The United Kingdom Chemical Database Service

David A. Fletcher, Robert F. McMeeking,* and Donald Parkin

Chemical Database Service, CCLRC Daresbury Laboratory, Warrington, Cheshire WA4 4AD, UK

Received February 27, 1996[⊗]

The Chemical Database Service (CDS) is a national service, funded by the Chemistry Programme of the United Kingdom Engineering and Physical Sciences Research Council (EPSRC). It provides access for UK academics to a range of chemistry databases in the areas of crystallography, synthetic organic chemistry, spectroscopy, and physical chemistry. Three post-doctoral chemists are available to assist users with problems, run training courses, and also give advice to the community on accessing other sources of chemical data and software.

INTRODUCTION

There are currently only two national centers in Europe that provide centralized chemical information services over networks. These are the CAOS/CAMM Center[1] in Nijmegen, The Netherlands, which was founded ten years ago, and the UK Chemical Database Service (CDS) at Daresbury. In addition the international data center operated by the Canadian Scientific Numeric Database Service (CAN/SND) of the National Research Council of Canada (NRCC) has been described in a recent article.[2]

The CDS has its origins in the late 1970s with a service based on the CSSR retrieval program to provide interactive networked access to the Cambridge Structural Databank (CSD)[3] compiled by the Cambridge Crystallographic Data Centre (CCDC). The CSSR program was developed from methods conceived by R. J. Feldmann of the National Institutes of Health, Bethesda, U.S.A. and tested and developed within the Environmental Protection Agency Chemical Information System (CIS).[4] Subsequently codes were adapted to provide access to ^{13}C NMR and fine chemical data.[5] In 1984, the then Science and Engineering Research Council (SERC) agreed to fund a dedicated VAX VMS[6] computer system to host the Service at their Daresbury Laboratory. The main focus of the service continued to be the provision of access to crystallographic data. Retrieval software was developed for searching the Inorganic Crystal Structure (ICSD) file[7,8] and the metals, intermetallics, and alloys file, CRYSTMET[2,9] as well as the NIST Crystal Data Identification file[10] and the Brookhaven Protein Data Bank (PDB).[11]

During the lifetime of the Service there have been dramatic changes in both chemical informatics and computer technologies. The quantity and diversity of chemical information available in electronic format has rapidly increased, and this information has become a key enabling resource for the chemistry community. With the advent of cheap storage devices, such as CD ROMs, a vast amount of data has become available for desk top PCs and workstations, and the need for central database facilities has been questioned. In 1992 the SERC Chemistry Committee set up a committee, initially under the chairmanship of Prof. K. R. Jennings of Warwick University, to review the whole issue of chemical database provision.

The committee recognized that there are a number of distinct advantages for central facilities including cost sharing, specialist information, access to large systems which would be beyond the means of individual users, support, and training. It was decided that a central Service should continue to be supported but put out to competitive tender. All UK Universities as well as the Daresbury Laboratory were invited to bid, with Daresbury being chosen to continue running the Service.

THE SERVICE

The specified aim of the Service was to ensure that the growing body of information from chemical research is conveniently accessible to UK academics in order to directly promote scientific and, through the supply of graduate personnel familiar with modern database systems, industrial advance. The Service was set a number of tasks and objectives. These comprise the following:

*to provide on-line access to up-to-date, high quality, comprehensive databases
*to provide training in the above databases
*to ensure that the Service is conveniently accessible
*to ensure that the UK chemistry community is aware of the service's potential for promoting scientific research
*to provide advice about other chemical information services.

Four chemistry disciplines were highlighted:

*Crystallography and structural databases, a fundamentally important resource to the chemistry community
*Synthetic organic chemistry databases, which play an important enabling role in elucidating possible reaction strategies
*Spectral databases, which are growing in importance as spectroscopic techniques are more widely used
*Physical chemistry databases, covering physical properties of chemical compounds.

The Service is currently funded by the Chemistry Programme of the Engineering and Physical Sciences Research Council (EPSRC) on the basis of a four year rolling grant. It provides access, free of charge at the point of use, for the

* Author to whom all correspondence should be addressed (r.f.mcmeeking@dl.ac.uk).

[⊗] Abstract published in *Advance ACS Abstracts,* June 15, 1996.

S0095-2338(96)00015-7 CCC: $12.00

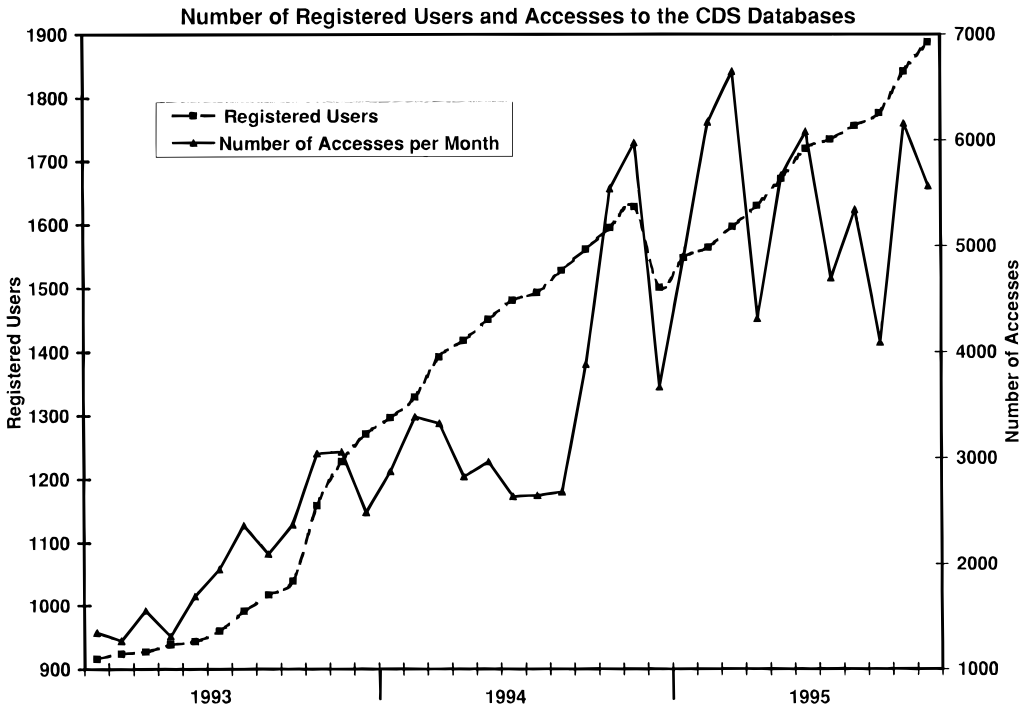

Figure 1.

UK academic community to a range of chemistry databases. Results obtained by using the Service may be freely published. Literature citations to the Chemical Database Service, and (where available) to the individual databases used in the study, should be included in such publications.

The Service currently has almost 1900 registered users from nearly 300 departments at 90 Universities. Of these, over 400 are active per month making around 6000 accesses to the Service. The user base has increased by 60% over the last two years and the trend continues, see Figure 1.

THE DATABASES AND SOFTWARE

The last 2 years has seen important new database facilities being added to the Service, in line with the recommendations of the Jennings Committee. A complete list of the databases available, together with their contents, is given in Table 1.

Major enhancements to the Service have been in the area of synthetic organic chemistry. Initially we acquired the REACCS[12] reaction information management software with associated databases.

Recently we have upgraded to the ISIS system and added the large, but selective, ChemInform RX database.[13] We do not aim to cover the literature comprehensively with reaction databases. Their purpose is to reflect new synthetic methods in organic chemistry with a focus on novel transformations.

ISIS has a client/server architecture. The client component, ISIS/Desktop, is used to generate queries and display results and runs on the chemist's local PC, Macintosh, or workstation. It interacts with the server, ISIS/Host, which performs the actual database searching and is mounted at Daresbury. We have acquired a UK wide academic site licence for ISIS/Desktop. Eligible users have to sign a conditions of use memorandum and are then allowed to electronically download the software from a disk area on the Daresbury system to their desktop computer. We are maintaining an overlap period with both REACCS and ISIS available to allow the community to make a smooth transition to ISIS.

Table 1. Databases Currently Available from the UK Chemical Database Service

Database	Description of contents	Owner/supplier
CSD	The Cambridge Structural Database. Crystal structure data for over 152,000 organic and organometallic compounds.	Cambridge Crystallographic Data Centre, UK
ICSD	Inorganic Crystal Structure Data File. Nearly 40,000 inorganic structures.	Fachinformationszentrum Karlsruhe, Germany
MDF	Metals Data File. The CRYSTMET Databank containing crystal structure data for about 54,000 metals, alloys and intermetallics.	National Research Council, Canada
PDB	The Brookhaven Protein Data Bank containing bibliographic and coordinate details for proteins and other biological macromolecules. There are currently over 4,400 coordinate sets.	US Government
CDIF	Crystal Data Identification File. Crystal class and unit cell data for over 200,000 crystal structures.	National Institute of Standards and Technology, USA
SpecInfo	Spectroscopy package. The database currently contains 99,059 ^{13}C NMR; 999 ^{15}N NMR; 856 ^{17}O NMR; 2,183 ^{31}P NMR; 1,825 ^{19}F NMR and 20,898 infra-red spectra.	Chemical Concepts GmbH, Germany
FNMR	A databank of about 6,000 ^{19}F NMR spectra and coupling constants.	Preston Scientific, UK
ELYS	Electrolyte Solutions Database. Thermodynamic and transport property data such as density, viscosity and diffusion coefficients. Currently contains about 10,000 entries.	Prof V.M.M. Lobo University of Coimbra, Portugal
ACDRX	Available Chemicals Directory, a database of suppliers of chemicals worldwide. Currently contains 180,000 unique compounds from 230 suppliers.	MDL Information Systems Inc. USA
Organic Reaction Databases		
REFLIB	Established literature data containing over 170,000 reactions from 1946 to 1991.	MDL Information Systems Inc. USA
CHC	Comprehensive Heterocyclic Chemistry, 42,376 reaction compendium.	MDL Information Systems Inc. USA
ORGSYN	Organic Synthesis, 5,016 reactions.	MDL Information Systems Inc. USA
JSM	Journal of Synthetic Methods, 1980-present, 48,776 reactions.	MDL Information Systems Inc. USA
CIRX	The ChemInform RX database. Current advancements since 1991, contains over 330,000 reactions	MDL Information Systems Inc. USA
SPG	Protecting Groups, 21,757 reactions.	Synopsys Scientific Systems Ltd. UK

Since the Jennings Committee recommendations our coverage of spectroscopy has been augmented by the addition of SpecInfo.[14] SpecInfo is a multi-technique spectroscopy software package with an associated high quality, validated database of ^{13}C and heteronuclear NMR and IR spectra. The package is designed to aid the chemist in spectrum interpretation and structure elucidation problems. It supports both substructure and spectral similarity searching, integrating differing types of spectral data. A novel feature of the software is a statistical analysis routine which makes use of the database to simulate spectra for given candidate structure. We have recently received SpecInfo 3.1, which has a much improved graphical interface as compared to version 2.1. SpecInfo 3.1 supports only X-Windows graphics. We thought at one stage this might be a problem, but a survey of our users has shown that practically all have access to a UNIX workstation or an X-Windows emulator on a PC or Macintosh.

The Service continues to provide support for crystallographic databases, and these are listed in Table 1. They are searchable using locally written software packages. In addition the Service now makes available the CCDC's own software system for use with the CSD.[15] This has many enhancements including the ability to search using both intra- and intermolecular geometric criteria. There are powerful molecular graphics display facilities, including the ability to produce crystallographic packing diagrams. A relatively recent addition to the CCDC system is VISTA, a package for statistical analysis of retrieved geometric parameters and graphical display of the results. The Service also supports the well-known molecular graphics packages Rasmol[16] and Xmol[17] together with the file format conversion package Babel.[18] There are also a variety of locally written utility packages including some for crystallographic file conversions which are not properly supported by Babel.

TRAINING AND SUPPORT

Help, support, training, and advice are key areas of provision for a central Service, since there is often no equivalent in small locally run systems. The Service employs three post-doctoral chemists to provide these functions. They are supported in their work by other Daresbury staff such as the User Interface Group, Operators, Systems Support, and Network Support.

A vital requirement for industrial advance is a supply of postgraduate chemists with appropriate training in chemical informatics. CDS provides training, free of charge, at Daresbury or University sites, when requested, which is tailored to the needs and resources of the University. Courses can range from introductory lectures giving an overview of the Service through to specialist courses on particular databases.

A large and increasing amount of material is now available via the Internet on the World Wide Web (WWW). The CDS's own Web pages (http://www.dl.ac.uk/CDS/cds.html), released in November 1994, cover the entire Service (details of databases available on the Service, newsletters, lists of user representatives, etc.) and point to other useful sites and information such as that provided by the National Information on Software Services (NISS) and the Combined Higher Education Software Team (CHEST). There is also information on other EPSRC funded Chemistry Service providers, chemistry software, programs, and publications. This is the principal way we give advice about chemical information not provided directly by the Service. There are about 300 accesses per day made to the CDS Web pages.

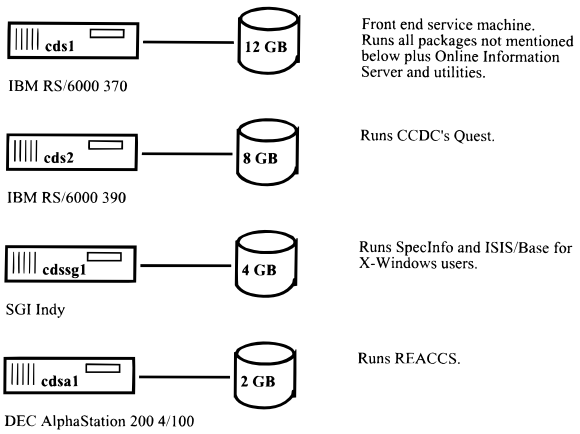

Figure 2. Current Chemical Database Service hardware.

A policy of making all documentation available on-line is being implemented. All major packages now have reference manuals on-line, as do some of the utilities. In addition, most of the packages have single sheet guides available to download and print locally. A number of manuals are also available to download. Importantly, an integrated hypertext based information source is now available—the On-line Information Server (OIS). This provides a single point of access for all of the available on-line information, including manuals. OIS uses html format files and users can view the information with a variety of WWW browsers. A CDS electronic mail discussion list is also available to the user community.

HARDWARE

Users login to the CDS via one machine (cds1.dl.ac.uk), but the task of running the various components may be farmed out to other machines (both VMS and UNIX[19]) as required. This process is transparent to the user. The "virtual computer" model, i.e., many computers appearing as one to the user, has several advantages. For example, the user has only one password to remember, a single area of disk space and only one computer operating system to deal with. The exception to this is ISIS, where the client software running on the user's desktop machine connects directly to the ISIS/Host machine. The virtual machine gives us a great deal of flexibility, allowing us to move systems from one machine to another to optimize load balancing. It is also easy to add new machine types if they are required for new database packages. The current configuration is shown in Figure 2.

In addition, Macintosh and IBM PC compatible clients are available to test the use of emulation and client software with the various packages.

CONCLUDING REMARKS

The nature of the Chemical Database Service has changed remarkably over the last two years to meet the needs of the academic community. The focus of the Service has moved from developing in house codes for database retrieval to mounting systems brought in from outside. Emphasis has also moved to support and training. We plan, however, to

develop an integrated graphical user interface, including online help, which will provide a seamless interface between component databases and graphical display programs.

Clearly new systems will become available and the Service must continue to adapt to new circumstances. We have to continually review what is desirable to provide via a central service and what is sensible to have on the chemist's desktop computer. There is a trend toward cheaper and more powerful computing which makes it feasible to mount systems on a desktop system which until quite recently could only be run on a large mainframe. This should be balanced against the improving quality of networks which allows faster and easier access to distributed, central facilities with pooled resources.

REFERENCES AND NOTES

(1) Theirs, A. H. M.; Leunissen, J. A. M.; Miller, T. M.; Schaftenaar, G.; Noordik, J. H. Computational Chemistry Network Services and User Interfacing. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 858−862.

(2) Wood, G. H.; Rodgers, J. R.; Gough, S. R. Operation of an International Data Center: Canadian Scientific Numeric Database Service. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 31−35.

(3) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C. F.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Version-3 and Version-4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187−204.

(4) Feldmann, R. J.; Milne, G. W. A.; Heller, S. R.; Fein, A.; Miller, J. A.; Koch, B. An Interactive Substructure Search System. *J. Chem. Inf. Comput. Sci.* **1977**, *17*, 157−163.

(5) Machin, P. A. The Chemical Databank System (CDS). In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; IUCr, Chester, 1987; pp 184−197.

(6) VAX and VMS are trademarks of Digital Equipment Corp., Maynard, MA.

(7) Bergerhoff, G.; Hundt, R.; Siever, R.; Brown, I. D. The Inorganic Crystal Structure Database. *J. Chem. Inf. Comput. Sci.* **1983**, *23*, 66−69.

(8) Bergerhoff, G.; Brown, I. D. Inorganic Crystal Structure Database. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; IUCr, Chester, 1987; pp 77−95.

(9) Rodgers, J. R.; Wood, G. H. NRCC Metals Crystallographic Data File (CRYSTMET). In *Crystallographic Databases*; Allen F. H., Bergerhoff, G., Sievers, R., Eds.; IUCr, Chester, 1987; pp 96−106.

(10) Mighell, A. D.; Stalich, J. K.; Himes, V. L. NBS Crystal Data − Database Description and Applications. In *Crystallographic Databases*, Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; IUCr, Chester, 1987; pp 134−143.

(11) Abola, E. E.; Bernstein, F. C.; Bryant, S. H.; Koetzle, T. F.; Weng, J. Protein Data Bank. In *Crystallographic Databases*; Allen, F. H., Bergerhoff, G., Sievers, R., Eds.; IUCr, Chester, 1987; pp 107−132.

(12) ISIS and REACCS are trademarks of MDL Information Systems, Inc., 14600 Catalina Street, San Leandro, CA 94577.

(13) ChemInform RX is compiled by FIZ Chemie GmbH, W-1000, Berlin, Germany.

(14) SpecInfo is a registered trademark of Chemical Concepts GmbH, Boschstrasse 12, D-69469 Weinheim, Germany.

(15) Allen, F. H.; Kennard, O. 3D Search and Research using the Cambridge Structural Database. *Chemical Design Automation News* **1993**, *8*, 130−137.

(16) Rasmol copyright Roger Sayle, BioMolecular Structures Group, Glaxo-Wellcome, Greenfold, Middlesex. Available electronically from ftp://ftp.docs.ed.ac.uk/pub/rasmol/.

(17) Xmol copyright Research Equipment Inc., dba Minnesota Supercomputer Center. Available electronically from ftp://ftp.msc.edu/pub/xmol/.

(18) Babel copyright Pat Walters and Matt Stahl, Dolata Research Group, Department of Chemistry, University of Arizona, Tucson, AZ 85721. Available electronically from ftp://joplin.biosci.arizona.edu/pub/babel/.

(19) UNIX is a trademark of AT&T, Murray Hill, NJ.

CI960015+