in a particular way. To understand those symbols and how they operate, it is necessary to read the search tips for each Web search engine. Many Web search engines also have an Advanced Search option in which the searcher can be more specific about how the search is conducted. (See section on “Search Engine Characteristics” for further information.)

## **RANKING**

Another differentiation to consider among Web search engines is the various ranking mechanisms used to display the resulting list of Web pages. These algorithms can vary greatly among different services. Criteria that are often considered in ranking include statistical information, link analysis, “clickthrough,” and whether, or how much, the owner of the Web page pays for their ranking (see *Buying Your Way In*, Sullivan, May 2003).

*Statistical analysis* can include:

- how often the search words occur (especially in relation to the total number of words in the Web page)
- how close in proximity the search words appear
- the uniqueness of the search word
- in which part of the Web page the words are found (i.e. if the words occur in the URL or title, then that Web page may rank higher than a page where the words occur only in the text).

*Link popularity analysis* looks at the other sites linking to a particular Web site. Links to a Web site are often weighted differently depending upon an analysis of the linking Web page and how “good” or important that site is determined to be. This form of link analysis is an integral part of PageRank™, Google’s technology for ranking retrieved Web pages (*Search Engine Ethics*, Aug. 2003). The link from a Web page determined to be reputable or significant is weighted more heavily than a site that is not considered to be important. Many search engines also consider the significance or quality of the linking site to be more critical than the number of links to a Web site (Sullivan, July 2003).

“*Clickthrough*” can be important with some Web search engines. For instance, if few people click on an entry in the results list, then that Web page may eventually fall in the relevancy ranking.

According to Danny Sullivan in *Buying Your Way In: Search Engine Advertising Chart*, “It’s important to remember that while search engines say that paid inclusion provides no ranking boost, there is no way for the general public to easily verify this. For this reason—and the fact that the amount of paid inclusion listings are growing—it may be that ultimately paid inclusion listings will be segregated and labeled in the way paid placement listings currently are.” In other words, until the search engines identify paid inclusions as a separate category, it is uncertain whether the engine’s hit ranking system is not skewed by these “advertisements.”

## **SEARCH ENGINE CHARACTERISTICS**

To determine how different search engines operate, their defaults, and tips on how to use them, refer to either the “Help” or “Search Tips” section of the particular search engine Web site, if available, for guidelines. Sometimes the site may also offer an option to specifically search images or special document types.

Other helpful resources that cover Web search engine attributes include two Web sites with detailed information on this topic. *Search Engine Watch* offers current and archived articles on a variety of topics concerning Web search engines, as well as tables and charts comparing different features. *Search Engine Showdown* provides detailed statistics and reviews of various Web search engines. Both sites are frequently updated and contain many articles, charts, tables, and links to other references.

### **INVISIBLE OR DEEP WEB**

The invisible or deep Web is comprised of Web pages that are not accessible through most Web search engines. These pages can include some dynamically generated pages, various special document types such as PDF, images, video, audio, information contained within databases, and sites that deliberately exclude spiders.

Most Web crawlers operate by following links from other Web sites they have crawled. Spiders may never find sites that are not connected via links to other Web pages crawled or spidered, and whose URLs have not been submitted to any Web search engine. Therefore, such sites may contain authoritative, relevant data, but will not be indexed or contained in any Web search engine's database of Web sites, and, consequently, will never be part of a Web search engine's results.

The question arises as to how much information is really missing from a search that does not tap into the hidden Web. Chris Sherman and Gary Price, well-known authors of *The Invisible Web: Uncovering Information Sources Search Engines Can't See*, estimate that the invisible Web is at least two to fifty times larger than the visible or surface Web (Schlein 2003, 123). The authors break down hidden Web resources into four categories:

**Opaque Web:** consists of data that could be, but for one reason or another is not, indexed by Web search engines (Sherman and Price 2001, 70-72; Schlein 2003, 123).

**Private Web:** "technically indexable Web pages that have been deliberately excluded from search engines" (Sherman and Price 2001, 73; Schlein 2003, 123)

**Proprietary Web:** only available to people who register with the site or organization. Since Web crawlers and spiders cannot perform even the simplest registration procedures, data contained within these Web pages is not accessible to them. Proprietary Web pages may include both free and fee-based information, such as Dialog services. (Sherman and Price 2001, 73-74; Schlein 2003, 123)

**Truly Invisible Web:** Web pages that for technical reasons cannot be spidered or indexed (Sherman and Price 2001, 74-75; Schlein 2003, 124)

There are many helpful resources, supply tips, and techniques on how to access the invisible or deep Web, and browsable directories of URLs from the invisible Web (a list of several references is included below).

Web search engines or site search tools are also useful to find searchable databases—one of the most valuable types of information stores in the invisible Web. Although the Web crawler cannot access data stored inside the database, it can often find the main page. Conduct a Web search on your research topic and include the terms “searchable database” OR archive OR repository in your search along with keywords for the subject you are researching. (Note that some Web search engines require Boolean operators to be in uppercase—see the Help or Search Tips for each search engine to determine its particular requirements).

**Example:** Find searchable databases or archived information on genealogy. Enter the command below in the basic search window of a couple of different search engines such as Google.com or AltaVista.com . Note that the search technique may vary slightly for different search engines depending upon their individual characteristics.

**("searchable database" OR archive OR repository) genealogy**

This search will pick up Web sites that offer searchable genealogical databases, archives, and/or repositories. Note that your search results vary among search engines due to the differing data universes and ranking mechanisms.

***Resources To Help Access The Invisible Web***

*Find It Online: The Complete Guide to Online Research* (Schlein 2003) presents a global perspective on conducting online research. Among the tips and tools provided in this reference is an overview of the invisible Internet and ten tips on searching the invisible Web.

*The Invisible Web: Uncovering Information Sources Search Engines Can't See* (Sherman and Price 2001) provides a comprehensive look at the history of the Internet and the Web, an explanation of how and why the invisible Web evolved, tips on finding hidden information on the Web, and lists of URLs.

*The Invisible Web Directory* is the companion Web site to the *The Invisible Web*, and supplies updated URLs along with new tips and news on Web searching.

*Invisible Web: What it Is, Why it exists, How to find it, and Its inherent ambiguity* is a tutorial on the invisible Web provided by the University of California at Berkeley through their Teaching Library Internet Workshops. This guide contains good explanations and examples.

The *Internet Primer2* is a Web search tool developed by Greta Marlatt, head of Information Services at the Naval Postgraduate School in Monterey, California. *Internet Primer2* is a PowerPoint presentation that contains valuable tips on Web searching and links with more information about the deep Web.

*Web Finding Tools* from the Dudley Knox Library for the Naval Postgraduate School includes a section on resources to help searchers access the *Invisible Web*, as well as links to other useful tools.

## **AUTHORITY AND BIAS**

Since anybody can publish a Web page, the Web can be rife with misinformation for the unwary searcher; therefore, evaluating the authority and potential bias of your sources and resulting data is of paramount importance. Below are several elements to consider when reviewing your Web search results.

**Credibility:** what experience or credentials does the source (author and publishing organization) have on the topic, and how can the credentials and experience be verified? Links to and from the Web page help to evaluate credibility, as do searches on the author(s) and organizations to find out more about the sources.

**Perspective or Purpose:** what is the purpose behind the Web page (i.e. what is the potential bias of the author or organization)? Proponents of a particular industry or topic are likely to present a positive slant on the information they provide while opponents will be more apt to offer a negative angle. Think about how the source is positioned.

As John Henderson of Ithaca College Library notes in his guide to critical thinking about Web pages, “Try to identify the reason the Web page was created in the first place. Determine if the main purpose is to inform, to persuade, or to sell you something. If you know the motive behind the page’s creation, you can better judge its content. And here is an important, if difficult, question to ask: What is *not* being said?” (Henderson 2003)

**Timeliness:** is currency important to the topic, or is a historical perspective needed? Does the Web page indicate when the information was last updated?

Many academic libraries provide resources to assist in evaluating the authority of Web pages. John Henderson, a reference librarian at Ithaca College Library, has developed “a guide to critical thinking about what you see on the Web” called *ICYouSee: T is for Thinking*. His guidelines contain both criteria to consider as well as examples. In addition, searchers can test their evaluation skills using his quick Pop Quiz or his more in-depth ICYouSee Homework Assignments (Henderson 2003).

Another excellent reference tool for evaluating Web pages, *Evaluating Web Pages: Techniques to Apply & Questions to Ask*, was created by Joe Barker for the UC Berkeley Library as part of their Teaching Library Internet Workshops. A series of exercises designed to illustrate the importance of evaluating the authority of Web resources is available at *Evaluating Web Pages: Experience Why It’s Important* (Barker 2003).

In general, many resources on the invisible web tend to be authoritative (Sherman and Price 2003, 95), or more focused than surface or visible web content (Schlein 2003, 124), largely because those databases and the specialized materials of institutions (universities, government agencies, publishers, and other information providers) are accessible only through the private or proprietary Web.

Dialog products and services are a good example of high-value content that is available through the proprietary Web. The searcher can count on particular information being available and verifiable, the source of the data is clear, the publishing date is provided, and consistent indexing is applied to the records.

**SUMMARY**

Even simple searching on the Web can be more effective if the searcher considers various elements during the process. Think about the attributes of the particular search engine(s) and utilize special features when appropriate. Use online directories of sites on the invisible Web, and search for databases on the Web combined with your topic. Remember, too, that the Internet and Web are always evolving, so reading current articles on Web searching and signing up for discussion forums or newsletters are important tools for maintaining peak searching effectiveness. And always, ALWAYS, consider the authority of the data you retrieve from your Web searches. Finally, the question is often asked by students, “how does Dialog compare to the World Wide Web?” Below is a table with comparisons between the Web and Dialog including the elements discussed in this paper.

## Comparing the Web to Dialog

	<b>World Wide Web</b>	<b>Dialog</b>
Data Universe	Depends on the search engine used. The invisible, or portion of the Web that is not crawled, is a very large percentage of the World Wide Web.	Dialog products offer over 900 databases containing more than 12 terabytes of data from many of the world's leading professional publications.
Copyright	Documents on the Web are protected by copyright, but it is easy to violate copyright regulations by copying Web pages and files and emailing them or posting on other Web pages. This type of copyright infringement is hard to control in an organization, yet could prove costly if violations are noted by officials.	Copyright protection can be ensured for your organization for most documents available on Dialog through use of the Electronic Redistribution & Archiving (ERA) feature.
Fulltext	When a fulltext document is available on the surface Web, it is difficult to establish whether or not it is really the complete document—and whether this is a legitimate or illicit copy of the document.	Dialog databases contain millions of fulltext documents and the e-Journal Linking feature available through various Dialog services provides access to more than 11,000 fulltext journals.
Abstracts	Web structure has no provision for abstracts and typically no abstract exists for most Web pages—and is not indexed as such.	Dialog consistently includes abstracts in addition to fulltext documents as an aid to quick evaluation of search results.
Indexing	No standard indexing is used for Web pages. Search engines attempt to infer indexing terms from the content and structure of the Web page.	Each data source is indexed using the information producer's uniform standards.
Search language	Search syntax is limited, varies from search engine to search engine, and changes over time.	Dialog search language is an extremely powerful tool to find information. It provides a uniform interface to each of the databases. Techniques such as truncation, word proximity, etc. typically produce more relevant "hits."
Ranking	Ranking for a particular search is established by an ever-changing weighting of page content, page popularity, and pay-per-ranking considerations.	TARGET command can be used to rank 50 most relevant documents.
Authority	Web sites may be created by individuals or organizations. It is often difficult to establish the credibility of relatively unknown individuals or organizations, especially since the author or source behind a Web page is frequently not identified. Web pages generated by universities are often authoritative sources of data, though the credibility of the source should still be considered (i.e. personal student Web page versus official university departmental research).	The source and/or author of each professional publication on Dialog is routinely identified and indexed. Credibility can often be established by searching on the source/author, or doing a cited reference search in certain databases.
Bias	Many Web sites are created to advocate a particular position. Care must be taken to establish the purpose or perspective of each Web page.	Each professional publication may have its own bias. Dialog attempts to act as a fair and impartial collator of published information.
Exclusionary Bias	Sometimes what a Web page excludes establishes more bias than what it includes.	Dialog strives to maintain access to everything that an information producer publishes.
Data Freshness	Web pages are updated irregularly. Often the date of creation and/or last revision is unknown.	Each Dialog database is updated on a published schedule. The date of every record is maintained.
Privacy	Depends on the privacy policy of each search engine. Searching for sensitive terms may be risky.	Privacy of customer searches is ensured. Secure Socket Layers (https://) are available for accessing many Dialog Web products.

## **BIBLIOGRAPHY**

- Alexander, Jan and Marsha Ann Tate. "Evaluating Web Resources." (1999) Wolfgram Memorial Library, Widener University. 14 Nov 2003. <http://www2.widener.edu/Wolfgram-Memorial-Library/webevaluation/webeval.htm>
- Barker, Joe. "Evaluating Web Pages: Experience Why It's Important." (22 Jan 2003) Teaching Library Internet Workshops, University of California at Berkeley. 14 Nov 2003. <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/EvaluateWhy.html>.
- Barker, Joe. "Evaluating Web Pages: Techniques to Apply & Questions to Ask" (12 Sept 2003) Teaching Library Internet Workshops, University of California at Berkeley. 14 Nov 2003. <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/Evaluate.html>.
- Barker, Joe. "Invisible Web: What it Is, Why it exists, How to find it, and Its inherent ambiguity." (28 Aug 2003) Teaching Library Internet Workshops, University of California at Berkeley. 14 Nov 2003. <http://www.lib.berkeley.edu/TeachingLib/Guides/Internet/InvisibleWeb.html>
- Beck, Susan E. "The Good, the Bad, & the Ugly or Why It's a Good Idea to Evaluate Web Resources." (30 Oct 2003) New Mexico State University Library. 14 Nov 2003. <http://lib.nmsu.edu/instruction/eval.html>.
- Dudley Knox Library. "Web Finding Tools." (30 Oct 2003) Naval Postgraduate School. 14 Nov 2003 <http://library.nps.navy.mil/home/netsearch.htm>.
- Henderson, John R. "ICYou See: T is for Thinking...A Guide to Critical Thinking About What You See on the Web." (21 Nov 2003) Ithaca College Library. 11 Nov 2003. <http://www.ithaca.edu/library/Training/hott.html>
- Kirk, Elizabeth E. "Evaluating Information Found on the Internet." (5 June 2002) The Sheridan Libraries, John Hopkins University. 14 Nov 2003. <http://www.library.jhu.edu/elp/useit/evaluate>.
- Marlatt, Greta E. "Internet Primer2." Naval Postgraduate Library. 14 Nov 2003. <http://library.nps.navy.mil/home/InternetPrimer2.ppt>.
- Notess, Greg R. *Search Engine Showdown*. 14 Nov 2003. <http://searchengineshowdown.com>.
- Schlein, Alan M. *Find It Online: The Complete Guide to Online Research*, 3<sup>rd</sup> ed. Tempe, AZ: Facts On Demand Press, 2003.
- Search Engine Ethics. "Link Popularity Analysis." (23 Aug 2003) L'il Engine. 14 Nov 2003. <http://www.lilengine.com/articles/link-popularity/118/3>
- Sherman, Chris and Gary Price. *The Invisible Web Directory*. 14 Nov 2003. <http://www.invisible-web.net>.
- Sherman, Chris and Gary Price. *The Invisible Web: Uncovering Information Sources Search Engines Can't See*. Medford, NJ: Information Today, Inc., 2001.
- Sullivan, Danny. "Buying Your Way In." (20 May 2003) Search Engine Watch. 14 Nov 2003. <http://www.searchenginewatch.com/Webmasters/article.php/2167941>

Sullivan, Danny. "How Search Engines Rank Web Pages." (31 July 2003) Search Engine Watch. 14 Nov 2003. <http://www.searchenginewatch.com/Webmasters/article.php/2167961>

Sullivan, Danny, Ed. *Search Engine Watch*. 14 Nov 2003. <http://www.searchenginewatch.com>.

Dialog has offices around the world. For more information on our full range of products and services, call one of our main offices below. Or, visit [www.dialog.com](http://www.dialog.com).

**Corporate Headquarters**

11000 Regency Parkway, Suite 10  
Cary, NC 27511  
United States  
+1 (919) 462 8600  
(800) 3 DIALOG

**Europe, Middle East, Africa**

Palace House  
3 Cathedral Street  
London SE1 9DE  
United Kingdom  
+44 (20) 7940 6900

**Asia Pacific**

20/F Sunning Plaza  
10 Hysan Avenue  
Causeway Bay  
Hong Kong  
+(852) 2530 5778