Chemie.DE – Setting Up an Internet Information Service for Chemistry

Hans Benedict, Holger Busse, Wolfgang Dreißig, Burkhard Kirste*, Thomas Richter and Claus Schröter

Fachbereich Chemie, Freie Universität, D-14195 Berlin

Abstract. A rapidly increasing amount of valuable chemical and biochemical information is offered on the Internet, but the task of retrieving specific information is virtually impossible without recourse to appropriate tools. General-purpose search engines cannot solve the problem. Project Chemie.DE, accessible at URL http://www.chemie.de/, is offering databases, retrievable by intelligent and chemistry-oriented search tools, currently for the following topics: classified chemistry-related Internet documents, a calendar of conferences and exhibitions, software descriptions, job offers and applications.

Introduction

A flood of information is already available in the Internet and in data bases dealing with chemical or biochemical problems, but individual scientists can hardly cope with the complex and time-consuming task of searching and evaluating relevant information which might be hidden among millions of information pages scattered all over the net. General-purpose search engines cannot solve the problem, but specific and intelligent search tools are required.

The aim of project "Chemie.DE", which started in December 1996, is not only to compile relevant information and to offer it in a structured and easily retrievable way, but also to offer useful tools and a platform for communication among scientists.

Aims

Chemie.DE[1] focusses on setting up databases, retrievable by intelligent and chemistry-oriented search tools, for the following topics:

- list of *chemistry information servers* with description of their main topics and some quality estimation;
- *meta index of all chemistry-related Internet documents* with classification on several attributes, e.g., 'contents', 'chemical topic', 'language', 'creation date' – there are also ranking schemes available;

[1] http://www.chemie.de/

- *software descriptions*, both public domain (PD) and commercial, with links to vendors, authors, distributors, mailing lists – most relevant PD software products are held on the project's file server;
- calendar of German and international *conferences*, workshops, meetings, and exhibitions;
- collection of *job offers and applications*;
- database of *chemical companies and their product lines*.

In addition, search tools, in particular for 2D and 3D molecular structures, will have to be developed. Small, specialized data bases (hazardous substances, thesauri for technical terms, molecular structures, occupational safety, waste disposal) shall be offered.

Current Status

A test version of the server http://www.chemie.de/ went public in May 1997. Meanwhile (November 1997) the design and functionality have been improved. Since the performance of the database system Postgres (PostgreSQL)[2] proved to be insufficient, it was replaced by the commercial product Solid[3].

The collection of links to chemistry-related documents currently amounts to about 14000 entries, 6000 of them have been classified manually. The calendar of events with about 600 entries is probably the largest world-wide in the field of chemistry. A mailing list for discussions of interesting problems has been opened. Server statistics may be viewed by means of a series of informative diagrams displaying the daily number of hits and the volume transferred as well as the number of hits by domains or for the most popular files.

Two tools are offered, namely a searchable dictionary of acronyms and abbreviations and a tool for the conversion of units.

How it Works

A dynamically configurable database system is employed for the meta index of chemistry-related Internet documents (collection of links), the software descriptions, the calendar of events etc. The layout database can be generated and modified dynamically by means of WWW front ends: Input and output masks are generated automatically, bilingual descriptions of attributes are given (German and English). External users may contribute to the information content of Chemie.DE; registration and user administration are required for that purpose. The integrated database interface PHP/FI allows different views of the data, such as search requests or hierarchical indices.

Depending on the category, information may come from different sources. For the link collection, an intelligent robot is tracking down chemistry-specific documents in the Internet and stores them locally. A filter program compares

[2] http://www.postgresql.org/
[3] http://www.solidtech.com/

the content of these documents with those already stored in the pool database with respect to ambiguities, unvalidated links and missing key information. A pattern recognition program tries to classify the documents according to predefined categories. In this process, meta data, title information and hyperlinks are extracted automatically. Additionally, an index of references is kept for each entry which may be seen as a measure of "popularity" of the document. Documents that have been unambiguously classified as being relevant to chemistry will contribute to the word list used for the pattern recognition system. This process may be controlled by means of local front ends. Thus, new documents or those which are difficult to classify can be added to the database semiautomatically. A background process (validator) checks the validity of the database entries in regular intervals and keeps the corresponding meta information abreast.

Technical Details

The *hardware* of the main server consists of a dual processor machine with two 200 MHz Pentium Pro CPUs, 128 MB of RAM and two 4 GB disks running in RAID0 mode. *Software:* The freely available operating system Linux[4] is used, and the web (HTTP) server Apache[5] is running with the CGI wrapper and database interface PHP/FI[6]. The commercial database Solid[7] is used.

Conclusions

After the first year of development, the server Chemie.DE is offering quite a few useful services for chemists. It stands to reason that the contents of the databases are presently far from complete, but this situation is going to improve within the next year. Currently an IO library is under development which will make it easier to set up new databases and the corresponding input/output forms. For instance, a database of chemical companies and one for job offers and applications are under construction.

Acknowledgment

We thank the students Rolf Claessen, Stefan Knecht, Markus Miertschink, Steffen Moeller, Manuela Scheuner, Christian Steiner and Kathrin Strahl for their contributions. The project Chemie.DE is supported by the Verein Deutsches Forschungsnetz (DFN) and by a financial grant from the German Bundesministerium für Bildung, Forschung und Technologie (BMBF) which is gratefully acknowledged.

[4] http://www.linux.org/
[5] http://www.apache.org/
[6] http://php.iquest.net/
[7] http://www.solidtech.com/