CCLRC DARESBURY LABORATORY
EPSRC NATIONAL CHEMICAL DATABASE SERVICE
ANNUAL REPORT 2005/6

1. REPORT OF THE PAST YEAR

1.1 The details of the EPSRC *Individual Grant Review* for our previous Grant have been made available to us. The overall rating for *Grant GR/R01866/01* was "Outstanding".

1.2 The CDS group operates as a well integrated four person team. It retains the same complement of full time staff (all have permanent contracts with the CCLRC). They include R.F. McMeeking (Group Leader), D. Parkin, D.A. Fletcher, and D. Osa-Edoh.

1.3 The *Management Advisory Panel* provides valuable input to the Service. Details of the membership of the current panel together with its *Terms of Reference* are given in *Appendix 1*. There was a face to face meeting of the Panel in September 2005. We made available online copies of the minutes of MAP meetings held since March 2004 at the URL http://cds.dl.ac.uk/map/recent_mins.html.

1.4 Universities outreach through the *CDS Roadshow* visits programme continued during the report year. The total number of visits completed for the current *Grant* cycle has now reached 44. The quota stipulated by the *Grant Appraisal Panel* prior to awarding the current 3 years funding was 45-50 visits in all.

1.5 We still offer *Roadshow* visits, but now mainly target bioscience & chemical engineering departments. The aim of the outreach programme is also shifting from simple presence and publicity to an emphasis on specific training. This involves providing tailored "hands on" training for the major database systems. In addition we assist local representatives in providing their own courses (see *Appendix 6* for further details).

1.6 All major aspects of the Service have shown sustained growth during the report period. Fuller information about the major database components currently supported, including usage details, is provided in *Section 2* and *Appendix 4*.

1.7 There is an inevitable lead time for any new system to be fully appreciated by the community. However, uptake of the DETHERM thermophysical database is now encouraging (see *Section 2*). Undoubtedly this has been helped by the various *CDS Roadshow* visits, especially those to Chemical Engineering departments.

1.8 The Chemical Engineering representative on our MAP, Mark Biggs, agreed to send letters to heads of departments, and this has also had a discernible effect on DETHERM uptake. In addition an article for *the Chemical Engineer* journal, to be written by an active DETHERM user, is in the pipeline.

1.9 An online User Survey was conducted in Autumn 2005. Registered users were alerted via email with a link to a simple web based form. The survey was available for completion over a period of a month. For details of the survey see *Section 5*.

1.10 We have removed dependency on a number legacy codes. For instance the old CSSR & ICSD codes are not required as background processes underpinning CrystalWeb. This is a key part of our plans for modular implementation and ease of future development for this web based system.

1.11 The CrystalWeb interface allows ready access to the full collection of crystallographic databases by a common web browser interface. It has been further refined and developed. In particular closer integration with the ICSD-WWW web interface is well established. As a result of this project a number of problems associated with the ICSD-WWW system, as supplied by the producers, have been identified and corrected.

1.12 We have improved CrystalWeb structure display (it is superior to that provided by ICSD-WWW for certain inorganic structures). There is better formula searching and structure export facilities have been enhanced with many more file formats available. These include Smiles strings and InChI code representations.

1.13 The Screening Compound database (compounds for activity screening and diversity synthesis) implemented by the CDS continues to grow with data now provided by 24 suppliers and the compounds available rising by 50% to *4.5M* during the *Report Period*. Datasets are received directly from the individual suppliers.

1.14 The Available Chemicals Directory (ACD) data is supplied by MDL and now has direct links to their Material Safety Data Sheets (MSDS) for around 15,000 compounds in the database.

1.15 We have produced a java applet version of the ACD web interface. It does not require use of the ChimePro plugin and works with a wider range of browsers and operating systems (including Mac OSX and Linux). The system also displays the 3-D (Corina generated) structures for retrieved entries. It also provides a facility to download the structural coordinates in Molfile and potentially other formats.

1.16 A new proton nmr dataset, containing over 53,000 assigned spectra has been added to the Service. These are all available via the SpecSurf web based system. This more than doubles the number of fully assigned H nmr spectra available, and brings the total number of spectra of all types within the SpecInfo collection to over 420,000. In addition two new hetero-nuclear collections have been added. There is a set containing 14,378 assigned ^{31}P nmr together with a set containing 23,600 assigned ^{19}F nmr spectra and structures.

1.17 The IUPAC International Chemical Identifier (InChI) continues to be actively developed and is expected to revolutionise chemical information retrieval, cheminformatics, and data mining. The latest version of the InChI code, which is now freely available, not only handles

organic, covalent structures but also inorganic and organometallic compounds. We have now done a good deal of work in this area but plan more in the future. For details see *Section 6.*

1.18 The service computer systems performed well with no major technical problems. A new Intel server has been acquired and installed with the Linux operating system. This machine (cds10) is now our main Oracle server. The old Oracle server (cds6) has been redeployed as an Oracle development machine. It can also act as standby machine in the event of hardware failures for cds10 or Oracle software upgrade problems.

1.19 We have arranged for a number of trials for prospective new database systems for the Service (see *Section 6*). These include the ACD/Labs I-Lab spectral and properties predictor. We have acquired a Microsoft Windows platform to run this software, and are currently working on the details of integrating this within the main CDS service environment.

1.20 A meeting was set up by the Service at Swindon to discuss issues relating to possible national arrangement to give cost effective access to the CAS SciFinder system by the UK academic community. This was attended by RFM & DP from the Service, David Knight (chair of the *Grant Appraisal Panel*, representing the academic community), and representatives from JISC, CHEST/EduServ and the EPSRC.

2. USAGE

2.1 Use of the Service is closely monitored, as is its efficiency in terms of a number of defined performance indicators (see *Appendix 5*). A monthly digest is sent to our Management Advisory Panel (MAP) and the EPSRC. Performance targets are kept under review and tightened up where appropriate. An archive of the digests since January 2004 is available at http://cds.dl.ac.uk/map/performance.html.

2.2. 726 new users were registered during the period 01/04/05 to 31/03/06. This represents a 30% increase in registrations compared to 2 years ago (pre *Roadshow*). After removing 533 "dead" usernames the total number of registered users at the end of March 2006 was 4,209 – an overall increase of 5%. The increase in registered users from engineering departments during the same period was 35%. For additional details see *Appendix 2.*

2.3 There has been sustained growth in the scope of the Service. Our existing datasets continue to grow at the rate of around 6% per annum. For instance there were over 40K new crystal structures and nearly 60K new reactions (for the year period 01/04/05 to 31/3/06). The Service now provides access to nearly 1.5M searchable reactions, and around 4.5M compounds. For a summary of currently available components see *Appendix 4.*

2.4 Use of all databases has risen steadily (see *Appendix 2*). There has been a change in pattern of usage partly related to faster machine response. We believe the users are making more efficient use of the more powerful software which has become available, and there has been greater uptake of the improved web browser access which is now available. This typically results in more intensive single sessions involving multiple searches.

2.5 A case in point is the Cambridge Crystallographic Database where the newer ConQuest software is truly interactive as opposed to QUEST, which was more like a pseudo-batch system. Conquest now accounts for 2/3 of the accesses and ¾ of the users to the Cambridge Crystallographic data compared to the legacy Quest and CSSR systems.

2.6 All uptime on the Service computers is available for CDS use. In general, the servers run 24 hours per day and 7 days per week. During the full report period there was effectively 100% service availability with no unscheduled downtime. There were no major instances of note except for instances of external network (JANET) problems, which may have affected certain users.

2.7 Accesses per month to the core database systems oscillated between 5,235 and 9,815 with an average of 6,981, representing an increase of 9.6% over the previous report period. The number of active users within any given month varied between 622 and 920 with an average of 777, representing an increase of 11.5%. All subject areas showed an increase in the number of active users and database accesses. DETHERM use had the biggest increase (165% increase in active users and 91% in accesses). Again fuller details are given in *Appendix 2*.

2.8 In this Report we do not attempt to match source of funding for individual users with their CDS access, but note that most use comes from departments with EPSRC Chemistry grants. The registration form now requests "main research funding source (or graduate funding body)". Currently around 70% of new users provide their funding details on registration. Such details can now also be added or updated via a simple web interface, and we will issue a general call for all users to do so as part a general exercise to update all registration details.

2.9 The use of the CDS amongst all chemistry departments continues to correlate well with the grade allotted by the Research Assessment Exercise (see *Appendix 3*). Users from chemistry departments with a 2001 RAE rating of 4 or higher accounted for 90% of all registered and 91% of all active users and 92.4% of all accesses from chemistry departments.

2.10 Our policy continues to be to grant access for eligible use to all UK academics who apply. User registration is now exclusively via an online form on the CDS web site. This is efficient in term of operational procedure and there are no discernible bottlenecks relating to resource availability. Registrations are completed and returned on the same day as receipt of the associated authorisation slip. A fax option as an alternative to receiving registrations by posting was introduced in 2002 and now accounts for ~60% of all application received. This excludes registrations at *Roadshows* - where all applications are fully completed there and then.

2.11 Users who have a valid "ac.uk" address can choose to be sent their username and password by email rather than post was introduced in 2004. This further reduces the turn round time and is also proving very popular, with over 90% of users being dealt with this way. Just over a third of registrations during the report period were from staff or research fellows and just under half were from graduate students.

3. PUBLICITY

3.1 As already discussed in *Section 1* the *CDS Roadshow* programme was a major element of our publicity strategy during the report period.

3.2 The CDS web site (http://cds.dl.ac.uk) continues to be a key component of the Service and a major vehicle for publicity. It served on average, 9,151 distinct requesting sites per month day during the period (01/04/05 to 31/03/06). There was an average of 3,853 successful requests for pages per day, an increase of 6% over the previous year.

3.3 There has been another overhaul of the web site including a redesign of the homepage with increased use of pull down menus. We have also attempted to better integrate the information on the web with that provided during *Roadshow* visits. For instance the databases are now presented grouped under more than the original four categories. Accordingly the organic chemistry area has been split into *organic synthesis* and *procuring chemicals*. There is also linking between all 3D *molecular* databases (in addition to the crystallography ones) such as ACD and NCI given they now also contain (Corina calculated) 3D structures.

3.4 Another important way of distributing information about new Service features to registered users is via broadcast email messages. Most users find this mechanism *useful* or better (see Survey Results in *Appendix 8*). Broadcast messages usually include links to fuller details on our web site. We invest significant effort in keeping our email lists accurate and up-to-date (using "ac.uk" addresses [over 91% of users] where practicable). This helps us to identify those users who move site and/or leave the academic community.

3.5 The web site homepage has also been modified such that it gives greater prominence to current news items. Automatic procedures have been put in place such that items are ordered according to their importance and currency, and "stale" news is removed when it is no longer of relevance. The *News* and broadcast *Email Archives* have also been substantially improved. The User Survey shows that 2/3 of users find the news items *useful* or better.

3.6 The Service produces a biannual *CDS Newsletters* (latest editions dated October 2005 and April 2006). The current version and back issues from 1994 are also available on the CDS web site (http://cds.dl.ac.uk/newsletters). The *Newsletter* is sent to all registered users and over 90% elect to receive it electronically (latterly via the web link) rather than in print format.

3.7 Traditionally users have been sent a copy of the *CDS User's Guide* and the latest copy of the *Newsletter* when they first register. We no longer do this for those who opt to receive their username electronically (currently about 90% of new registrations), but point to the location of this material on the CDS web site.

3.8 Publicity material continues to be sent to User Representatives for display within departments. There is a *CDS Overview Powerpoint* show (also used in standard *Roadshow* visits), which is available for display and/or downloading from the CDS web site at http://cds.dl.ac.uk/overview.ppt. This is advertised as being made free for copying and distributing within departments.

3.9 Members of the CDS team attended 6 academic conferences and similar events. DP remains as a committee member of the RSC Chemical Information Group (CIG) and attended all its meetings in this capacity. DP is also on the organising committee of the International Conference on Chemical Structures which held a successful seventh meeting in Noordwijkerhout, The Netherlands in June 2005. There was also a CDS presence at the British Crystallographic Association Meetings.

3.10 A *User Forum* was held in July 2005. This allowed useful discussions with attendees, which has provided us with valuable feedback from the wider community.

4. RESEARCH HIGHLIGHTS

4.1 Seven research highlights were used this year, and these provide a snapshot of *some* of the fields of study which have benefited from use of the Service. The articles themselves are included in the separate *Research Highlights* document with this Report.

4.2 It is in the nature of things that Service personnel do not have direct contact with the majority of the work carried out. We ask users to acknowledge the Service in their published work, and we also have a 1996 review article which we encourage them to cite*. This allows us to track published paper using the *ISI Web of Science*. There were 66 citations from journals published in the 2005 calendar year. Acknowledgements are not as easy to pick up, but an exercise using Google indicated approximately twice the number acknowledged use of the Service as cited our 1996 article.

4.3 The User Survey included a question as to how many papers the respondee had published in the last 3 years citing or acknowledging use of the CDS. This recorded a total of 211 papers. It could be argued, perhaps optimistically, that this would scale up to around 1300 papers per year (excluding duplication) if applied to the total community. Respondees recorded that they had published 489 papers in the last 3 years which did not cite or acknowledge, even though they did make use of the CDS in their work. This implies we are considerable under represented by the actual citations made.

5. CDS USER SURVEY

5.1 As indicated in *Section 1* the Service conducted an online User Survey in Autumn 2005. Registered users were alerted via email with a link to a simple web based form. The survey was available for completion over a period of a month. Full details of the questions and a

---

* "The United Kingdom Chemical Database Service", D.A. Fletcher, R.F. McMeeking, D. Parkin., *J. Chem. Inf. Comput. Sci.* (1996), 36, 746-749
In addition there is a more recent review of the Service in the book "*Cheminformatics Developments: History, Reviews and Current Research*" (Ed. J. H. Noordik), IOS Press, Amsterdam, Chapter 2, pp 37-67, 2004.
Links to both articles in electronic format are available from the URL http://cds.dl.ac.uk/cds/acknowledge.html

breakdown of the responses, including user added comments are given in *Appendix 8.* Details are also available from the URL http://cds.dl.ac.uk/hd/survey_results2005.pdf.

5.2 A total of 3,621 emails were sent and 195 forms were completed. When asked "How useful is the CDS to your work?" those that replied said it was useful (21%) or very useful (35%) while 25% said it was vital. Over ¾ of replies were also positive when asked if CDS should continue to acquire and support specialist databases.

5.3 Results of the survey gave us useful information on the preferences the community had for receiving new about the Service. It also indicated that there is general enthusiasm for the proposed simple global database interface we are developing. Users also endorsed the case for maintaining a number of specialist database which may have an appeal for a subset of users.

5.4 The survey also included spaces for respondees to include comments about specific topics and the Service in general. Comments on the issue of citing the Service raised a number of points, which have been discussed in *Section 4.* The majority of comments on the general value of the Service were very positive. Use of the Service complements access to major systems such as CAS SciFinder and Beilstein/CrossFire.

5.5 It is also clear that a section of organic chemists who are Mac users have been experiencing problem. We are aware that ISIS support for the Mac/OSX operating has been sub-optimal, and we are actively addressing this problem. For instance we already produced a new platform independent interface for the Available Chemicals Directory (see *Section 1.15*).

6. FUTURE PLANS

6.1 The Service has been working in-house with InChI related software, and has applied the technology to molecular structures included databases ranging from the Available Chemicals Directory to the Cambridge Structural Database. We plan to extract data/meta-data for all molecules in the CDS collection of databases and also extracting such data from any external sources which become available to us.

6.2 eScience technologies will assist us this aim, but InChI codes will provide the main glue for the integration process. The aim is to be aware of and implement the best standards of interoperability between distributed systems. We are, however, mindful of not making the Service a hostage to unproven technologies. In the first instance we plan to use well established web services techniques. We have already explored commercial solutions such as IBM's WebSphere information integrator.

6.3 Commercial solutions have real advantages, but a big problem is cost. We are seeking funding in conjunction with IBM and others to cover these. We have already made two bids which unfortunately have been unsuccessful. The first was in response to a JISC call with an emphasis on Shibolleth authentication. The second was a DTI call for innovative technology.

This involved other commercial and academic partners and would have supported an OpenQSAR project, a core aspect is use of the CDS federated database system.

6.4 ACD/ILab (ACD Interactive Lab) is a web-based gateway to spectroscopic information, compound name generation, and property prediction. We believe that this system has the potential to complement our current SpecInfo system and to some extent DETHERM. The trial is scheduled for summer 2006. The full database and software system will be mounted locally (see *Section 1.19*). Coupling with local ACD/Labs systems which may be installed in UK departments and other national services may also be a possibility in the future.

6.5 We have also scheduled a trial for the SPRESI system for chemical structure and reaction information. This a large database system with over 5 million structures, 3.7 million reactions, and 28 million pieces of factual information. The software has been developed by InfoChem GmbH. It has a very impressive, easy to use, platform independent web browser interface. Access to this technology to provide access to other components of the Service may be possible given future interactions with InfoChem.