DOCUMENT RESUME

ED 411 824 IR 056 651

AUTHOR Inlenfeldt, Wolf-D.
TITLE Chemical Structure Search on the World Wide Web.
PUB DATE 1996-00-00
NOTE 9p.; In: Online Information 96. Proceedings of the International Online Information Meeting (20th, Olympia 2, London, England, United Kingdom, December 3-5, 1996); see IR 056 631.
PUB TYPE Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Access to Information; *Chemistry; Foreign Countries; *Information Retrieval; Information Sources; *Online Searching; *Reference Materials; Search Strategies; User Needs (Information); *World Wide Web
IDENTIFIERS *Chemical Structure; *Germany

ABSTRACT

Chemists have been very active in utilizing the World Wide Web as an information distribution medium and much interesting scientific chemical information is already offered on it. Various classical text-based search engines have made locating information on the Web easier. However, keyword-based searches are often insufficient for chemists interested in structural features of chemical compounds, especially since the naming of chemical structures is far from simple or unique. Chemists have special search requirements and rely on non-textual structure-oriented search methodology. An increasing number of chemical structures can be found in various computer-readable structure exchange formats as MIME attachments to Web pages on the Internet. If they are cataloged and made searchable with the structure-oriented search methodology that chemists are used to, they can lead to valuable chemical information sources on the Internet which are difficult to locate with other, text-oriented search methods. The Computer Chemistry Centre, University of Erlangen-Nurnberg, (Germany) has implemented a system for the collection, recording, search and context-aware retrieval of chemical structures from the Web. (Contains 18 references.) (Author/AEF)

Chemical Structure Search on the World Wide Web

By:

Wolf-D. Inlenfeldt

Chemical structure search on the World Wide Web

Wolf-D. Ihlenfeldt
Computer Chemistry Centre, University of Erlangen-Nürnberg, Germany

Abstract: *Chemists have been very active in utilising the World Wide Web (WWW) as an information distribution medium and much interesting scientific chemical information is already offered on it. Various classical text-based search engines have made locating information in Webspace much easier. However, keyword-based searches are often insufficient for chemists interested in structural features of chemical compounds, especially since the naming of chemical structures is far from simple or unique. Chemists have special search requirements and rely on non-textual structure-oriented search methodology. An increasing number of chemical structures can be found in various computer-readable structure exchange formats as MIME attachments to WWW pages on the Internet. If they are catalogued and made searchable with the structure-oriented search methodology that chemists are used to, they can lead to valuable chemical information sources on the Internet which are difficult to locate with other, text-oriented search methods. We have implemented a system for the collection, re-coding, search and context-aware retrieval of chemical structures from the WWW. We assume it is the first instance of a non-textual search engine for WWW data.*

Keywords: Web spider, Web search, Web database, chemical structures, MIME

1. Introduction

Chemical information is unique in its contents. While of course research results are described in words, the information the chemist is most interested in is often non-textual. The focus of interest is the chemical structure — either as a schematic entity depicting the connectivity pattern of the atoms, or increasingly as a three-dimensional object. The latter type creates the expected difficulties of displaying the subtleties of a specific molecular conformation by flat and still mostly monochrome images in printed media.

Nearly all chemical structures are systematically extracted from literature and stored in large databases such as CAS Online or Beilstein Online, together with abstracts and references to the original literature. Structure extraction is performed manually by necessity. These databases are searchable by structural criteria (for example full-structure and substructure search) and guide chemists to the original printed articles. Structure-oriented searches are especially important because the official naming schemes for chemical compounds are extremely complicated, error-prone and, even if the name is constructed according to the rules, far from unequivocal. Therefore, the search for a specific compound by name or name fragment is often impossible, or prone to missing many database entries where a different name has been used.

Compared to traditional printed media, the WWW offers some notable advantages to the chemical community (Ref 1). It has therefore been readily accepted as an interesting new information distribution medium, especially for conference proceedings where the problems of copyright, long-term storage and so forth are not so severe. There have even been virtual Internet-based conferences without any face-to-face meetings. Generous use of colour, video and so on is no problem with HTML pages, but the most important feature only possible in electronic media is the attachment of computer-readable structure files as links to hypertext documents (Ref 2). Following these links, chemists can download the original structures into their desktop computers and examine them in detail with helper applications, for example by rotating three-dimensional structures in molecule viewers. A set of MIME types for chemical structure exchange has been proposed to the standards committees (Ref 3). Even before the standard has gained official blessing, the number of chemical papers on the Web which contain attached structure files has exploded recently. Estimations are that at the time of the conference, their number will have surpassed 10,000 — a number which makes indexing mandatory and an interesting problem. Textual indexing services can be provided by the usual WWW search engines, but there are currently no structure-oriented indexing services (Ref 4). Manual compound registration by the authors is certainly not practicable outside the scope of small conferences (Ref 5). Semi-automatically linking compound name occurrences on hand-picked sites to a manually generated structure database is not much more promising (Ref 6). We therefore set out to determine the possibility of creating a fully automatically generated structure-oriented and therefore non-textual database with information from WWW sites (Ref 7).

The information content of the Internet is completely uncontrolled. Therefore the search for specialised infor-

mation types employed only by experts, such as structure data, has the advantage that the scientific content of the hits is generally much higher than in plain text searches. This is especially notable if chemical compounds are considered whose name happens to be known to a broader audience. Textual searches in WWW search engines for scientific information on vitamins, plant protection agents, drugs and so forth are tedious, because the comparatively few scientific information sources are far outnumbered by pages describing health food, cosmetics and similar merchandise or containing political statements. Text in the form of HTML pages is a medium shared between scientists, commerce and interest groups, and text pages are not automatically classified by type of contents in any search engine we know about. As a contrast, we are currently not aware of a single instance of a structure file not related to science in the wider sense, a few abuses as adornments for conference and departmental home pages notwithstanding.

2. Information gathering

The system we have developed consists of four major components. The first is a kind of Web spider, which traverses the Web in an attempt to locate chemical structure files. The second is a sophisticated converter programme, which unifies the various file formats into a common representation suitable for inclusion in a structure database. A third component, the verifier, checks whether the sources of the database records are still accessible at their original locations and that their contents are unchanged. The final major system component is the structure database proper, which is accessible via Internet channels either by a custom client or by means of a form-based Web interface.

It is fruitless to attempt another complete indexing of Webspace (the sum of all WWW-accessible data). Rather than search randomly, our spider starts around a set of promising start pages and expands them two or more link steps. Promising start pages in this context are pages which contain words indicating chemical content. We currently use a manually generated word list of some 100 terms. Typical words contained in the list are 'chemistry', 'molecule' or 'phenyl'. A start site generation script presents these words to standard textual Internet search engines and extracts the URLs from the replies. Currently the script decodes the answer pages of Lycos and Alta Vista (Ref 8). Unfortunately, none of the bigger search engines allows the complete retrieval of large answer sets any longer. Therefore it is impossible to obtain more than only a small portion of, for example, the URLs containing the word 'chemistry', which is present in more than 400,000 instances in the June 1996 Alta Vista index. By expanding the vocabulary to other more specialised chemistry-related terms and automatically merging the search engine results, we routinely update a database of starting URLs which currently contains about 15,000 entry points (Figure 1).

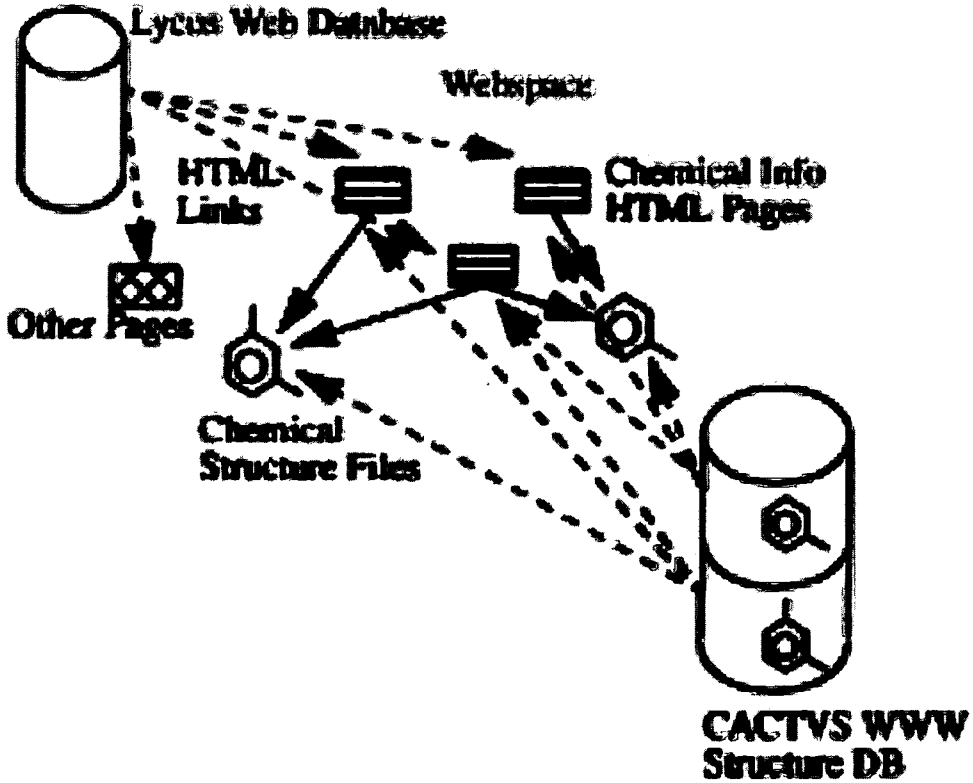

Figure 1: Relationship between textual Web databases (for example, Lycos) and the Web molecule database.

The spider program expands the links in and around these pages up to a specified depth, which is initially set to two for unclassified sites. We do not store the page contents. Only the extracted links and the link texts, plus some administrative information such as the time of the last visit, are collected in a set of gdbm databases. The 15,000 starting points expand to about 750,000 different URLs. Host name aliases are resolved to primary names, but it is technically impossible to detect aliased directories. The gatherer is regularly running at European night-time on a small cluster of three to five workstations, and comprises about 30 independent processes. This is the optimum in the current implementation because after that the read/write synchronisation and locking of the gdbm files becomes the limiting factor. For performance reasons, it is mandatory to cache information from the network wherever possible. This is especially true for host name resolutions, which can take a comparatively long time. If a host name in a URL is in the persistent shared cache and the entry is reasonably recent, the WWW page retrieval is conducted directly with the host's known IP address instead of its name. Another example of useful caching is the standard Web robot exclusion file *robots.txt* from the server root directory. It is cached in pre-parsed form (Ref 9). The spider program, including the client/server communication routines plus all additional scripts used in the information collection phase, were written for convenience in Tcl/TclX (Ref 10). Performance is limited by the network connections and NFS file locking, not the page data analysis.

From time to time, the links to structure files are sifted out from the link database. The basic recognition criteria are the file name suffixes which are more or less consistent and characteristic for the various popular chemical structure exchange file formats. Additionally, a clustering step is inserted here. Sites which contain a notable number of structure files whose URLs have a common part are detected. Another expansion step with an extended link depth of three to five levels is performed, starting with the common directory part. During these scans auxiliary checks are performed to avoid wandering off the cluster site. This second sweep of cluster sites typically yields another set of structures comparable in size to the initially located set. Files suspected to be structure files are downloaded and stored into another gdbm database with their URL as key. At the time of writing, a total of about 3000 structures have been found. This is very little compared to the more than 20 million entries in databases like CAS Online, and the 30 million WWW pages in the Alta Vista index of June 1996. Also, the scanning effectiveness is still small, comparing the number of structure files to the locally expanded 750,000 links in the spider database (0.4%). However, the number of structure files is explosively growing and has been approximately doubling every two months in the last year. We think it is necessary to develop techniques to index this kind of information now in order to be prepared for the arrival of the expected flood of new data of this kind in the near future should the growth continue — and there are no reasons why it should not, especially since now a number of refereed electronic chemical journals have gone online. Three thousand compounds is a number which definitely is no longer manageable manually but small enough to be handled by a simply structured experimental database.

It has been asked why we did not try to locate the structure files directly by textual search with their characteristic endings as query. The main reason is that links to these files are necessarily unidirectional. None of the common chemical structure file formats has provisions to include a URL as a back link to a referring page. The structure itself, while it is the primary vehicle for the location of information for chemists, is of value only in the context of the pages describing its properties and the experiments performed with it. Therefore, not only the structures must be registered but also the links leading to it. These can be located only by a search starting at text pages. Only very recently search engines such as Alta Vista have implemented search capabilities to locate referring pages. Once this functionality becomes stable and the search engines allow (again) the retrieval of an unlimited number of hits, it may be possible to construct the link tree by such queries instead of building it oneself. However, facing a constant overload, these search engines become continuously less responsive and therefore it is not clear whether an approach based exclusively on search engine results would actually be faster than a restricted local Web traversal.

3. Information re-coding

The second phase in the production of the database is the re-coding and unification of the structure files. This is a prerequisite for structure-oriented searchability. This task turned out to be magnitudes more difficult than the collection of the primary data and link relationships. There are about two dozen important structure file types in the MIME set and we have implemented I/O modules for most of them. Unfortunately, many of them — and unfortunately the most popular ones — are not very well suited for the coding of connectivity-based structure representations which are the base for all connectivity-oriented structure search algorithms. Among the structure file formats, the two most common ones are the formats of the Brookhaven Protein Database (PDB) and the XYZ format. The XYZ format is truly minimalist in its information content. It contains only three-dimensional atom coordinates, which are normally sufficient for simple 3D visualisation tasks, but the files do not contain any connectivity at all. PDB files can contain connectivity but often do not, and worse, there are no provisions to encode bond order — a very important concept for structure-oriented searches. We encountered numerous cases where PDB files contained connectivity only for selected parts of a molecule which were related to a certain problem. Furthermore, due to the ill-defined nature of the PDB format which is several decades old and was originally designed for a specific database of proteins, it has often been severely abused to encode data the format was not designed to hold. In PDB files, it is not even possible to determine unequivocally whether an atom labelled CE is an e-carbon atom (of the amino acid lysine, for example) or cerium — the two very different meanings are normally discernible for chemists looking at the overall nature of the compound

but difficult to recognise programmatically in unknown files without context information. Both meanings were actually encountered.

In contrast to connectivity reconstruction, as it is for example performed when electron density patterns from x-ray crystal structure analysis are decoded, several important pieces of auxiliary information are missing in many of the structure records. For example, often the structures are not guaranteed to be electroneutral and the status of hydrogen atoms (whether they are all included, or only at selected atoms, or — following the custom with certain compound classes — not at all) is generally unclear. It was a major effort (about 20,000 lines of code) to implement the machinery which now for more than 95% of the samples reconstructs a structure with a complete hydrogen set connectivity including bond order, stereochemical descriptors and 2D display coordinates which is identical to structures reconstructed manually by an experienced chemist. The system relies in its functionality on the chemical data handling capabilities of the CACTVS system (Ref 11). It uses a multi-step approach. At several processing stages the introduction of heuristics was unavoidable. They were derived from general chemical knowledge, geared to the problem domain. The most important processing steps are, in the order listed, atom type recognition, hydrogen status guess, generation of the first set of single bonds, ring detection, tentative assignment of aromaticity to rings, aromatic ring Kekulé structure generation, multiple bond order detection, radical and charge equilibration, neutralisation of improbable charges and radical centres, display coordinate generation, stereo centre detection, stereo descriptor assignment, wedge bond assignment, and finally the generation of display attributes for double bonds inside rings, suppressed hydrogen atoms, carbon atoms displayed by convention only as node and not as symbol etc.

Reconstruction of structures is currently a batch process. The time scale ranges from less than a second for small molecules without complex aromatic ring patterns to half a minute for large proteins (several thousand atoms) with very incomplete information. These figures were measured on a 200 MHz SGI Indigo II workstation. All structures are converted into a common format and augmented with display coordinates, stereochemical descriptors, names (if present in the original file), the original URL and the two-level tree of referring URLs. We do not retain the original file after conversion. We also compute a group of very discriminative 64-bit hash codes for all structures (Ref 12).

The hash codes are used by the verifier and for full-structure searches. The verifier is a separate program which regularly checks whether the structures in the database are still accessible and have not changed. The verifier operates first by downloading the structures using the stored original URL. If successful, it repeats the conversion process and compares the hash codes of the original and the database entry. In case of changes, all URLs in the link database which share a common name part with any of the URLs referring directly or indirectly to the changed or disappeared structure are marked for a rescan. This strategy helps to detect new structures because often the contents of changed sites were globally edited, not just with respect to the single structure where the change was detected.

4. Database access

The database is managed by a customised Postgres95 object-relational database server (Ref 13). It employs dynamically loaded custom function modules to provide the chemistry-specific structure search facilities. The implementation of the substructure search follows the traditional model. It uses a 256 bit fragment-based screening vector which filters out most of the records in the database from the candidate list. Only the survivors are submitted to an atom-by-atom matching step (Ref 14). Full-structure search is provided without a subsequent atom-by-atom match by simple comparison of one member of the aforementioned set of 64-bit hashcodes. The hashcoding procedure is so reliable that there are currently no known examples of collisions. The algorithm has been subjected to tests with documented problematic cases from the literature (Ref 12). The hashcode is pre-computed for the database records in several variations, such as with inclusion or omission of stereochemistry, which correspond to the various full-structure search modes.

The performance of the database, though certainly not optimised for this kind of application, is sufficient for the current number of entries. Typical retrieval times are from one to five seconds on an Indigo II workstation. A noteworthy feature of our database is that there are no molecule size limitations. The molecule sizes in this dataset vary dramatically, from two to several thousand atoms. Commercial molecule database products typically impose a rather low limit on the number of atoms they can handle, and many structures from our comprehensive collection could not be stored in such a database. In our approach, the storage model of the structures in the database is dynamic and transparently switches from an opaque field in the structure table proper to an external large object if a size threshold is exceeded.

The structure database is freely accessible via the Internet as a research prototype. We provide two basic methods of access: by means of a custom client, and from a standard WWW page.

The custom client is part of the CACTVS tool distribution and can be obtained free of charge for the major Unix operating systems (Ref 15). The advantage of the client is that it is tightly integrated with other tools from the distribution and can use complex means of graphical visualisation. Drag and drop of structures and search fragments in graphical representation and other convenient operations are supported (Figure 2). This degree of functionality is impossible to achieve with standard HTML pages. From within the client, the Mosaic WWW browser can be instructed via its CCI interface to display the referring text pages (Ref 16).

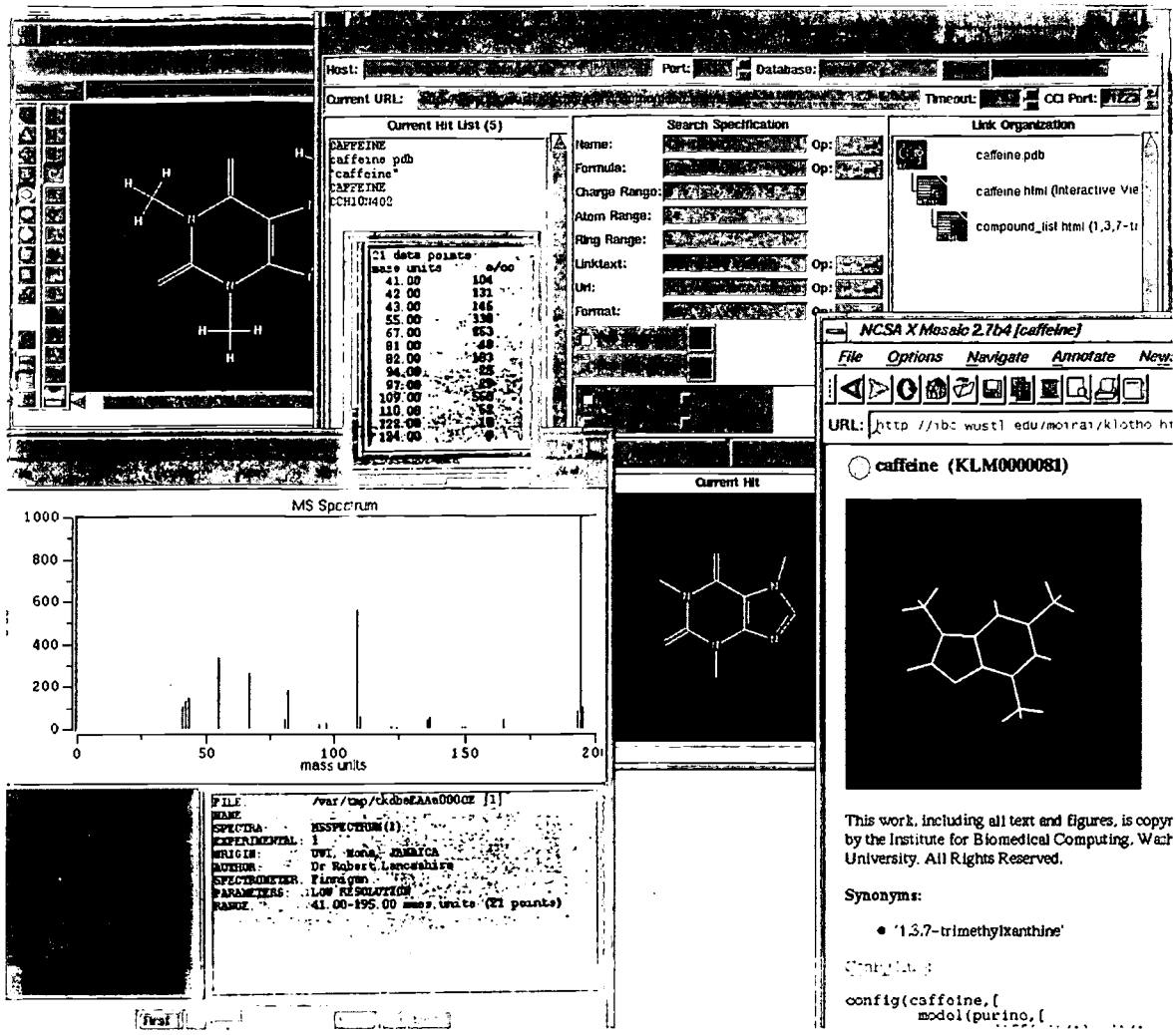

Figure 2: Searching for chemical information about caffeine on the WWW with the database client (upper right). The structure of caffeine was drawn in a structure editor (upper left) and dragged into the full-structure search pocket of the client. Five hits were found. By selecting one of those hits (highlighted yellow), the corresponding structure is retrieved and its link context displayed in the right-hand window of the client. Referring HTML pages can be displayed in a standard WWW browser by selecting them from the context tree (lower right, on this page the biochemistry of caffeine is elucidated). Chemistry-specific information is viewed with helper applications. As an example the mass spectrum of caffeine is displayed on the lower left.

Access by means of the alternative forms-based WWW interface (*http://www.schiele.organik.uni-erlangen.de/services/webmol.html*) is somewhat less convenient, but avoids the use of specialised software (Ref 17). Search fragments must be input manually as SMILES strings, or drawn with an external structure editor and pasted into text fields as ASCII structure files. Response pages are dynamically generated by CGI scripts on the server side. These pages include GIF images of the structures matching the query, so this mode of access generates a notably bigger load on the network than the compact transfer of the information to the custom client, which handles structure display and query command status updates autonomously. The CGI program at the server side is essentially a GUI-less version of the client program. It outputs HTML pages and GIF images instead of the normal GUI-based query result representations. Both programs are scripts written in Tcl/TclX and are executed on a general-purpose interpreter program for chemical structure handling. The only major difference between the two interpreter programs is that the client version is linked with the Tk GUI toolkit and the CGI program is not (Figure 3).

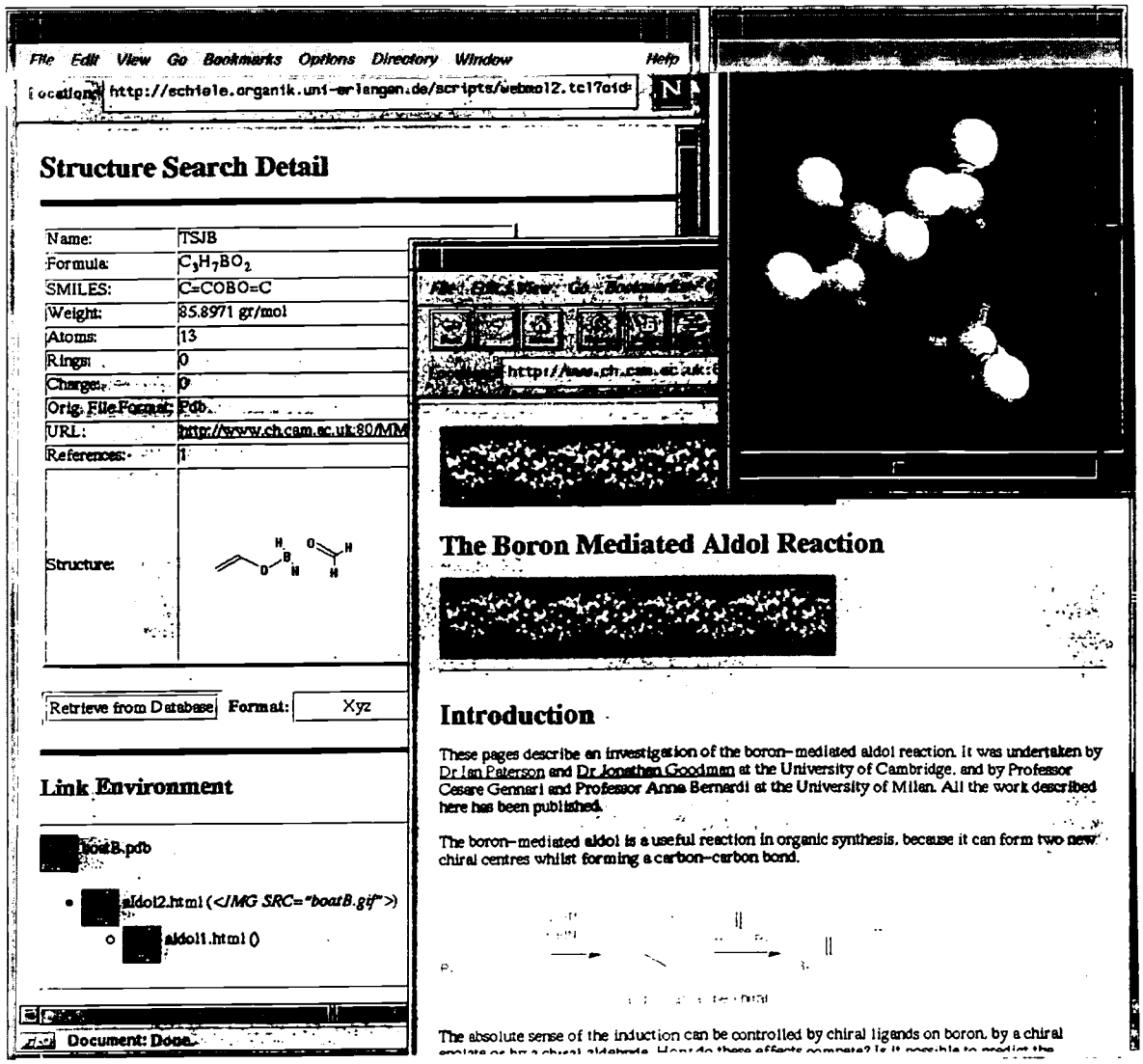

Figure 3: Searching for information in WWW documents about the mechanism of boron-mediated reactions. Here the forms-based WWW interface is employed. Details about a specific hit are displayed in a dynamically generated HTML page (left). The link environment of the hit is shown on the lower left. The icons are linked to the original pages and an original paper was loaded into another browser window on the lower right. The structure can be retrieved as 3D model for close inspection with the aid of molecular visualisation helper applications (upper right).

5. Query Example

A search for coffeine or caffeine yielded more than 10,000 matches in the Alta Vista June 1996 catalogue (Figure 1). As expected, truly chemical information is extremely sparse among these hits. The corresponding structure-based search in our catalogue resulted in only five hits. A string search on the URL of the registered files yielded another hit: The mass spectrum of caffeine was coded in a JCAMP-DX file without an attached structure. All references contain information highly relevant for chemists and therefore a complete suppression of irrelevant information was achieved. However, an examination of some of the hits obtained by the keyword search revealed that not all interesting information had been retrieved, since there are a number of information sources dealing with caffeine and its chemistry which include drawings of formulae and reaction schemes as images only, and do not yet link structure files to their texts which could be indexed by our robot. It remains to be hoped that once the value of structure attachments for chemistry-based indexing becomes more widely known, the majority of sites focusing on chemists as clients will provide these attachments. Since non-scientific information on the WWW is certainly growing more rapidly than scientific information, the selectivity of text-based search engines will continue to degrade. Automatically indexed structure attachments are a convenient method for information providers to help them to stand out and to draw visits from the targeted scientific readership.

6. Summary

We have implemented a structure-oriented search and retrieval system for chemical compounds found as MIME attachments to textual HTML information on the WWW. As far as we know, this system constitutes the first non-textual search engine for WWW information (Ref 18). We expect that the importance of non-textual specialised information will grow very rapidly for the Internet community. We have demonstrated that the collection, re-coding and database production of molecular information from the Internet is not trivial but absolutely feasible. Searches on this data provide valuable information for users with scientific interests which is not available, or much more on target than results obtained, by other available Internet search methodology.

Dr Wolf-Dietrich Ihlenfeldt
Computer Chemistry Centre
University of Erlangen-Nürnberg
Nägelsbachstraße 25
D-91052 Erlangen
Germany
Tel. +49 9131 856579
Fax +49 9131 856566
E-mail: wdi@eros.ccc.uni-erlangen.de

Acknowledgements:

This work was supported by FIZ Chemie, Berlin.

References

[1] Rzepa, H.S., B.J. Whitaker and M.J. Winter (1994) Chemical applications of the World-Wide-Web system, *Journal of the Chemical Society*, Chemical Communications 1907-1909.

[2] Krieger, J.H. (1995) Chemistry sites proliferate on the Internet's World Wide Web, *Chemical and Engineering News*, **73**(46), 35-46

[3] Chemical MIME proposal: http://www.ch.ic.ac.uk/chemime2.html

[4] Ihlenfeldt, W.D. and J. Gasteiger (1995) Beyond the hyperactive molecule: search, salvage and visualization of chemical information from the Internet, in L. Hunter and T.E. Klein (Eds.) *Pacific Symposium on Biocomputing 96*, World Scientific, pp. 384-395.

[5] The ECTOC Hyperglossary: http://www.ch.ic.ac.uk/ectoc/glossary/

[6] The ChemFinder database: http://chemfinder.camsoft.com/

[7] Ihlenfeldt, W.D. and J. Gasteiger (1996) Treasure hunt: on the track of reusable structural information in the World Wide Web, *Software Development in Chemistry 10*, Springer, in print.

[8] The URLs of the cited search engines are:
Alta Vista: http://altavista.digital.com/
Lycos: http://www.lycons.com/

[9] Robot exclusion standard: http://info.webcrawler.com/mak/projects/norobots.html

[10] Ousterhout, J. (1994) *Tcl and the Tk Toolkit*, Addison Wesley. See also http://www.sunlabs.com/research/tcl/docs.html

[11] Ihlenfeldt, W.D., Y. Takahashi and H. Abe (1994) Computation and management of chemical properties in CACTVS: an extensible networked approach toward modularity and compatibility, *Journal of Chemical Information and Computer Science*, **34**(1), 109-116.

[12] Ihlenfeldt, W.D. and J. Gasteiger (1994), Hash codes for the identification and classification of molecular structure elements, *Journal of Computational Chemistry*, **15**(5), 793-813.

[13] Postgres 95 database home page: http://epoch.cs.berekely.edu:8000/postgres95/index.html

[14] Hicks, M.G. and C.J. Jochum (1990) Substructure search systems, 1. performance comparisons of the MACCS, DARC, HTSS, CAS Registry MVSSS and S4 substructure search systems, *Journal of Chemical Information and Computer Science*, **30**(2), 191-199.

[15] CACTVS system home page: http://www.schiele.organik.uni-erlangen/cactvs/

[16] Mosaic CCI interface: http://www.ncsa.uiuc.edu/SDG/Software/XMosaic/CCI/cci-spec.html

[17] WWW chemical structures database Web page: http://www.schiele.organik.uni-erlangen.de/services/webmol.html

[18] Other interesting non-textual information sources on the WWW are maps, VRML scenes and images. See for example Wilensky, R. (1996) Toward work-centered digital information services, *Computer*, **29**(3), 37-44.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC®

NOTICE

REPRODUCTION BASIS

☒ This document is covered by a signed "Reproduction Release (Blanket)" form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").