The Online World resources handbook

Chapter 10:
Finding a needle in a bottle of hay

There is little doubt that the databases of the online world contain nearly everything needed to complete a major research project, fuel an information-needy business, or just help get the school homework done.

Online research is faster, provides more depth and is cross-referenced to help researchers locate obscure resources. It makes you an "instant expert" on a subject matter. The main problem is learning how to get a confident grip on the searching process.

Prepare by clipping

Experienced users regularly "clip" news from online services, and store selected parts of what they get on their personal computers' hard disks. They use powerful tools to search their data, and know how to use the information in other applications. (More about clipping in Chapter 11.)

Regular clipping of news is highly recommended. It is often quicker and easier to search your own databases, than to search online. Your data is a subset of previous

searches. Therefore, the stories on your disk are likely to have a high degree of relevancy.

There are many good programs for personal computers that let you search your personal data for information. See Chapter 14 for ideas.

While secondary research can never replace primary information gathering, it often satisfies most information needs related to any task or project. Besides, it points in the direction of primary sources from where more in- depth information may be elicited.

When your personal database fails to deliver

Regular "clipping" can help you build a powerful personal database, but it will never satisfy all information needs. Occasionally, you must go online for additional facts.

When this happens, you may feel like Don Quixote, as he was looking "for a needle in a bottle of hay." The large number of offerings is bewildering. To succeed, you'll need a sound search strategy.

Your first task is to locate **useful sources** of information. The next, to decide how best to find that specific piece of information online. You must **plan** your search.

Although one source of information, like an online database, is supposed to cover your area of interest, it may still be unable to give you what you want. Let me explain with an example:

> You're tracking a company called IBM (International Business Machines). Your first inclination is to visit forums and clubs concerned with products delivered by this company. There, you plan to search message bases and file libraries. The search term IBM will probably give so many hits that you almost drown. To find anything of interest in these forums, your search terms must be very specific. General news providers, like Associated Press, may be a better alternative. Usually, they just publish one or two stories on IBM per week. Don't expect to learn about details that are not of interest to the public. AP's stories may be too general for you. Maybe you'll be more content with industry insiders' expert views, as provided by the Brainwave for NewsNet newsletters OUTLOOK ON IBM, or THE REPORT ON IBM.

The level of details in a given story depends in part on the news providers' readers, and the nature of the source. The amount of "noise" (the level of irrelevancy) also varies. In most public forums, expect to wade through many uninteresting messages before finding things of interest.

Try the following strategy:

Step 1:	Locate sources that provide relevant information,
	Selecting sources is half the battle in making a good search! You probably won't find what you need if you're not looking in the right place.
Step 2:	Check if the information from these sources is at a satisfactory level of details, and that the volume is acceptable (not too much, nor too little).
Step	Study the service's search commands and procedures,

3:         PLAN, and then SEARCH.

Locating interesting sources

Step 1 is not an easy one. There is such an abundance of directory services and pointers.

My **current favorite** is Google. By December 2001, it claimed to index over 3 billion documents, i.e. Over 2 billion web pages (html, pdf, other formats), 700 million Usenet posts and 300 million images. Its Usenet postings date back to May, 1981.

Compare notes with Searchenginewatch.com, and Searchengineshowdown.com. They regularly gauges the comparative sizes of each major search engine's index.

Google rates sites based on who links to whom. Ranking depends on the number of links to a site and its rating position, thus giving a type of peer review of the Web itself. It puts search terms in context by displaying an excerpt of the text that matches a particular query with the search terms included in bold. Several languages.

If you worry about search engines selling data collected from your searches to third parties, try the Google-powered Topclick search engine. They use no cookies, no banner ads, and strives to protect your privacy.

Google lets you use a plus sign (+) to include words or a minus sign (−) to exclude words in the search, as in *+online +world −computer*. This search will only return hits containing the words "online" and "world" but not "computer".

You can limit your search to .pdf files by using the syntax filetype:pdf. A Boolean "OR" is available with Google. For it to function, capitalize the OR. Note that Google only makes searchable the first 110 k of a page. Long documents may have substantial content invisible to Google.

One nice feature in this age of dead links is that Google saves a copy of the indexed page. So, if you are unable to get to a desired page, select the cache copy.

It's only worth using Google if you bear in mind the sort of material which might be posted in your subject area. Since anyone can publish almost anything on the Web, pages vary – from personal pages set up by any student who has Internet access, to those set up academic or research institutions, those set up by not-for-profit organizations, and those from commercial organizations.

Warning: The largest search engines index less than 1/10th of the web! *"Two scientists from the NEC Research Institute in Princeton carried out a study on the Net's loudest search engines and found that not only do they not index the best part of the Net but they are most likely to index commercial over educational, US over European and popular over relatively unknown,"* reported Nua Internet Surveys in July, 1999.

Search engines are not indexing their databases daily. In early 2002, Google refreshed their database once per month! A web page made available on the web on Thursday might not appear in any search engine for weeks.

Simply because many pages from a site are available does not mean that the engine makes every page of an entire site searchable. This is a good

reason for not relying on one search engine only.

Fast Search and Alta Vista are second choices. Sometimes, I play Google against Alta Vista or Fast Search for maximum result. If I want a query to contain **a string from a Web address**, Alta Vista would be my first choice. In other cases, network access speed will decide. If getting to one of them takes to long, I go to the other. Arnoldit.com

For international search engines, check Arnoldit.com.

Disabled Internet surfers may want to search these offerings through SETI-search.com. This search service is particularly interesting foor those who are blind, or have very low vision as it works smoothly with their assistive technology devices.

Special

FindSame allows users to search for documents using large pieces of text rather than keywords. It treats your search query as an entire document and returns a list of "documents that contain any fragment of that document that is longer than a certain length. That length is about one line of text." Alternatively, users can enter the URL of a document and FindSame will return pages that contain at least a few sentences that appear on that page.

Meta-searching

Meta-search agents let you search several search engines in one operation. For example, Super Searches searches major search engines like Alta Vista, Excite, Galaxy, HotBot, Lycos, Web Crawler, Yahoo, WWW Yellow pages, Meta crawler, Google Groups, Aliweb, Hotbot, Lycos, and more. Also, try Dogpile, Highway61

One word of warning: The meta-search agents treat the product of search engines as data: changing it, organizing it, and making it simpler to use for the consumer, without understanding that this information is more like a publication than raw data.

Usually, these services do not support Boolean, temporal, or proximity operators, and therefore give very poor precision. Set building is not possible. Most well-known meta-engines send a query to many entirely "pay for placement" engines.

Searching a topic area

Narrowing a search down to a specific topic area can be a challenge with the general search engines. Sometimes, you may be better off using a more targeted search service.

There are many services linking you to topic area search engines. Example: SEARCH.COM links you to search services within areas like Arts, Automotive, Business, Computers, Directories, Education, Employment, Entertainment, Finance, Government, Games, Health, Housing, Legal, Lifestyle, News, People, Politics, Reference, Science, Shopping, Sports, Travel, Usenet, and Web.

Langenberg Search is a gateway to some of the most popular search engines for a variety of subjects grouped under : Acronym, Area Codes, Books&Pubs, BusinessFinder, Cooking, Dictionary, Encyclopedia, Entertainment, Government, Jobs, Maps, Medicine, Metasearch, Misc, Money&Stocks, News&Sports, PersonFinder, Religion, SearchEngines, Shipping, Translation, Travel, Usenet,

Weather, Zip Codes.

SearchEngineGuide.COM offered links to 2341 search engines sorted by area in December, 1999. The BIG Search Engine Index may also be worth your visit.

Some other interesting offerings:

- Today's news
- Acronym Finder (over 225,000 acronyms, abbreviations and definitions – Jan. 2002)
- Agricultural Search Engine (English, French)
- Archives of yesterday's news
- Airport Search Engine
- Animals
- Asia (Search Asian Studies WWW VL Web Space)
- Browser applications, all the small application programs that run in web browsers.
- Computer companies, hardware, software, peripherals.
- Computers, etc.
- Country music
- Clip art, icons, background images, animations, sound clips
- Currency exchange
- Education
- The Internet guide to engineering, mathematics and computing (EEVL)
- English (to find the meaning of this very special English word)
- Films (Online Short Films)
- Financial-only content
- Food recipes
- Frequently Asked Questions (FAQs)
- Games on the Net.
- Games II
- 'Geo' Industry Search Engine (GIS, GPS)
- Geography – blank map outlines
- Glossaries – Arts & Culture, Business , Careers & Employment , Computers & Internet , Economy & Finance , Education , Entertainment , Family & Relationships , Government, Politics & Military , Health, Medicine & Fitness , Humanities & Social Sciences , Law and Justice , Lifestyle , Media , News & Weather , Reference , Science , Sports & Recreation , Technology , Transport , World, Regions, Countries & Travel
- Health
- Health
- Historical web pages. – "The Internet Archive's" WayBack Machine lets you surf the net as it was. Also, includes Usenet postings, archival movies, Arpanet historical documentation, and more.
- Html, dhtml, Perl, Java: for Web developers and programmers.
    - Sourcebank – Programming resources on the internet, source code for Java, C, C++, research papers and online magazine articles.
- Image Search (Google)
- Locating mailing lists with interesting discussions
    - Topica
    - Liszt
    - L-Soft (the LISTSERV vendor)
    - Mailbase
    - Onelist
    - Neosoft
    - New Lists
    - Catalog.com

- o Findmail
- Mail-order catalogs
- Maps
- Marketing: For Online Advertisers, Marketers, and E-Commerce
- Multimedia files such as movie scenes, pictures, music clips, concerts, sporting events: Lycos, Scour.net, arribavista.com, Altavista.
- Music (MP3 format)
- Mythology and folklore
- Non-Profit Organizations
- .pdf documents documents (Adobe)
- Old newsreels: more than 3,500 hours (1896-1970).
- Postal codes around the world
- Sample sounds and sound effects.
- Scientific information. | Scirus
- Software - shareware and public domain: Lycos, Tucows
- Time related links . An extensive collection.
- Weblog searches
- Web searches. In addition, searches in million of articles from 5,400 premium sources, such as books, magazines, databases, and newswires not available elsewhere.
- WebData.com. Access to specialized online databases.

Finally, check AltaVista's Search Guides for guidance about how to search for some types of information. W3engine is a search engine's search engine. It helps you find search engine sites; meta search sites; index directories; specialty engines; yellow pages; and more. You can also locate search engines by country. InfiniSource has a portal with links to specialized search engines, as has Universitet Leiden in Holland.

Searching for non-US information

No search engine indexes the whole Web, and most US based services tend to be best at US contents. US services focusing on other geographical areas tend to miss local organizations having registered .com, .org, or other global addresses.

For contents in other geographical areas, you may be better served by engines specialized on these areas. To locate such engines, try Search Engines Worldwide (with frames). 850 search engines and directories from 130 nations (April 2000), listed alphabetically by country.

Some examples:

Africa

Europe: Euroferret

India-Related Search Engines: Utexas.edu, KHOJ

Israel

Middle East

Russia

Scandinavia

South Africa

United Kingdom

For links to search services in other countries, try **Search Engines Worldwide**, **Search Engine Colossus**, and **Country Specific Search Engines**. **The Financial Times Global Archive** is another interesting offering. It has over 10 million articles from 2000 publications. Their news database is updated on a 24 hours / 7 days basis from selected international publishers and agencies. Search the five year archive of the Financial Times Newspaper as well as archives of European, Asian and American business sources.

In the **comp.infosystems.search** newsgroup, discussion is focused on web searching: "*Discussion about the different aspects, ramifications and use of search engines and associated technology.*"

Non-English language searches

There are major structural differences between languages. An indexing system built for English text may therefore not be suitable for a text written in the language you're searching, and in particular if the other language uses special fonts. Using special purpose search engines may be the way to go in such cases. Some options:

- **Arabic**
- **Chinese**
- French (**http://lokace.iplus.fr/** and **http://www.ecila.fr/**)
- German (**http://www.aladin.de/** and **http://www.dino-online.de/suche.html**)
- **Italian**
- **Japanese**
- **Spanish** (**Spanish language search engines**).

Another problem using the English language search systems is that you don't just have to understand English to get the most out of them, you'll have to understand English well.

Searching Usenet

After searching the Web, my next step is usually **Google Groups**, a large indexed database of archived Usenet news dating back to 1981. We're talking about over 700 million messsages here (2001)! It is by far the most complete collection of Usenet articles ever assembled and a fascinating first-hand historical account.

You can use the service for research, or to locate interesting newsgroups worth your subscription.

Google Groups' powerful filter lets you limit what records will be searched by a query. A search may be limited by date, author, newsgroup name (using wildcards, or range operators), wildcards (compan* matches companies, company, and much more.

The order of the records in the hit list may be set to reflects how often the words you're searching for appear, as well as the importance you have given the posting date. This scoring will give you the records that best match your search at the top of the list.

Once you have found an interesting message in a hitlist, you can retrieve the thread as it appears on the screen.

InfoSeek lets you search many Internet newsgroups, news and business information from real- time newswires, publications, broadcast programs, financial and government databases, World Wide Web pages, mailing list archives, and technical support information (including over a year of Computer Select database of

the full-text and abstracts of about 100 computer magazines).

     Queries can be entered as plain English, or by just entering key words and phrases. There is a Japanese language version at http://japan.infoseek.com/.

Searching Mailing lists and Web forums

BoardReader claims to let you "search 732,456 forums and message boards from the Web, information that most search engines do not retrieve." The forums and boards are sorted in these top-level topic groups: Art & Antiques, Automobile, Boating, Computers, Console Games, Hobbies, Home & Garden, Motorcycles & ATV's, Health, Entertainment & Music, Outdoor Recreation, Sports, Social, Travel.

     Reference.COM (Chapter 11) indexes messages posted to several mailing lists and Web forum. This includes Kidlink's announcement lists.

     Several mailing lists let you search their archives of postings through the Web. For example, all postings to the TOW mailing list since 1993 are searchable. Hits can be filtered by strings found on the subject line, strings in the author's email address, or by giving a date range.

     Microsoft lets you search several of their mailing lists, like those on ATL, ActiveX, Active Server Pages Scripting, Authenticode, CIFS, Client Scripting, Cryptographic API, Distributed COM-Based Code, Internet Explorer Html.

     Some other mailing list archive sites:

- American.edu
- Anet (via rutgers.edu)
- Buffalo.edu
- Asu.edu
- Emory.edu
- Gmd.de
- Hea.ie – many lists focusing on Irish, Celtic and Gaelic culture and languages.
- Indiana.edu
- Lsoft
- Nih.gov – mainly medicine and health related lists
- Nodak.edu – carries all public Kidlink lists
- Polito.it
- Psu.edu
- Spc.edu
- Stjohns.edu
- Temple.edu
- Uta.edu
- Yorku.edu

Catalist is the official catalog of LISTSERV mailing lists. This site lets you search for mailing lists of interest. It guides you to their web archive interface, if available. The LISTSERV web archive interface allows you to search the list's archive, and browse postings chronologically.

Searching specialized databases

If you are looking for more specialized databases, try **The Internet Sleuth**. It links to over 3,000 searchable databases on the Internet on a wide variety of subjects.

     Sleuth's categories include: Agriculture, Economics, Internet, Regional, Education, Legal, Sciences, Astronomy, Employment, Literature, Shopping, Aviation,

Engineering, Mathematics, Social Sciences, Biology, Physics, Entertainment, Medicine, Software, BioSciences, Environment, Arts, Music, Sports, Business, Finance, News, Technology, Business Directories, Food & Drink, People, Trade & Industry, Chemistry, Genealogy, Travel, Commercial Databases, Government, Politics, Usenet News, Companies, Health, Computer Related, Recreation, Veterinary, Humanities, Reference, Web Search Engines.

Database Central links to over 4,000 database resources. The resources vary widely, from software, shareware, and middleware to tips, tutorials, and white papers to books, magazines, and discussion forums. You may browse by category or use a keyword(s) search engine.

The "Deep" Web

Then, there's the "deep" web, also called the Invisible Web. These are the terabytes of information available in digital form through hidden databases that cannot be seen or searched directly by most Web search engines. They include databases, archived material, and interactive tools such as calculators and dictionaries. About 95 percent of this information is available to the public and is not subject to subscription fees.

In early 2002, experts estimated that the "surface Web" contained 1 billion to 2 billion documents, while the "deep Web" could contain as many as 550 billion. Put another way, the surface Web contained about 19 terabytes of information, while the deep Web contained about 7,500 terabytes.

A terabyte is a measure of data storage. One terabyte is the equivalent of about 1,600 CDs or 1,000 gigabytes.

Reasons for the invisibility of the deep Web include that search engines cannot find them (they identify "static" pages rather than "dynamic" pages used by large databases), have made a conscious decision not to index them, or that the information is stored in a format that search engines are unable to index. For example, search engines can record a database's address, but can tell you nothing about the books, magazines or other documents it contains.

Links to some Invisible Web resources:

- The Direct Search page
- InvisibleWeb.Com
- CompletePlanet
- The Big Hub

Your "last" resort

If your success is still meagre, consider asking other onliners for advice. Actually, as this may often be a fast way to interesting sources, you may even want to put it higher on your list.

When looking for information about agriculture and fisheries, visit forums and conferences about related topics. Ask members what they are using.

If you want information about computers or electronics, ask in such conferences.

When you do not know where to start your search, ask others!
Their know-how is usually the quickest way to the sources.

Google Groups will help you locate relevant newsgroups for your questions. To find interesting mailing lists, check Topica, or its subsidiary, the Liszt Index of Electronic

Mailing Lists. Liszt can also be searched by email.

The Liszt Index lets you enter any word or phrase to search their directory of over 90,095 listserv, listproc, majordomo and independently managed mailing lists (as of March, 1999). It will not allow you to search the message bases, but it sure will help you locate potentially interesting discussions.

The Listserv home page lets you sort LISTSERV discussion groups by 1st letter of list name, by country, by server name, and more. The description pages of the individual discussion groups, however, is not to much help. Try Publicly Accessible Mailing Lists for an alternative.

Also, there are over 250,000 Web based discussion forums (June 26, 1998). By November 25, 1996, the number was just 37,000. Search for discussions of interest at http://www.forumone.com/.

Note: There is much free information on the Internet, but be prepared to pay for current and relevant information. Your payment is for **filtering**, **sorting**, and **emphasizing** of what matters to you.

Read the user manuals

Some online services let you retrieve their user information manuals by modem for free. Others send them to all users, while some charge extra for them. If they do, buy! They're worth their weight in gold.

User manuals from commercial services like CompuServe make good reading. The latter two also publish monthly magazines filled with search tips, information about new sources, user experiences, and more.

Whenever it is possible to retrieve these help texts in electronic form, consider doing that. It is often faster to search a help file on your disk, than to browse through a book.

Monitor the offerings

Professional information searchers watch the activity in the online world. They subscribe to announcements about new offerings, regularly search databases for new sources of information, and read about new services.

On most online services, you can search databases of available offerings, and a section with advertisements about their own 'superiorities'. Keep an eye on what is being posted there.

There's an announcement-only service called NET-HAPPENINGS. It is a favorite for monitoring Internet's offerings.

The service distributes announcements about tools, conferences, calls for papers, news items, new mailing lists, electronic newsletters like EDUPAGE, and more.

Net-happenings is also at comp.internet.net-happenings. Their archives can be searched at http://scout.cs.wisc.edu/index.html.

NEW-LIST regularly distribute notices about new discussion lists (conferences). You can search the postings. Also in the bit.listserv.new-list newsgroup.

"Seidman's Online Insider" is an informative newsletter. You can subscribe to have it delivered weekly to your mailbox. Subscription information at http://www.clark.net/pub/robert/listserv.html.

Heriot-Watt University Library (Scotland) publishes the free INTERNET RESOURCES Newsletter. Emphasis is on Engineering, Science, and Social Science related sources in the United Kingdom. You can subscribe to have an alerting message, plus the table of contents sent via email, each time a new issue appears.

The Usenet newsgroup alt.internet.services focuses on information about services available on the Internet. Services for discussion include:

- things you can telnet to (weather, library catalogs, databases, and more),
- things you can FTP (like pictures, sounds, programs, data)
- clients/servers (like MUDs, IRC, Archie)

Every second week, a list of Internet services called the "Special Internet Connections list" is posted to this newsgroup. It includes everything from where to retrieve pictures from space by FTP, how to find agricultural information, public UNIX, online directories and books, you name it.

On The Well, read the "News from Around Well Conferences" topic to learn about developments.

The LINK-UP magazine is an interesting paper source. In North America, contact Learned Information Inc., 143 Old Mariton Pike, Medford, NJ 08055- 8707, U.S.A. In Europe: Learned Information (Europe) Ltd., Woodside, Hinskey Hill, Oxford OX1 5AU, England.

Two monthly magazines, Information World Review and FULLTEXT SOURCES ONLINE from BiblioData Inc. (U.S.A.), are also available through Learned Information. (BiblioData, P.O. Box 61, Needham Heights, MA 02194, U.S.A.) Learned Information's "Learned InfoNet" is at http://info.learned.co.uk/

More sources about sources

Scott Yanoff updates an interesting, selected list of Internet resources twice per month.

John December's "Information Sources: the Internet and Computer-Mediated Communication" has pointers to information describing the Internet, computer networks, and issues related to computer-mediated communication. It lists Internet texts for new users, comprehensive Internet guides, and specialized and technical information.

The Gale Directory of Databases contains detailed descriptions of over 11,500 publicly available databases accessible through an online vendor or batch processor or for purchase on CD-ROM, diskette, or magnetic tape, or as a handheld product (Feb, 1999). It is a comprehensive guide to the electronic database industry worldwide. They also offer listings of database producers and vendors.

For lists of electronic journals about the Internet ("E-zines" or "Ejournals"), click at http://www.edoc.com/ejournal/

Several electronic journals and newsletters are available through the Internet, covering fields from literature to molecular biology. For a large list, try http://www.meer.net/~johnl/e-zine-list/.

The NEWSLTR list distributes various network newsletters. Offerings include: Edupage, Hitek, HPC, Infosys, IAT Inforbit, and many more.

The Argus Clearinghouse offers over 1,000 topical guides to the Internet's information resources. The guides are created by librarians and other information professionals, and cover a diverse range of topics, from Theatre, Law, and Chemistry to Midwifery.

Interested in CD-ROM? The database at http://www.microinfo.co.uk/ offers details about thousands of information products and services – mainly CD- ROMs. Products are classified in 27 topics ranging from agriculture and food to theology.

Search: [   ] [go]

Illustration by Anne-Tove Vestfossen