Searching for Chemical Information on the WWW

Lou Coury, Ph.D.
Director of Research
Bioanalytical Systems, Inc.
2701 Kent Avenue
West Lafayette, IN
47906-1382

coury@bioanalytical.com

The amount of chemical information available for free on the World Wide Web (WWW) has experienced an explosive growth over the past few years. In the early days of the web, search engines were inefficient and the material they uncovered during scientific searches often tended to be home pages of research groups presenting information that may not have been subjected to the peer review process. By contrast, it is now possible for anyone with an Internet connection and a web browser to gain free access to the titles, authors, citations and abstracts for hundreds of thousands of reviewed papers published in the primary literature. This article will discuss a few sites that are accessible without charge used by scientists here at BAS in their daily work of providing scientific services, contract research, new equipment and software to the scientific community.

Search Engines

The most important tools for finding information in the vast expanse of cyberspace are search engines. These sites provide increasingly sophisticated mechanisms for sorting and selecting web resources based on keywords entered by the user. The most reliable engine we have found is called Google^{SM} (URL: www.google.com). Our experience has been that Google^{SM} turns up relevant information within the first few results displayed a large percentage of the time. Through a combination of ranking the importance of individual web sites based on the number of times they are linked to other pages, along with evaluation of search results based on the degree of proximity of the keywords entered, high quality results are generated.

There are several other welcome features of the Google^{SM} site. When results are displayed, short excerpts from the web pages found are displayed by each result, with the user's keywords highlighted. This allows for quick evaluation of multiple "hits" without the need to load the entire page. This is extremely helpful for those who access the web via a modem and thus may not have the time to download each search result to review it.

Second, many of the web pages indexed by Google^{SM} are stored (cached) at their site. Because of this feature, it is often possible to read a web page from the Google^{SM} cache that would otherwise be inaccessible due to a server being down or even due to the original page being corrupted or deleted after it was originally accessed by Google^{SM}. Finally, the Google^{SM} site has a button named "I'm Feeling Lucky^{TM}." By clicking on this, users are linked directly to the first (highest ranked) site located using the search parameters entered. Google^{SM} is so accurate when a properly posed search is conducted that there is often no need to even view the search results!

There are, of course, other search engines that can be helpful at times. These include earlier entrants onto the web search engine playing field like Yahoo!®, AltaVista®, and Lycos® (see ***T1***). Rather than conducting separate searches using each of these engines, it saves considerable time to use a *metasearch engine*. These tools allow the user to type one or more keywords into a dialog box, and then execute a search. The meta engines pass the search terms along to several different search engines, and the user is presented with a list of the most relevant "hits" at each of these external sites. Dogpile (www.dogpile.com) and Metacrawler (www.metacrawler.com) are two of our favorites. Give them a try the next time you are researching a subject on the web!

Literature Searches

Search engines provide a quick way of turning up information from a wide variety of sources. Commercial sites, university information pages and amateur home pages are examples of the types of pages likely to be located with a generalized web engine. The quality of the information provided, however, is usually un-

T1

Uniform Resource Locators (URLs) of Web Sites discussed in this Article

Category	Name	URL
Search Engines	Google ^{SM}	http://www.google.com
	Yahoo!®	http://www.yahoo.com
	AltaVista®	http://www.altavista.com
	Lycos®	http://www.lycos.com
MetaSearch Engines	Dogpile	http://www.dogpile.com
	Metacrawler	http://www.metacrawler.com
Literature Search Resources	PubMed	http://www.ncbi.nlm.nih.gov/PubMed
	PubSCIENCE	http://pubsci.osti.gov
	Beilstein Abstracts	*access via* http://chemweb.com
	UnCover	http://uncweb.carl.org
	The Library of Congress	http://catalog.loc.gov
	Library of Congress Z39.50 Gateway	http://lcweb.loc.gov/z3950
Other Resources	ESTIR	http://electrochem.cwru.edu/estir
	NIST WebBook	http://webbook.nist.gov
	IBM Intellectual Property Network	http://www.patents.ibm.com
	US Patent & Trademark Office	http://www.uspto.gov
	Code of Federal Regulations	http://www.access.gpo.gov/nara/cfr
	OSHA Analytical Methods	http://www.osha-slc.gov/dts/sltc/methods
	MSDS Links	http://siri.uvm.edu
	General Safety Links	http://hazard.com/links.html
	NFPA Labels	http://www.orcbs.msu.edu/chemical/nfpa/nfpa.html
	ToxNet	http://toxnet.nlm.nih.gov

known and hence often cannot be cited with confidence by scientists preparing manuscripts and grant proposals for review by fellow scientists. Fortunately, there are ways to find citations (and sometimes full abstracts) of papers published in peer reviewed journals.

PubMed. Perhaps the most useful and extensive resource available is the public access version of MedLine called *PubMed* (www.ncbi.nlm.nih.gov/PubMed) developed by the National Center for Biotechnology Information (NCBI) at the National Library of Medicine. This database contains approximately 9 million records dating back to 1966, comprised of bibliographic citations and author abstracts from approximately 3,900 current biomedical journals published in the United States and 70 foreign countries. Full abstracts are accessible for the most recent citations, and titles and authors are available for many older papers.

Some of the useful features of this resource include the ability to use wildcards when conducting searches. For example, to obtain abstracts of papers that deal with either "electrochemistry" or "electrochemical" studies of the enzyme "sulfite oxidase," the search terms entered might be *electrochem** "*sulfite oxidase*". The wildcard character * instructs the search algorithm to find all citations containing the stem "electrochem." The use of quotation marks around the word pair "sulfite oxidase" is an example of proximity searching. The quotation marks link "sulfite" and "oxidase" together to help limit to citations discussing this enzyme and not, for example, sodium sulfite and glucose oxidase. Note also that if the example search is run without the quotation marks, citations mentioning "sulphite oxidase" are also found. Thus, the quotation marks also limit the search to references using the American spelling of sulfite. PubMed is a fabulous resource, and a commendable way for the U.S. government to use tax dollars to serve the public interest. In addition to its obvious utility for research scientists, this database allows the general public to read abstracts from the medical and clinical literature, and thus educate themselves about health-related issues. Bravo!

PubSCIENCE. On October 1st, 1999, the Department of Energy (DOE) teamed with the U.S. Government Printing Office (GPO) to launch a resource modeled after PubMed that is focused on the physical sciences. The new database is named PubSCIENCE and it is accessible on the web at the URL: http://pubsci.osti.gov. This database currently includes about 1,000 journal titles, and gives users access to abstracts of the papers indexed there. For example, journals published by the American Physical Society and The Electrochemical Society are included. Paradoxically, the American Chemical Society (ACS) does not make abstracts from its journals available, although the Royal Society of Chemistry and the National Research Council of Canada both participate. Thus, US tax dollars are being spent by the DOE and GPO to make information published in British and Canadian chemistry journals (but not ACS journals…) freely available worldwide. Kudos to our colleagues in the UK and Canada!

Beilstein Abstracts. Fortunately, several ways to gain free access to titles and/or abstracts of papers published in ACS journals already exist. My personal favorite is the port to *Beilstein Abstracts* provided by a web site called *ChemWeb.com*. This resource is owned by the European publishing giant, Elsevier Science, and is accessible at the WWW URL: chemweb.com. New users will need to register, but membership is free, and provides a whole host of benefits in addition to Beilstein access (*e.g.*, job listings, a conference diary, and limited-time *full* access to a number of Elsevier Science journals as well as contents searching abilities for all Elsevier Science journals).

Once logged in as a registered user, it is only necessary to follow the links and pull up the Beilstein search form. Terms may then be entered into dialog boxes set to search the Authors, Title, Abstract or Abstract + Title fields of records in the database. Note that the database in question is *not* the famous Beilstein Handbuch, but rather a separate resource containing the full abstracts of papers published in over 140 of journals in organic and related areas of chemistry, dating back to 1980. This ChemWeb.com service is a much appreciated alternative to performing expensive Chemical Abstracts Services searches, which are beyond the budgets of many smaller companies and academic research groups.

UnCover. Another way to obtain free citations, authors and titles (but not abstracts) of papers published in ACS journals is to use the service known as *UnCover*. The URL is: http://uncweb.carl.org. UnCover indexes 18,000 multidisciplinary journals and contains brief descriptive information for over 8,800,000 articles which have appeared since the Fall of 1988. The UnCover search mechanism supports both wildcards (*) and Boolean operators (*and, or*). However, unlike PubSCIENCE and PubMed, full abstracts are not available for free. However, users who register and provide a credit card number can order full articles and have them faxed to their homes or offices. One disadvantage of *UnCover* is the relatively high frequency of spelling errors in the database which we have noticed in the course of using it over the past few years. This illustrates an important point that applies to all of the resources discussed in this article: the usefulness of a database is dictated by the quality (fidelity) of the information stored there. For example, using "fluorescence" as a search term will usually not find references in which "flourescence" has been typed during record transcription!

The Library of Congress. Finally, an easy way to find the titles and/or authors of books dating back as far as 1898, is to access the Library of Congress Online Catalog (catalog.loc.gov). This free database contains approximately 12 million records describing books, computer files, manuscripts, maps, and art work. Searches may be conducted based on subject, author name, serial title, or call number. A feature called Guided Keyword Search allows for fairly complicated searches to be executed and supports both wildcards (in this case ? rather than *) and Boolean operators. Also linked from this site is a service called the Z39.50 Gateway. This service allows for searches to be conducted at over 300 other libraries from within the Library of Congress web pages without the need to know the search syntax that is used by the other systems. It is accessed at the URL: lcweb.loc.gov/z3950.

Other Resources

The Electrochemical Science and Technology Information Resource (ESTIR). All electrochemists should be aware of a unique and tremendously useful site on the Internet called *ESTIR* (electrochem.cwru.edu/estir). This resource is hosted by the Ernest B. Yeager Center for Electrochemical Sciences and the Chemical Engineering Department of Case Western Reserve University, and is maintained by Dr. Zoltan Nagy of Argonne National Labs. Dr. Nagy has organized the site into subsections listing electrochemical information such as WWW sites, newsgroups and mailing lists related to electrochemistry; public domain bibliographies, software and data banks; an electrochemical dictionary; a list of suppliers of electrochemical equipment, accessories and consulting services; a list of over 2000 review articles and chapters on various topics; a list of all known books ever published on the topic of electrochemistry; a compendium of proceedings volumes published in conjunction with various symposia over the years; a list of graduate schools and science/technology societies; sections on handbooks, electrochemical nomenclature and standards; and a list of scientific meetings of interest to electrochemists. When conducting a general topic search related to electrochemistry (e.g., "fuel cells," or "photoelectrochemistry"), *ESTIR* is invariably the place to start looking for information.

NIST WebBook. A useful site that contains physical and chemical data is the NIST *WebBook* (webbook.nist.gov). Billed as "a gateway to the data collections of the National Institute of Standards and Technology" (NIST), this resource provides thermochemical data for over 5000 organic and small inorganic compounds (e.g., enthalpies of formation and combustion, heat capacities and vapor pressures). Spectroscopic data is also available here, such as IR spectra (>7500 compounds), mass spectra (> 10,000 compounds), UV/Vis spectra (> 400 compounds), and electronic and vibrational spectra (> 3000 compounds). A sophisticated search mechanism has been incorporated that allows for retrieval of data based on compound name, chemical formula, CAS registry number, molecular weight, or selected ion energetics and spectral properties.

Patents. Useful chemical information resides in the patent literature. There are now at least two user-friendly ways to access these resources over the Web. The first is called the *IBM Intellectual Property Network* (www.patents.ibm.com). U.S. patents are collected into groupings of Front Pages, Front Pages & Claims, Titles & Abstracts, and Inventors & Companies that can be searched by keywords separately. Alternatively, patents can be retrieved using the patent number. Abstracts of Japanese, European and recent World Intellectual Property Organization patents are also available.

After performing a search, a text version of the abstract as well as the list of claims can be displayed. Also, graphic images (in the tagged image file (tif) format) of entire patents are often available. These can be viewed, although they are typically not high enough in resolution to print a legible copy. Registered users can purchase copies of the patents in the Adobe portable document format (pdf) using a shopping cart paradigm. These can either be downloaded or faxed to the purchaser.

The United States Patent and Trademark Office also maintains a publicly-accessible database, located at www.uspto.gov. The full text of patents can be retrieved using either Boolean search syntax or by entering the patent number. Scanned images (300 d.p.i. resolution) also in the tif format can be downloaded and displayed using web browsers equipped with an appropriate plug-in. The usefulness and comprehensiveness of this site have increased dramatically over the past year. The USPTO web team is to be congratulated on a job well-done!

The Code of Federal Regulations (CFR). The National Archives and Records Administration provides web access to the entire Code of Federal Regulations through their Internet site www.access.gpo.gov/nara/cfr. This database can be searched by keyword or individual sections may be retrieved. Results may be viewed as a summary (text), as a full text document, or in Adobe portable document format (pdf). This resource is used daily here at BAS to retrieve copies of FDA, OSHA, EPA and DOT regulations. Also of note is the separate web site maintained by OSHA which includes, among other things, links to "fully validated" analytical methods for many organic and inorganic compounds. The latter are most easily accessed at the URL: www.osha-slc.gov/dts/sltc/methods.

Material Safety Data Sheets (MSDS) and Other Safety Links. The jobs of safety directors everywhere have been made considerably easier by the now widespread availability of material safety data sheets on the web. Several excellent sites exist, but my personal favorite is run by the University of Vermont and has the URL: siri.uvm.edu. The same site has a page of links to other safety related web pages (hazard.com/links.html) including those covering ergonomics, industrial hygiene, radiation safety, and fire safety. For an explanation of National Fire Protection Association (NFPA) warning labels see: www.orcbs.msu.edu/chemical/nfpa/nfpa.html. Lastly, a cluster of toxicology-related databases maintained by the National Library of Medicine is available at the URL: toxnet.nlm.nih.gov. This resource provides information in more depth than is supplied in a typical MSDS.

The above paragraphs are obviously an incomplete and idiosyncratic list of Internet sites, but they represent those that have proven quite useful to my colleagues and I here at BAS.

No one knows everything, but the future belongs to those who know where to look for information they need to know. Let us know some of your favorites, and in a year or two, we'll prepare an update of this article for *Current Separations*.