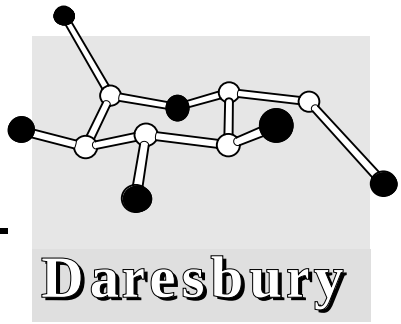
Chemical Database Service

NEWSLETTER

Issue 3 March 1995



Thank you to everyone who responded to the recent CDS survey. As well as giving us a lot of information about our user base, the replies gave us valuable statistics to use in our recent request for continued funding of the service. An analysis of the replies is given in the article on page 10.

The next CDS user forum will be held on May 24th at Daresbury.
All users are invited to attend. More details will be available soon.

Cds1 is now the main service machine for the Chemical Database Service, and users should only use DLVB for accessing SpecInfo and REACCS. New features available on cds1 include the Online Information Server (OIS) as well as chemical file format converters and display packages (see page 4).

The OIS provides access to the latest information on the available packages and utilities, online versions of manuals and newsletters (including this one) and a variety of other information about the service. See the article on page 2 for more details.

A new CDS users guide is now available. The online version can be read via the OIS, see page 2 for details.

The CDS World-Wide-Web server is now operational. The URL is http://www.dl.ac.uk/CDS/cds.html, though if you are logged onto cds1 and using X-windows, simply typing "xmosaic" will access it.

CDS queries can now be entered directly into the automated Computer Help Desk via the "dlquery" command. This saves effort at our end and should speed up the processing of your query.

Contents

Some CDS communications, including this newsletter, can be sent by electronic mail instead of paper. If you would prefer this method of communication, please send us your CDS username and a current email address.

The cds1 Online Information Server

The cds1 Online Information Server (OIS) is the main method of accessing the available online information provided by the Chemical Database Service. Currently it provides the following:

Up-to-date details of the packages and utilities available on cds1
Manuals for CSSR, ICSD, MDF, CDIF, CSDS, PDB and some utilities
Details of what help is available for users
Online newsletters - PDB and MDL newsletters as well as all CDS newsletters
Access to the CDS gopher and WWW servers
Listings of known errors in the databases

In addition there are smaller sections on specific topics such as file format conversions.

The documents are all written in hypertext format (HTML), which allows selectable links between the documents (hyperlinks). Thus the initial screen (see below) has a menu of highlighted options and selecting any of these calls up the appropriate document. Use of the HTML format means that the documents can be read using World-Wide-Web browsers. Currently there are two different ones available on cds1. These are **lynx**, a text based browser for VT100 compatible terminals, and **xmosaic** for X-windows systems. Since almost all terminal emulators are capable of emulating a VT100 terminal, this should cover everyone.

Text based access to OIS is started by simply typing "lcds1" at the UNIX % prompt. A screen similar to that shown below will then be displayed.

```
                                                        CDS1 Guide (p1 of 9)


                                  GUIDE TO CDS1

 Welcome to the cds1 online information server.

 Click or select here for introductory information.

 Information is available on the following topics, simply click on or
 select the topic you wish to view.

    * Help available
    * Packages and utilities available
    * Online manuals
    * Newsletters
    * File format conversion
    * Login messages
    * Training courses
    * Ordering documentation
-- press space for next page --
  Arrow keys: Up and Down to move. Right to follow a link; Left to go back.
H)elp O)ptions P)rint G)o M)ain screen Q)uit /=search [delete]=history list
```

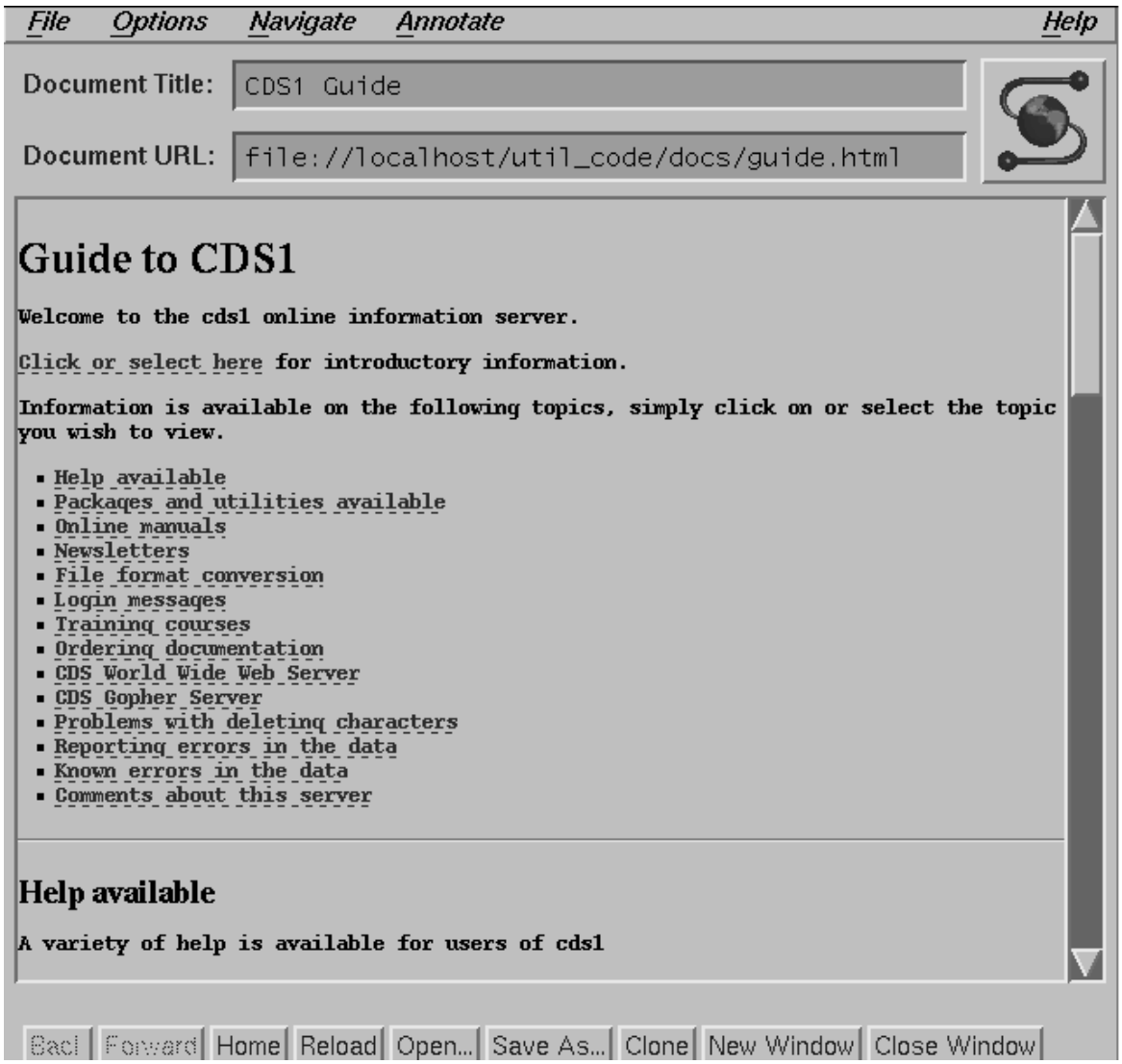
The bold text will appear highlighted in some way on your screen and is used to mark the hyperlinks to other documents. The up and down arrows will move the inverted text cursor between these links and pressing RETURN or the right arrow key will 'follow' a link and bring up the new document. The left arrow key will follow a link backwards to the previous document. Type "Q" to quit and return to the % prompt or "H" for the lynx online help.

Xmosaic based access to OIS is started by typing "xcds1" at the UNIX % prompt, though you need to have set the DISPLAY environment variable beforehand with the following command:

setenv DISPLAY <IP address>:0

where <IP address> is the IP or Internet address of your terminal display. This variable may have already been set as part of your login procedure.

Once started you will see the following screen:



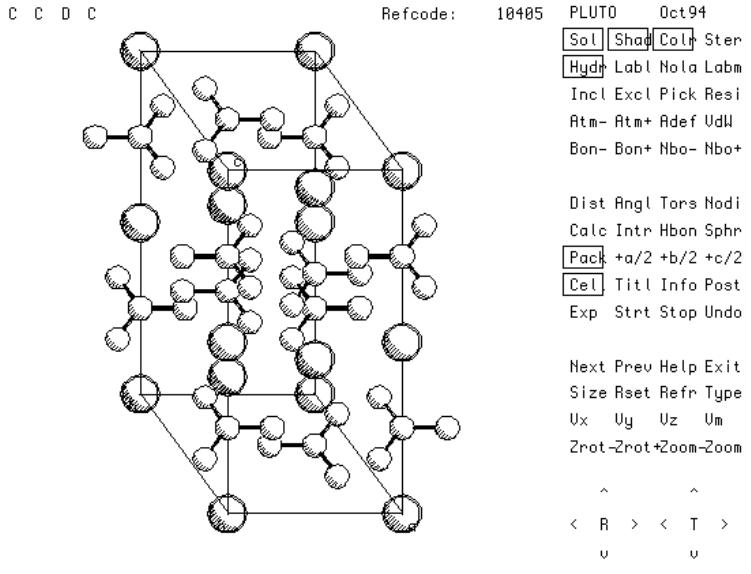
As with the lynx browser, the highlighted bits of text are hyperlinks and are selected by simply clicking on them with the mouse. Back navigation is possible by clicking on the 'Back' button at the bottom of the screen.

Crystallographic display

The various packages on cds1 handle files of crystallographic coordinate information in differing formats. The local retrieval systems (CSSR, ICSD and MDF) produce data in .XR format, which can be read by the utility programs, **CRAD** and **VIEW** (on DLVB). The Cambridge Structural Database System (CSDS) supports its own formats with the retrieval program, QUEST, currently generating coordinate files in either FDAT or Model format. QUEST's companion crystallographic manipulation, display and geometry calculation program, **PLUTO**, reads its data in FDAT format. The coordinates data sets we make available for the Brookhaven Protein Data Base are in yet another special format (PDB).

The .XR, FDAT and PDB formats can be read by many molecular modelling and visualisation packages, but we know of none which handles all three correctly. Also many of the packages available on workstations, PCs and MACs have their own specific preferred file formats. We currently provide the molecular display programs **xmol** and **rasmol** on cds1 in addition to PLUTO. Xmol and rasmol are both X-based systems, but PLUTO supports a number of graphics devices. We do not intend to port the old display program, VIEW, from DLVB to cds1 as all its functionality is available within the new systems.

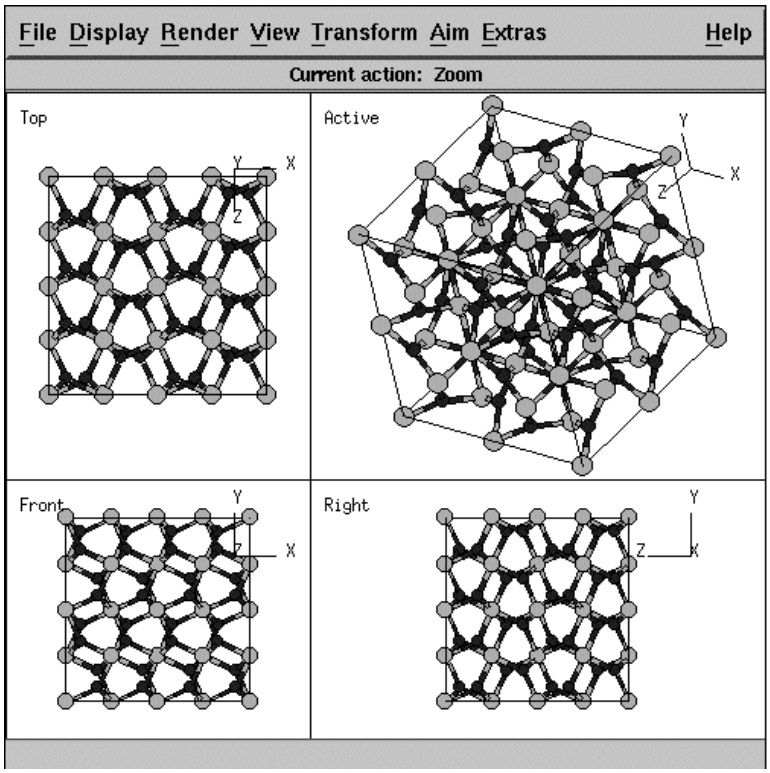
Both xmol and rasmol will read data in a variety of formats and xmol has a format conversion capability. They both read PDB but unfortunately currently not .XR and FDAT formats. To extend the available file conversion facilities we have installed the package **babel** on cds1. Babel will convert between a wider range of formats. In particular, it will convert to and from .XR format but currently only from FDAT. To simplify use of the packages we have set up two new utilities, **plotxr** and **rasplotxr**, which automatically convert a .XR file to an intermediate XYZ format using babel and start up either xmol or rasmol with the XYZ file as input. Two additional format converters, **cod2cam** and **num2cam**, produce FDAT format output. The input for cod2cam is a file of refcodes from QUEST or CSSR, whilst num2cam (a replacement for **getfdat**)takes a file of CSSR reference numbers as input.



This is a PLUTO display of a calcite ($CaCO_3$) structure retrieved using ICSD. The atom coordinates in .XR format were converted to FDAT format using xr2cam. There is no chemical connectivity information in the ICSD database and in this case there were only three atoms in the asymmetric unit. The PLUTO options Pack, Calc and Cell were used to expand out the structure, generate the chemical connectivity and show the unit cell. The cell was rotated to give a clearer view.

Some care should be exercised when using babel as it does not handle some aspects of crystallography very well. For instance, **do not use** babel to convert from an FDAT to a .XR file and then use that for input into CRAD - the results will be nonsense. We have written two utilities, **xr2cam** and **cam2xr**, which convert between .XR and FDAT format whilst retaining full crystallographic symmetry, and they should be used in preference to babel.

The **fillxr** utility program generates all the atoms needed to fill up a unit cell or block of units cells. In general, the output files produced by CSSR, ICSD and MDF include only coordinates corresponding to the asymmetric unit. To get all atoms in the unit cell you need to apply the symmetry operations for the crystal space group. Fillxr takes .XR files as input and outputs the expanded coordinate list in the same format. It is particularly useful if you wish to display crystal packing diagrams using the xmol or rasmol packages, neither of which have the capability of performing crystallographic transformations.



This shows the use of xmol with the split screen display option. The structure is taken from ICSD. It is from a theoretical study and is a hypothetical high pressure structure for SiO_2. It has the cubic FeS_2 structure type with 6-coordinate silicon. There are only two atoms in the asymmetric unit. Fillxr was used to the generate the coordinates for a block consisting of 8 unit cells and the structure viewed using the plotxr utility. The active window shows the view along the 3-fold axis

Summary of the ways available to display different format files

PDB format:		use rasmol or xmol
	or	use babel to convert to .XR format and then run xr2cam and use PLUTO (restricted to less than 1000 atoms)
.XR format:		use plotxr or rasplotxr
	or	run through xr2cam and use PLUTO
FDAT format:		use PLUTO
	or	run through cam2xr and use plotxr or rasplotxr

Brookhaven Protein Databank (PDB) on cds1

The Brookhaven Protein Databank contains bibliographic and coordinate details for over 3200 proteins and other biological macromolecules. It has been available on DLVB for some years, but has now been ported to cds1. **Functionality has been increased and the whole database is now updated weekly** (if new data is available). In particular X-window based display utilities are available to view the 3D structures.

It is available to registered PDB users. If you wish to register contact the Computer Help Desk / UIG at the address given on the back page.

Accessing the data

There are now four different ways to search the data:

1. The brook command.

This is an interactive menu-driven system with online help. It allows searching of the header file as well as the actual PDB files themselves. It also has a number of additional features. Typing "brook" starts the system and presents you with the following options:

```
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*          WELCOME TO THE BROOKHAVEN STRUCTURAL DATABASE                *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *
*                                                                       *
*        A......List.The.Brookhaven.Directory                           *
*        B......List.A.Group.Of.PDB.Files                               *
*        C......List.SOURCE.&.COMPOUND.Of.All.PDB.Files                 *
*        D......List.SOURCE.&.COMPOUND.Of.Group.Of.PDB.Files            *
*        E......Look.At.A.PDB.File.(with co-ordinates)                  *
*        F......Look.At.A.PDB.File.(without co-ordinates)               *
*        G......Copy.A.File.To.Your.Own.Disc.Space                      *
*        I......Look.At.The.PDB.Index.File                              *
*        J......Search.The.PDB.Index.File                               *
*        K......Search.For.A.String.In.All.PDB.Files                    *
*        L......Look.At.The.PDB.Index.File.Using.An.Editor              *
*        M......Look.At.Any.PDB.File.Using.An.Editor                    *
*                                                                       *
*               Type    Q   to   QUIT                                   *
*                       ?   for  HELP                                   *
* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *

Please type in your choice and hit return__
```

Most of the options are self explanatory, though the online help function gives more details for each. The header file, which can be searched, is a single file made up of summary information about the contents of each PDB file, including file name, protein and source, authors and reference.

2 The pdbfind command

This is a simple command, based on the UNIX grep command, to search the header file for a simple string. For example, typing "pdbfind phosphate" at the UNIX % prompt displays every entry in the header file that contains the word 'phosphate'.

3 Searching PDB via QUEST

The bibliographical and sequence information, but not the 3D coordinate information, is searchable via QUEST. This option is only available to registered CSD users (this includes anyone registered for CSSR). QUEST searches a different version of the PDB data, which we only receive every 6 months with the Cambridge Structural Database updates.

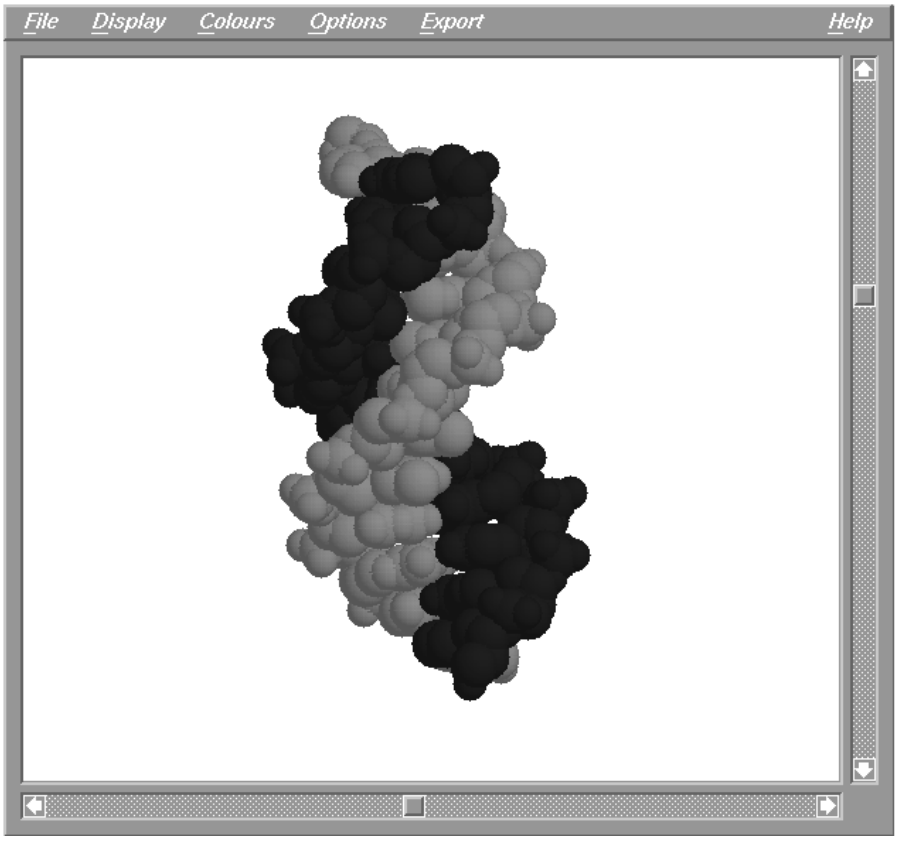
To search the PDB files, use "questp" in place of "quest" when starting the program. More details are given in the CSDS Getting Started Manual.

4 Searching PDB via gopher

The PDB files are also searchable via gopher directly from Brookhaven. Type "gopherp" or "xgopherp" to start. This implementation is very similar to the 'brook' command described above. This method may be slow when the network is busy.

Display of PDB files

Once found, PDB files can be displayed with **rasmol**, a 3D molecular display program.



This example shows a section of DNA colour coded by chain using the spacefilling display mode. Other display formats available include wireframe, ball and stick, and backbone.

Rasmol is an X-windows application and is started by typing "rasmol filename" where filename is the name of the PDB file you wish to display. The program has a variety of display formats and allows rotation of the structure in 3 dimensions. A number of output formats, including GIF and PostScript are available. Online documentation is provided for the program and a quick reference guide, in PostScript format, can be downloaded from /util_code/ftp/rasmol.

Other display options are discussed in the article on crystallographic display on page 4 of this newsletter.

Downloading PDB files

Individual PDB files can be downloaded to your local machine using the file transfer utility, **ftp**, for users on the Internet. Details of using ftp can be found in the Guide to UNIX on cds1 or by typing "man ftp". In addition we have a version of **kermit** available for non-Internet users.

Cds1 problems

The move of the service from VAX/VMS to UNIX has given rise to a variety of problems for users. Some of the common ones and their solutions are listed below.

Deleting

UNIX uses ^H (control-H) for deleting as opposed to the VAX which used DEL. If your terminal was set up for a VAX you may find that you cannot delete characters properly on cds1. Many emulators will allow you to change this setting, but an alternative is to add the following line to your .login file:

stty erase "^?"

Note: there is a mistake in the Guide to UNIX on cds1. On pages 5 and 7 the erase key is given as DEL. This is incorrect, it is actually ^H.

Crashing out of EDT

When you use EDT via X-windows you get a new window to edit your file. When you exit you need to type the ^Z (control-Z) in this window. If the focus has been shifted back to the original xterm window, typing "^Z" will suspend the EDT process but the file will not be saved and the EDT process will not be stopped. If you type "^Z" in the wrong place by mistake, type "fg" to bring the EDT process back, but also see the section on logging off with suspended jobs, below.

View and fill

View and fill were programs available on DLVB. On cds1, fill is called fillxr and the function of view has been taken over by PLUTO and other display utilities. However, there are UNIX utilities called view and fill, so if you type "view" or "fill" by mistake you will start up a new process. To exit from view, type "qq" followed by the return key. To exit from fill, simply type "^C" (control-C).

Core files

You may occasionally find a file called 'core' in your user space. This is a core dump file and is created when a program crashes on you. This file is of no use; it merely uses up some of your disk quota. It can be removed by typing "rm core".

Logging off with suspended jobs

If you have any suspended jobs (caused by crashing out of EDT, for example) and you try to logout, the computer will tell you that there are stopped jobs and will return to the % prompt. If you immediately try to logout again, you can. However the suspended jobs will not necessarily be terminated, and in the case of EDT will consume CPU time. You can use the "fg" command to bring the suspended jobs back and terminate them correctly before logging out. If you do logout, you can remove the jobs with the "kill" command. Log back on and type "ps -u xxxx", where xxxx is your username. You will see and output like this:

```
UID   PID   TTY   TIME  CMD
11653 15773 pts/1 0:00  csh
11653 18849 pts/1 0:00  ps
11653 21864 pts/8 1:39  edt.x
```

To remove the suspended job type "kill -9 PID", where PID is the process number of the job, listed in the PID column, in this case 21864.

X25 Removal

From May, the recommended way to access DLVB will be via the Internet. All JANET accesses will come in via a gateway, which should be transparent to users. The TRANSFER and PAD commands will no longer work (use FTP and TELNET or RLOGIN instead) and there will be changes to the mail facility. All users are advised to use their cds1 accounts for email and all email to any of the CDS staff should use the addresses given on the back of this newsletter. If you experience any problems, contact the Computer Help Desk at the address given on the back.

CDS Questionnaire Analysis

Thank you to everyone who replied to the CDS questionnaire last year. We received 279 replies (about 20%) and a lot of useful information about our user base as well a variety of specific comments. The overall message from the questionnaire is that the Chemical Database Service is an important part of UK academic research in chemistry and provides a good service to its users. The bottle of champagne was won by Dr. Jones from Keele.

We were pleased to learn that nearly 90% of you rated the service as good or very good overall, with only one person rating us as poor. The quality of help/support and the availability of databases were also rated highly (84% good or very good), whilst online response time was rated lower (70% good or very good). Online response time should be improving now, with all of the service except REACCS and SpecInfo on the new machine, cds1.

REACCS and the Cambridge Structural Database (QUEST and CSSR) were the most popular databases and the most important. Of the other databases only FNMR and ELYS were used by less than 25% of people and not rated as very important.

We received a large number of comments and queries. We have attempted to reply to people who asked specific questions, apologies to anyone we have missed, please let us know if you think you should have had a reply. Areas which could be improved have been highlighted and will be addressed.

Many of the comments were requests for improvements or new services. Some of these have already been implemented (e.g. PLUTON, babel and the FDAT to XR format converter, cam2xr) whilst others are in the process of being improved (such as the online help on cds1 and updating of some of the manuals). The requests for new databases will all be examined and where possible we will get an evaluation version for users to try out.

The most common hardware and software used to access the service are PCs and Tektronix emulators, though about 2/3 of users have access to X-windows. X-windows will become more important in the future, as access to some packages is only possible via X-windows. Already some the utilities, such as rasmol and xmol, are X-windows only and we would expect this trend to increase.

Our training is valued by users and in strong demand. Some 10% of those who responded had already attended a CDS training event and about half of the respondents would like to. Our training program is ongoing, with courses being planned for Sheffield, Bristol, Bath and UCL. Details will be advertised online when the dates have been fixed.

Approximately half of the people who replied said that use of the service had generated new leads or approaches in their research work. A similar number had published papers making use of the service.

Summary of databases available from the Chemical Database Service

A brief description of all the databases currently available from the CDS at Daresbury.

REACCS A chemical reaction information management system allowing search, retrieval and display of molecules, reactions and their associated data. Currently contains 244,425 searchable reactions and supplier information on 170,593 different compounds.

SPEC SpecInfo is a multi-technique spectroscopic database package intended to cover NMR, IR and mass spectra. The package is designed to aid the chemist in spectral interpretation and structure elucidation problems. The database currently contains 66,063 ^{13}C NMR; 1,000 ^{15}N NMR; 856 ^{17}O NMR; 2,182 ^{31}P NMR; 1,830 ^{19}F NMR and 28,464 infra-red spectra. It does not, at present, contain any mass spectra.

CSD The Cambridge Structural Database. Crystal structure data for 126,353 organic and organometallic compounds. Currently available under both Cambridge's QUEST retrieval software and our in-house software (CSSR). The QUEST software has a number of features not available in CSSR, such as 3D searching.

ICSD Inorganic Crystal Structure Data File. About 38,000 inorganic structures - the companion file to the Cambridge organic file.

MDF Metals Data File. Crystal structure data for about 52,500 metals, alloys and intermetallics.

PDB The Brookhaven Protein Data Bank containing bibliographic and coordinate details for proteins and other biological macromolecules. There are currently over 3,200 coordinate sets.

CDIF Crystal Data Identification File. Crystal class and unit cell data for 197,612 crystal structures.

FNMR A databank of about 6,000 ^{19}F NMR spectra and coupling constants.

ELYS Electrolyte Solutions Database. Thermodynamic and transport property data such as density, viscosity and diffusion coefficients. Currently contains about 11,000 entries.

Information

General queries and registration of new accounts:

Computer Help Desk / User Interface Group (UIG)
Daresbury Laboratory
Daresbury, Warrington
WA4 4AD
(01925) 603351
Email uig@daresbury.ac.uk

Hours are:
Mon-Thur 9:00-12:00, 13:00-17:00
Fri 9:00-12:00, 13:00-16:00

Otherwise 24 hour answer phone.

General database queries:

Chemical Database Service (CDS)
Room C18, Daresbury Laboratory
Daresbury, Warrington
WA4 4AD
(01925) 603162
Email cdsbb@daresbury.ac.uk

For REACCS specific problems:

Dr. D. Parkin
Email
D.Parkin@daresbury.ac.uk
(01925) 603162

For SpecInfo specific problems:

Dr. D.A. Fletcher
Email
D.A.Fletcher@daresbury.ac.uk
(01925) 603162

For QUEST specific problems:

Dr. R. F. McMeeking
Email
R.F.McMeeking@daresbury.ac.uk
(01925) 603162

Documentation:

Documentation is available through the orderdoc utility or from UIG at the above address.

Access:

The service is available free of charge to UK academic research groups through Janet or Internet.
Individual user names are issued on request to UK academics.

cds1:

Internet number: 148.79.160.113
Internet name: cds1.dl.ac.uk
DTE address: 00000100026012
NRS name:
UK.AC.DARESBURY.CDS1

DLVB:

Internet number: 148.79.160.7
Internet name: dlvb.dl.ac.uk
DTE address: 000001003000
NRS name:
UK.AC.DARESBURY.DLVB

Comments:

All comments, questions and suggestions about this newsletter should be sent to:

Dr. D.A. Fletcher
Chemical Database Service (CDS)
Room C18, Daresbury Laboratory
Daresbury
Warrington
WA4 4AD
(01925) 603162

Email
D.A.Fletcher@daresbury.ac.uk