USING NETWORK INFORMATION RESOURCES IN CHEMISTRY

Primož Škulj and Peter Krajnc

Laboratory for Organic and Bioorganic Chemistry, Faculty of Chemistry and Chemical Technology, University of Ljubljana, Aškerèeva 5, Ljubljana, Slovenia

Abstract: Retrieving chemical information from local and global networks was studied. The advances in computers and their interconnections have enabled various services based on client-server approach. The possibilities of specialised information services like the Scientific and Technical Information Network providing databases as Chemical Abstracts On-line and local services were compared to global and general server platforms on World Wide Web. Fast development is obvious and hints are given for successful conduct of searches or browsing primary and secondary literature.

Introduction

Informatics has thoroughly influenced the way everyday's activities are conducted. Computers have been used for calculation, later to increase personal productivity use, for presentation and finally, communication.

Because the search for scientific literature and data is of great importance for a research worker and not much has been said about the new methods of information retrieval which have emerged, we here present some of network accessible resources for study and research work.

Chemists were amongst the earliest users of computers for generating, acquiring or searching data what could not have been done without access to remote computers i. e. networking. [1-3] The most known network is certainly the Internet, but it is merely a network which connects many WANs (Wide Area Network) and

LANs (Local Area Network). The physical connections are made by different technologies, though the core is protocol TCP/IP which packs all transmitted data into packets and routes them to the addressee. The most known tasks on the network are electronic mail, transfer of files, access to Web servers. WANs and specially the whole Internet give the possibility of accessing information on machines anywhere. [4] Today the client-server technology prevails and its widest implementation is the World Wide Web developed at CERN. [5] Different network tasks are accomplished using various protocols - language by which machines communicate. Information on the Web are retrieved by Hypertext Transfer Protocol (`http`)[6], files are sent over using File Transfer Protocol (`ftp`)[6], on most UNIX platforms the mail is routed using Simple Mail Transfer Protocol (`smtp`)[6]. URL[6] (Universal Resource Locator) for instance: `http://pubs.acs.org/hotartcl/index.html` means that the computer named `pubs.acs.org` is accesed with protocol `http` and that the requested file `index.html` can be found in directory `/hotartcl/` on that computer. Every computer on the Internet has its own IP address that correspond to its name and the task of finding it is done by the nearest Domain Name Server (DNS) – a server specially for that purpose.

The computer being rather a communication tool than a typewriter substitute, enables access to information services and databases. Information is either public and (free of charge or per-pay) available to all as the Web and Scientific and Technical Information Network (STN) or some service run locally like Intranets or services providing access to University Library databases. The mass of available information is overwhelming so strategies were devised to make them manageable. More or less complete pieces of information about one subject are either made accessible using special software (for example STN providing CA online) or by publishing it on the Web. The advantage of on-line search is clearly evident from the possibility to combine or limit searches to achieve a manageable number of hits, which one can list and browse.

We present the following network accesible services: Current Contents Service, Scientific and Technical Information Network and chemical recources and chemical journals available on the Web[7].

Current Contents Service [8]

The profusion of original papers is so great that the publications that merely list the titles and abstracts of current papers find much use. One of those is the Current Contents Physical, Chemical and Earth Sciences, which has been released weekly since 1967 and has provided data and abstracts on articles published in more than 4000 journals for exactly one year back and is available also on-line. There exist various user interfaces to browse Current Contents. With the example search for textword `fibre` we retrieved more then 700 hits and for textword `polymer*` (here is `*` used as a truncation sign so program looks for all words with the same root, for example polymeric, polymerase, polymerised,...) hits counted to few thousand, but the combination of search terms would display only articles which contain both two matches. Further limitation to *latest update*, or *list of journals* would result in perhaps ten hits, which can be displayed with title, author(s), address of the author, number and year of publication, its page and abstract.

Scientific and Technical Information Network [9]

Databases in chemistry are available from several organisations, but by far the most important one is STN International (The Scientific & Technical Information Network), a service operated jointly by CAS in North America, by the Japan Science and Technology Corporation (JST) in Asia, and by Fachinformation Zentrum (FIZ) - Karlsruhe in Germany for users in Europe. STN charges for each use, depending on databases searched, for how long and what kind of information is retrieved.

Chemical Abstracts and other databases on-line

CA On-line is counterparting the printed Chemical Abstracts. It covers all areas of chemistry, biochemistry and chemical engineering. Sources include journals, patents, technical reports, books, conference proceedings, and dissertations. Bibliographic terms, indexing terms, and CAS Registry Numbers are searchable. Over 87% of the records also contain CA abstracts, the text of which is searchable. Unfortunately database is complete only from 1967 to present, nevertheless has many advantages.

Not only is it a great deal faster, but one can do kinds of searches online that are simply not possible using only the printed volumes. Furthermore, online files are regularly updated so one finds information well past the appearance of the latest semi-annual indexes, even before the library has received the last issue of CA. Nevertheless, CA File is just one of the many databases provided by STN; they have established a comprehensive supply of databases in science and technology. For references prior to 1967 there is a file CAOLD, which contains 695.000 records for the period from 1957-1966. Other important databases include: BEILSTEIN, which contains organic chemical structures, preparation and reaction information, and numeric property data. The source for the BEILSTEIN database is the Beilstein Handbook of Organic Chemistry; REGISTRY File is a chemical structure and dictionary database that contains unique substance records identified by the Chemical Abstracts Service (CAS) Registry System. Each record contains the CAS Registry Number, CA index name, commonly used synonyms, a structure diagram, and a molecular formula, all of which are searchable. Substances containing rings may be retrieved using ring system data; alloys may be retrieved using alloy composition information; and protein and nucleic acid sequences may be retrieved using codes for the amino acids or nucleotides; CASREACT contains information on reactions of organic substances, including organometallics and biomolecules, it also contains single and multi-step reaction information for reactants, products, reagents, solvents, and catalysts. The source for CASREACT is the Chemical Abstracts Organic Sections (21-34). [2,9,10]

A Messenger - common command language is used in all databases, though its major disadvantage is that it is almost fully command line oriented and allows no intuitive user interface. Free access to some of online learning databases is available on `telnet://a45.fiz-karlsruhe.de:4050`. It is limited to fifteen minutes and maximally 5 simultaneous users but allows one to learn basic commands and search strategies. Once connected to STN and with selected database one is presented with a command prompt to enter the query. With the command `SEARCH BENZENE` in learning database LBEILSTEIN we have got the answer:

```
L1 1678 BENZENE
```

The L1 means that this is line 1, the 1678 means that system has 1678 abstracts that contain word benzene. The word may be in the title, an index entry or a keyword. Compounds can be also searched for by using the Registry number. To display the first three matches for benzene one enters the command

```
DISPLAY L1 1-3
```

It is possible to specify the output format (bibliographical data, text of the abstract, abstract number only...).

The real power of online searching is its ability to combine and filter different queries. Operators of Boolean algebra (AND, NOT, OR) provide useful mean to extract the information needed. Instead of querying many words with single root, one can truncate the search term and get all the matches at once. There is no need to search several annual and monthly volumes because databases are unity as a whole. On the other hand it is possible to limit the search only to specific area, time period, journal, etc.

An alternative to terminal accessed STN International is STN Easy which provides point-and-click access to STN International. STN Easy operates through World Wide Web[9] and has two modes of searching; Basic Search - the easiest way to locate relevant answers by simple keyword searching and Advanced Search with greater flexibility in setting search criteria for author searches, Boolean operators, index browsing, etc.

World Wide Web

More than ten years ago an idea was persuived at CERN laboratories, Switzerland, to establish a common mean of accessing data on different computers running different software platforms [5]. Such tasks are usually achieved using server-client technologies, meaning that users retrieving information run client software on their machines and computer they are accessing is running server software. A client is equipped with what is needed to process and display data received from the server. Such system has many advantages; it lowers the burden (the needed bandwidth) on network connection and also on the server, because server processor power is not used for the displaying and formatting.. `Http` protocol [6] is the basis for WWW (World

Wide Web); a network of servers which can be all accessed from any computer. With WWW there is no need for specialised client software, a Web browser is sufficient. Fast development of the Internet can be illustrated by the number of computers, connected to Internet, which exceeded the 10.000 mark in 1987, 100.000 in 1989 and 1.000.000 in 1992.[11] While in January 1996 over 1000 servers, dealing with chemistry, existed[3], over 277.000 sites were found in January 1998 using the AltaVista search engine. This information clearly shows the rapid increase in the use of Internet as an information source for a scientific worker. As the number is huge we reviewed some of them and summarised the ones for which we think are most resourceful. In *Table 1* we list their URLs. Due to the numerous links they offer one can explore further on the Web to find the sites and topics of specific interest. Sites accessed frequently can be bookmarked [12] easily by most browsers.

Table 1: URLs of chemical sites on the Internet

• **University of Hertfordshire**	`http://www.herts.ac.uk/` `lrc/subjects/natsci/chem/chemweb/`
• **American Chemical Society**	`http://www.acs.org/`
• **Royal Society of Chemistry**	`http://www.rsc.org/`
• **University of Sheffield**	`http://www.shef.ac.uk/chemistry/chemdex/`
• **Chemcenter**	`http://www.chemcenter.org/`
• **Tennessee State University**	`http://acad.tnstate.edu/~chemnet/www.html`
• **Department of Chemistry, Imperial College of Science, Technology and Medicine, London**	`http://www.ch.ic.ac.uk/`
• **Faculty of Chemistry and Chemical Technology, University of Ljubljana**	`http://www.uni-lj.si/www/kem/`
• **Faculty of Chemistry and Chemical Engineering, University of Maribor**	`http://www.uni-mb.si/new/fkkt/okv_an.htm`
• **Slovenian Chemical Society**	`http://www.kemijsko-drustvo.ki.si/`
• **Slovenian National Institute of Chemistry**	`http://www.ki.si/`

Chemical journals on the Web [13]

Staying in touch with the latest developments in the field of research is of great importance for a scientific worker. Libraries have thus always been a special place for a scientist. However, with a growing number of journals on line [14], classical libraries are losing a bit of magical atmosphere since the computers are winning the battle of access speed and, even more important, most publishers now offer their own search engines to facilitate efficient or specific information retrieval. An article, dealing with one's topics of interest, can therefore quickly be located and viewed. Typical time of access to a bookmarked journal would be less than a minute provided the local net and server are not very crowded. In *Table 2* URLs of some renowned chemical journals are summarised.

Table 2: URL's of some chemical journals

● **Journals published by the Am. Chem. Soc.** (J. Org. Chem., J. Am. Chem. Soc., Chem. Rev., …)	`http://pubs.acs.org/`
● **Chemical Communications**	`http://www.rsc.org/is/journals/current/chemcomm/cccpub.htm`
● **Tetrahedron Information System** (Tetrahedron, Tetrahedron Lett., …)	`http://oxford.elsevier.com/tis/`

To access most of them, a username and a password are required, while some are at the moment still free of charge. Most of the journals offer their articles as both: HTML (Hypertext Markup Language) and PDF (Portable Document Format) [15], of which the later enables the viewer [16] to view on the monitor and print the pages exactly as they appear in the printed form of journal, which is of course important for schemes, tables and figures. In February 98, 387 chemical journals were found available on line [17], some of them complete, while others only with abstracts of articles. At `http://www.chemconnect.com/library/journals.shtml` there are links to probably all chemical journals on line. Advantage of following journals on-line can be clearly demonstrated by ASAP (As Soon as Publishable) service of American Chemical Society publishing peer-reviewed journal articles 2-11 weeks

before they appear in the print journal. Recently some journals emerged, published only on-line, without the printed version.[18]

Conclusion

Scientists have always stridden for an immediate access to scientific literature. The technological advances have enabled steady improvements and the advent of World Wide Web has rekindled popular interest in those issues. Traditionally the information retrieval was a task for professional librarians, who conducted the searches and reported results to querying scientists. The rise of the Internet has made most of the mediators obsolete and access to databases is enabled to a wider audience. Today the state of technology allows users to interact effectively with information distributed across the network. Network information systems in various forms support search and retrieval of items from organised collections. In their historical evolution, the mechanisms for retrieval of scientific literature have been particularly important. With move from syntactic to semantic and searching concepts rather then words the utilisation of network resources is becoming more pertinent and user friendly.

References and notes

[1]   B. R. Schatz, *Science* **1997**, *275*, 327-334
[2]   D.D. Ridley, *Online Searching: A Scientist's Perspective*; John Wiley and Sons, New York, **1996**
[3]   P. Murray-Rust, H. S. Rzepa, B. J. Whitaker. *Chem. Soc. Rev.* **1997.** *26*, 1-10
[4]   O. Kirch, *Linux Network Administrator's Guide,* **1996**, *http://www.linux.org*
[5] B. Segal, *http://wwwcn.cern.ch/pdp/ns/ben/TCPHIST.html*
[6]   Used terms are defined in request for comments (RFC) documents which are accessible at
       ftp://ftp.arnes.si/standards/ftp
       http - Hypertext Transfer Protocol (RFC 2068, RFC 1945), ftp - File Transfer Protocol (RFC 430),
       smtp Simple Mail Transfer Protocol (RFC 876), url - Universal Resource Locator (RFC 1738)
[7]   Popularity and simplicity of the Web are resulting in more and more services becoming reachable
       and searchable by an Web browser.
[8]   Institute for Scientific Information *http://www.isinet.com*
[9]   Fachinformationzentrum Karlsruhe *http://www.fiz-karlsruhe.de/*
[10] J. March, *Advanced Organic Chemistry*; John Wiley and Sons, New York, **1992**
[11] S. M. Bachrach (Ed.), *The Internet: A Guide For Chemists*, American Chemical Society, **1995**
[12] To *bookmark* a site is to store its URL adress on your computer's harddisk so you can later access
       this site without typing its URL adress.
[13] *Web* and *Internet* are both mistakenly often used to describe only the World Wide Web.
[14] On-line journals is an accepted phrase for journals appearing in an electronic form accesible via the
       Internet.

[15] PDF is a format developed to create and edit cross-platform documents by Adobe Systems Incorporated *http://www.adobe.com/*

[16] The word *viewer* instead of *reader* was used to distinguish between reading a printed article and article appearing on the computer screen.

[17] *http://www.chemconnect.com/library/journals.shtml*

[18] *http://www.ijc.com/*

Povzetek: Vzpostavitev računalniških omrežij je omogočila dostop do različnih novih ali spremenjenih virov informacij. Predstavila sva informacijske servise, namenjene prvenstveno kemikom. Informacijski viri, dostopni raziskovalcem, so bodisi lokani ali globalni. Pregledala sva njihove lastnosti in uporabo ter podala nekaj napotkov za uspešno iskanje infromacij iz omrežnih virov.