CCLRC DARESBURY LABORATORY
EPSRC NATIONAL CHEMICAL DATABASE SERVICE
VISION FOR THE FUTURE OF THE SERVICE

What a central Chemical Database Service can/should do and why should the UK have one.

Things the current CDS provides

- Access to a wide variety of high quality databases in the following areas of chemistry; structural chemistry, synthetic organic chemistry, chemical procurement, spectroscopy and physical properties.
- Powerful, yet user friendly, access mechanisms to these databases.
- A helpdesk and support in the use of the databases, both technical and chemical.
- Training in the use of the databases.
- A publicity/outreach program for potential new users.
- Access to a number of smaller, specialist databases which appeal mainly to a subset of the user community.
- Integration of multiple data sources via a single interface for structural chemistry.
- Future developments – CDS looks out for relevant new databases and does some limited development on interfaces.

Why these things are needed

The current Service has 4,200 registered users. Of these up to 900 are active in any given month making a total of over 80,000 accesses per year to the databases. This shows a demand from the user community for the current Service. In addition, our 2005 survey of users indicates that 60% of users find the Service very useful or vital to their research, indicating that it is not just of peripheral interest.

The Service supports research published in the peer reviewed scientific literature. The ISI Citation Index found 70 publications published in 2005 which directly cited a CDS descriptive paper. In addition, there are an unknown number of papers which acknowledge the Service, but do not cite our reference. Data from the user survey suggests that the average active user publishes one such paper every three years as well as a further 2-3 papers which have made some use of the Service but do not cite or acknowledge it.

The survey data shows that users were over 5 to 1 in favour of the retention of the smaller, specialist databases which mainly appeal to a subset of the user population. Such datasets are realistically only going to be available through a centrally funded service.

Some of the communities supported by the Service are still experiencing massive growth. The number of users from Chemical Engineering departments has grown by

50% in the last 2 years, coincident with the introduction of the Detherm database. Usage of this database is currently doubling on an annual basis. A central service is vital in order to get a critical mass of users within a community to maximise exploitation of a data resource. There are other communities, particularly within the Life Sciences, where this expansion has not yet taken place.

How CDS fits in with complementary services

There are many electronic sources of chemical information available to the UK academic community today. Perhaps the two most known services are the Beilstein/CrossFire service hosted at MIMAS and the CAS databases, predominantly now accessed through SciFinder Scholar (SFS). In addition, there are numerous websites of varying quality, including institutional data repositories, which are likely to become more important sources of data in the medium term. Finally, users may have access to local copies of data, such as library CD-ROMs.

Beilstein is a large molecule database which contains a lot of data, including single step reaction data. As such it can be viewed as a competitor to the CDS reaction databases. However, it is different to the current main CDS reaction database ChemInform, which contains both single and multi-step reactions selected for their synthetic utility. So if you are looking for a published synthetic route to compound X, the Beilstein database is a good place to search. However, if you want to develop a synthetic route to an unknown compound X, ChemInform is far more useful. The Beilstein/CrossFire service also requires a subscription, which some institutions cannot afford.

The CAS registry database can be considered the premier chemistry database and the majority of our users will also want access to it. However, it does not contain the raw data itself; you would not use this to find the structural co-ordinates of a small molecule. Also it is very expensive to access via SFS.

These two major resources are thus complementary to the service provided by CDS. User survey data shows that around 50% of our users also use the Beilstein/Crossfire and SFS services. This shows that our users also think these services are complementary to CDS.

Web resources are many and vary greatly in quality. Whilst there are some very good data sources available (such as the NIST Chemistry WebBook), they are not a substitute for the large, high quality datasets available through the current Service.

Institutional repositories, designed to hold the research output from a university, will likely contain a variety of chemical data in the medium term. New, distributed, search tools will need to be developed to interface across these data stores and are likely to emerge from the e-Science program. We would envisage incorporating such tools into the global molecule search interface, described below. Again, these repositories will not be a substitute for the comprehensive (in some fields) and historical data CDS provides.

What else is needed

There are a number of developments which would benefit the community in the short to medium term.

The first is a single global interface to all of the data held by the Service. This has strong support from the current user base (>80% would like to see it) and CDS has already done some enabling development in this area. The goal is a simple to use, web based, molecule search interface which returns either data or hyperlinks to data on the requested molecule. The system should be easily extensible to include new data sources, both internal and external (such as the data repositories for the other EPSRC supported chemistry national services).

We know that support for the Macintosh platform is important for a significant minority of the current user population. New interfaces, such as that described above, should support this. Older systems, where possible, should be enhanced to provide support for Macs where they do not already do so.

Adoption of a Shibboleth based access and authentication system, which is likely to be the new standard within the UK academic community for resources, would facilitate access to the Service.

Central service justification

The need for the academic community to have access to this chemistry data is demonstrated above, and the reasons to provide it through a central service are presented below.

Perhaps the most compelling argument is the cost of licensing the data. High quality databases are not cheap, though a single central license is typically substantially cheaper than multiple single licenses as well as being administratively much cheaper for both the academic community and the supplier. In addition, the bulk purchasing power of the central service can often result in a better price. Two current examples serve to illustrate these savings.

Institutions which subscribe to the current Beilstein/CrossFire service now have the option to subscribe to MDL's DiscoveryGate service, which would give them access to most of the synthetic organic chemistry databases available through CDS. The annual cost for this would be in excess of £500k, i.e. more than the total cost of the current CDS, and this is only for the current CrossFire subscribing institutions not the whole academic community.

The Detherm database of thermophysical properties was purchased by CDS in April 2004 for €173k (including 3 years maintenance). The alternative access mechanism for the community would be via the Dechema online service, which costs €80 for each search (returning a modest 30 data points). Already, the searches performed by CDS users would have cost some €220k if performed via the online service, and usage of the database is still growing. In future years the annual maintenance costs of the centrally held data will be some €17k compared with an online search cost of at least €150k.

A central service has sufficient size to become a centre of excellence with expert knowledge available for solving both technical and chemical problems. Similar knowledge will be patchy or more likely non-existent without a central service. This allows for a high quality helpdesk for the community as well as the resources and expertise to run training courses.

Centralisation leads to significant economies of scale in the hardware used to run the databases and maintenance effort required. In addition, the centrally maintained databases are regularly updated when new data becomes available. In the distributed model this is not always the case – for example many users in the community with local copies of the Cambridge Structural Database do not apply the interim data updates which are made available throughout the year between major releases of the database.

A central team can also keep abreast of future developments in the field and be aware of the release of new databases and systems which might be of value to the community. This includes the facilitation of new national deals for large services (the current CDS was instrumental in enabling the current CrossFire/Beilstein deal, being in a position to quickly setup and run a trial service).

One other point in favour of a central service is that it is much easier to integrate different data sets when they are all in the same place. Search tools can be developed to allow seamless access through a single global interface to all of the data held by the central site.

Summary – how the future Service should look

A central service run by a team of experts who:

- Provide integrated access to a wide variety of high quality databases in the area of chemistry.
- Provide simple global access through a single web interface to all the data as well as the more powerful, proprietary interfaces.
- Link and integrate with relevant high quality external sources of data.
- Provide a helpdesk and support in the use of the databases, both for technical and chemical problems.
- Provide training courses in the use of the databases and training materials such as tutorials.
- Run a publicity/outreach program for potential new users and new communities.
- Provide access to smaller, specialist databases which appeal mainly to a subset of the user community.
- Keep abreast of developments in the field such as new databases and systems, and arrange trials and evaluations of these.
- Enhance and develop the in-house interfaces in response to user demand.