THE CDS PERSPECTIVE OF THE EPSRC CROSS SERVICE PANEL RECOMMENDATIONS OF JUNE 2006

21 September 2006

EXECUTIVE SUMMARY

This report provides a considered response to the findings and conclusions of the EPSRC Cross Service Panel at its meeting on June 5-6, 2006, and the associated report from the panel that was eventually provided to CDS on July 11, 2006. Much of the discussions with EPSRC in the aftermath of the report have focused on the specific issue of the need for a National Cheminformatics Service, and has not delved into the serious consequence of the report in its recommendation that the CDS service itself should terminate on completion of the current grant period i.e. on April 1, 2007. This separation of the two major conclusions from the report was an agreed approach between EPSRC, the Chairman of the Service's Management Advisory Panel (MAP) and the PI on the current CDS grant, and was taken in the best interest of de-coupling the issue of strategy regarding the need for a National Cheminfomatics Service, and the more emotive issue of the impact on the existing service.

This document returns to the issue of the impact on the Service itself by presenting a formal response to EPSRC around the whole process that has culminated in the present situation. This report has been prepared by the CDS and discussed in full with the Service MAP, and is fully endorsed by that body. The major conclusions presented in detail in the body of the report may be summarised as follows:

1. The Service believes that the Review process itself and the subsequent conclusions are seriously flawed for a number of compelling reasons:
   - No community input had been sought in spite of the major impact of the ensuing decision.
   - There are numerous factual errors and clear examples of the mis-interpretation of provided data in the report itself.
   - There are repeated instances of the panel being unclear about the data to hand, but no attempt was made to seek clarification on these issues from the Service itself prior to the Panel reaching its conclusions.
   - The lack of synthetic organic chemistry representation on the panel was a serious omission that had a major impact on the conclusions reached.
2. The impact of implementing the conclusions of the Panel i.e. to terminate the existing service and to continue the provision of a limited number of offerings from that service by an alternative, yet to be specified mechanism, will have a major negative impact on 60% of the current users of the service. Annually, more than 1200 individuals from 75 institutes use services that are to be stopped. This number is increasing at 10 per cent per year. Alternative arrangements by these users would end up costing the UK significantly more.
3. Many research projects as well as cheminformatics courses are already in progress or are about to start with the implicit assumption that CDS is available to support them. The transition to alternate arrangements will adversely affect all of them, obviously more so in the cases where no alternate access is available.

1. Introduction

This report provides a considered response to the findings and conclusions of the EPSRC Cross Service Panel at its meeting on June 5-6, 2006, and the associated report from the panel that was eventually provided to CDS on July 11, 2006. Much of the discussions with EPSRC in the aftermath of the report have focused on the specific issue of the need for a National Cheminfomatics Service, and has not delved into the serious consequence of the report in its recommendation that the CDS service itself should terminate on completion of the current grant period i.e. on April 1, 2007. This separation of the two major conclusions from the report was an agreed approach between the EPSRC Chemistry programme, the Chairman of the Service's Management Advisory Panel (MAP) and the PI on the current CDS grant, and was taken in the best interest of de-coupling the issue of strategy regarding the need for a National Cheminfomatics Service, and the more emotive issue of the impact on the existing service.

This document returns to the issue of the impact on the Service itself by presenting a formal response to EPSRC around the whole process that has culminated in the present situation. This report has been prepared by the CDS and discussed in full with the Service MAP, and is fully endorsed by that body. Our analysis is presented below under the following section headings:

2. The Assessment Process regarding any Future National Service
3. The Assessment Process relating to the Current Service
4. Advantages of the CDS versus Commercial Alternatives
5. Specific Points raised by the Cross Service Panel Report
6. Summary and Conclusions

CDS Usage Statistics referred to in this Document are provided as an Appendix

2. The Assessment Process regarding any Future National Service

1. The current review process attempts to mix an assessment of the current Service with recommendations on how to best meet the future needs of the chemistry community. This has placed the Service in an invidious position in preparing its submission.
2. The CDS MAP meeting on 16/9/05 expressed disquiet at the procedure which was being adopted by the EPSRC to decide whether they would continue to support a national Chemical Database Service. The MAP unanimously agreed (reflected in Section 3. of the Minutes) that it was not an appropriate role for the CDS to argue this case.
3. There is the conflict between preparing an annual report covering recent developments in an objective fashion, proposed new developments and arguing the case about the need for a future Service. This problem is further increased given that we have repeatedly been told that any future provision would be subject to a tender exercise, and there would be the real possibility that our ideas could well be used by competitors if, as proposed, reports are made publicly available.
4. Given the major impact this decision will have on the academic chemistry community, it seems surprising that they were not consulted during the review process.

5. The case should be analysed and argued by the EPSRC – no doubt with the assistance of some sort of expert committee. The EPSRC should also explicitly solicit input from the wider UK academic chemistry related communities. There is precedence for this with the Jennings Committee exercise in 1992/3.
6. A key issue in this assessment appears to be the usage of the organic chemistry databases. This is a mature component of the CDS, having been available for over 12 years. These databases are used by the majority of CDS users (over 60%), and the absolute number of users is currently growing at around 10% annually (see the Appendix). Previous EPSRC Cross Service Reviews have made incorrect statements (Section 2.3 of 2005 Minutes) about the static nature of organic chemistry database usage.
7. Misinterpretation of the usage statistics presented in our Reports has continued with the 2006 Cross Service Review. In fact there has been healthy growth in the use of the organic chemistry databases over recent years (see Appendix). In some ways this has surpassed that of the systems which are recommended to be retained.
8. The increase in organic chemistry usage is despite the easier availability to the community of systems such as Beilstein and CAS/SciFinder. This supports our contention that the organic chemistry information made available via the CDS is regarded by experienced synthetic organic chemists as being largely complementary to that available via the other major database systems. It is unfortunate that the 2006 Panel did not include any recognised specialist synthetic organic chemists.
9. Having said this, it understandable why some misreading of the data might have happened. We provided a great deal of information, and it was difficult to predict which particular aspects were likely to be selected. A revised set of figures illustrating what now appear to be the salient features is provided in the Appendix to this document.
10. It has become abundantly clear in retrospect that the continuation of a national service was to be judged almost solely on the success of the CDS in producing some sort of substantial increase in the use of its synthetic organic chemistry component.
11. In addition, the Cross Service Report makes no mention of the value of spectroscopic database systems. There is the current SpecInfo system, which has moderate usage. More importantly there is the current ACD/Labs I-Lab trial which, with the availability of central service facilities, will make important new features available to the community as a whole.
12. The impact of any decision to drop provision of all organic chemistry components of the Service to the UK academic community, if implemented, would be drastic. Users of the ISIS and SpecInfo components represent 61% of the active user base. There have been 1,896 active users of these systems in the last 3 years. 80% if these users are from RAE grade 4, 5 and 5* chemistry departments.
13. Alternative arrangements for the bulk of these users will cost the UK about £ 1/2 million a year by the cheapest options[1,2]. This would still miss out a number of

[1] Access to the organic chemistry databases via MDL's DiscoveryGate service through the most recent JISC brokered deal for the 42 institutes which currently make significant use of them would cost £340,000 annually.

important components[3] and disenfranchise a proportion of the users completely (we estimate this number to be several hundred).

14. It should perhaps be regarded as a bonus that the latest Cross Service Review recommended that support be continued for the crystallography and thermophysical components. It is, however, in no way clear that they had a coherent view as to how this might happen.

3. The Assessment Process relating to the Current Service

1. The Cross Service Panel awarded RED grades for a number of Service activities. We find many of these curious, but will make just a few more general observations.
2. There is a clear tendency to award a RED mark where the committee experienced difficulties in understanding aspects of the CDS Report. We believe this may well indicate possible failings in our submission rather than any real problems with performance or management.
3. There are some clear anomalies in the markings. Demand was awarded an AMBER last year. This year there were appreciably more users and accesses in all areas to an increased data holding, but the mark awarded was RED.
4. We feel strongly that the award of a low rating for performance of Service provision was unjustified given that the statements in the Cross Panel Report simply criticised a lack of clarity in our presentation.
5. Also the committee seems to conflate management with performance. It is a most curious situation when “the service was clearly working and delivering access to databases” and yet was given a RED performance rating. We were also criticised for not providing specific benchmarks without actually having been asked to provide them.
6. The recurring obsession with precise pigeonholes of individual group member's roles is strange. A perfectly good categorisation of these roles has already been provided.
7. The Panel recognised the difficulties in producing a simple metric to measure the research quality of the work enabled by the Service. They acknowledged the importance of the Service facilities in underpinning a wide range of research.
8. The Panel, however, took the response rate of our recent user survey as meaning a lack of support and an indicator of low importance in the delivery of high quality research. This was a misunderstanding of the purpose of the survey. It was mainly an attempt to find information on user requirements with a view to making Service improvements. It was not an attempt to extract an endorsement from the community. The response rate was, in fact, perfectly respectable falling into the 5-10% range we typically see for this type of survey.
9. The above is not intended as a particular criticism of committee members but of the process itself. This all relates to the apparent mixing up of the appraisal of the performance of the current providers with need for future provision.

---

[2] Access to spectroscopic data via ACD/Labs I-Lab system for the 42 institutes which make significant use of them is estimated to cost around £150,000 annually.

[3] There would be no access to the Accelrys databases (Protecting Groups, Solid Phase Synthesis and Biocatalysis) or to the Chiral Separations database.

10. Given these important issues we would have expected that the Panel would look to question the CDS itself directly to seek clarification on contentious issues, or at least to have had the Service in attendance in case such issues arose. We would request EPSRC give more serious consideration to the interaction between review panel, and those it is reviewing; the distancing of the two groups serves little purpose and inhibits decision making when all the salient facts are not to hand.

4. Advantages of the CDS versus Commercial Alternatives

1. The Cross Service Panel Report states “The advantage of the service provision verses commercial provision of the data is not clear”. There are many advantages to provision via CDS which were presented in the submissions to the Panel, and these are reiterated below.
2. Cost of access – for instance, subscribing to the MDL DiscoveryGate service (excluding Beilstein) via the latest JISC brokered deal for the academic community would cost around the same as the whole of the current CDS, and this would only provide some of the data which is proposed to be dropped. It also requires institutes to sign up to a 5 year commitment, which many are reluctant to do.
3. Some suppliers, such as Dechema or STN, provide pay-per-view access to the data. This can be confusing when in Euros or Dollars and can work out very expensive for the inexperienced user, all of which is likely to put them off.
4. Commercial providers only provide access via their own proprietary software and do not combine any other provider’s data. CDS provides CrystalWeb (access to all the crystallography databases) and the reaction data contains Accelrys data as well as MDL’s own data.
5. Since CrystalWeb was introduced, the number of accesses and users has been steadily growing, doubling in the last 3 years. It now accounts for almost a quarter of the crystallography accesses and one third of crystallography users.
6. Commercial providers usually provide accesses through one route only and often only support PCs, sometimes UNIX and not Macs. Use of Macs is wide spread amongst academic faculty members, especially within the synthetic organic chemistry community. Supplier provision tends to reflect what is the norm within large pharmaceutical organisations.
7. CDS provide different access routes (client software or web) for both PC and Macs. Cambridge does not as yet provide web access to their data.
8. MDL only supplies web access to the Available Chemicals Directory (ACD) or their Screening Compounds Database (SCD) via a package deal (DiscoveryGate). The CDS provides alternative modes of access. In addition the Service compiles an SCD implementation from data provided directly by chemical supplier worldwide. Our version is significantly more up to date than that available from MDL.
9. Commercial suppliers do not support secondary software which may be required to access the data. For instance, Quest and ConQuest require X-Windows access from PC’s and Mac’s but Cambridge do not supply information about emulators. CDS provide help and information on a range of X-Windows emulators.
10. Software provided by one provider is also used by CDS to enhance other databases such as the use of LitLink which is available in both reaction and crystallography databases.

11. Training – most providers do not provide training and are not based in the UK. Those that do (e.g. MDL) are very expensive. They provide little in the way of on-line information unlike CDS that provides a great deal including Flash based demonstration movies and tutorial sets.
12. Outreach - commercial suppliers are less likely to run a publicity/outreach programmes targeted at, for instance, graduate students. They rely on exhibitions at conferences which in many cases are sparsely attended by academics.
13. Commercial suppliers would much rather deal with a single large customer than lots of individual users and not host it themselves. For example, ACD Labs have come to CDS to run their trial and host their data.
14. CDS acts as a focus for trialling new systems which may be of benefit to the community. The CDS infrastructure can be used to rapidly mount new systems, inform users and collect feedback. Without CDS, the successful Beilstein trial would have been much more difficult, the current I-Lab trial would not be taking place and the highly regarded Detherm database would not be available to the community.
15. CDS will also be vital in enabling systems using new e-science technologies and facilitating access to distributed resources.

5. Specific Points raised by the Cross Service Panel Report

1. The Cross Service Report addresses issues around the need for a central database service for chemistry. Here we refer to 13 specific points made in Section 4 of this document that we would seek to challenge; in many cases we believe these to be flawed, and would appreciate the opportunity to seek further clarification from EPSRC as to their substance, and in some cases, accuracy. Quotes from the report are given in italics.
2. *The provision of such data has changed radically over the past 10 years and continues to do so with the developments in GRID and e-science*:

   This is true, but these developments have not yet in most cases resulted in mature services readily usable by the practicing research chemist.
3. *The developments in data-provision via GRID/e-science mean that a centralised service no longer appears appropriate*:

   This relates to the point above, but is currently visionary and not a justification for dismissing the role of a centralised service; indeed one might argue that it is fanciful in the extreme. It should be reiterated that these are immature technologies that do not currently support viable alternative services. This said, the CDS recognises the importance of GRID and Web based technologies and continues to work actively in these areas. Central organisations still have an important role in enabling these new technologies and in data validation. As a national facility the CDS is potentially in a unique position to work in concert with data generators such as the other EPSRC chemistry services. Fully distributed data systems are unlikely to happen spontaneously without some sort of central focus as might be provided by a national Chemical Database Service.
4. *New models for accessing data sources will become available using these emerging technologies and there is the prospect of fully distributed data systems*.

   However, they are not available at present and experience shows that such systems will not initially be very user friendly. The existence of a central

informatics service is likely to be critical in ensuring that the potential of these e-initiatives are actually delivered to those not expert in IT and chemical information. There also remains the problem of data quality with distributed data systems which is unlikely to be addressed by individual users.

5. *The service appears to be focussing on the smaller/specialist databases:*

   This is untrue. The provision of specialist data represents only a part of the Service. However, the breadth of coverage of the service has increased with time and it may well be sensible for this trend to increase further in the future.

6. *major key databases provided through other routes*

   This is presumably a reference to the CAS databases. These were available 10 years ago as they are today, but not via CDS. Why has this suddenly become an argument against CDS as a national service? Experienced users fully recognise the complementary nature of the various systems.

7. *Other than cost there appears to be little reason why individual researchers and institutions cannot access all their data requirements through direct contact with the vendors:*

   This is not the case. Aside from cost there is a major economy of scale of effort in doing this centrally, and in some cases significant expertise (e.g. Detherm) is required in setting up systems. In other cases, such as CrystalWeb there is no equivalent commercial system. It is also much more convenient for a typical user to access a full range of available systems from a single access point. Such users are unlikely to be particularly computer literate or have great knowledge of or interest in the information technology scene.

8. *Most significantly the level of usage of synthetic organic community had not shown a substantial level of increase.*

   Organic synthesis data is a mature component of CDS (available for over 12 years). There seems to be an expectation - not explained, justified or quantified - that growth should be bigger than it is, plus an expectation that organic usage should be larger. Where does the data which supports this assertion come from? Organic synthesis is a well established sector of the Service which caters for some 1200 active users a year and is growing at a healthy rate (see Appendix) of some 10% per year.

9. *The only two key elements of data provision, which continue to have significant usage, are the structural and DETHERM databases:*

   This is not the case. The usage of the organic databases dwarfs that of Detherm and is about the same as that of the structural databases.

10. *it would appear that the demand for these other databases is not sufficient:*

    As an example, the Available Chemical Directory is used by around 1000 chemists a year. Just how many is deemed to be sufficient?

11. *there is now co-operation of institutions on a regional level and more specialist databases could be provided regionally*

    This may well be some sort of palliative option open to some. It would, however, be more expensive and involve more effort overall.

12. *The increased provision of usable interfaces provided by the vendors and the opportunity to contact the data provider directly, makes the case for a training programme less clear.*

This is in part true. However, there appears to be a blanket assumption that database manufacturers will provide training. There is also no real consideration of the cost implications for this or whether such training is actually available (not all the database manufacturers are service providers). There is a later assertion that "a page of FAQs and downloads" on a manufacturer's website is sufficient, and yet, somewhat inconsistently, the Service is rated AMBER in Section 2.7 of Cross Service Report for not providing enough hands-on training. It would appear the panel has no coherent view as to how much training is required by the community or the best way to provide it.

13. *The service is potentially unique in offering a central portal and the ability to automate the searching of data across a number of databases; however, with communities such as organic and physical chemists using data provided outside the service then this is of less use.*

    The Service has acquired data which is verified and of high quality. Its portfolio includes important commercial data from a number of providers and the Service is probably in a unique position to make cross database linking. Because alternative data source are, in some cases, available does not mean such an opportunity should be passed over. Indeed current developments (e.g. PubChem, DiscoveryGate, CSLS and the linking of SpecInfo data in SpresiWeb) in the wider informatics world suggest that such linking of data sets is both valuable and important.

14. *The negotiation of a major regional or national licence could be carried out by another body such as JISC/Eduserv*

    Negotiating licences has been one of the many functions performed by the CDS, but we are perfectly happy when aspects of this process can be taken on my others. However, none of these other bodies have any particular knowledge of chemical database systems or a direct means of appraising their potential value with direct links to practising chemists. It should be remembered that the current highly successful Beilstein/CrossFire system would not have been made available to the UK academic community, in the timely manner it did, without the crucial involvement of the CDS. Setting up and arguing the case for Detherm was another example where a national specialist chemistry database centre was able play a unique role. The current ACD/Labs I-Lab trial is yet another case whether the CDS can enable a national deal with real benefits to the community at large. All these instances have involved direct support of the users and the resolution of key technical issues.

6. Summary and Conclusion

1. We believe that the process used to determine whether a National Cheminfomatics Service is required is deeply flawed.
2. As indicated in Section 2, there will be a large segment of the community, especially amongst the synthetic organic chemists, who will be badly affected by the EPSRC proposals for a drastic reduction in database support. We have quantified this impact above.
3. We also believe the Cross Service Panel badly underestimated the effect on other key sections of the UK chemistry community.
4. Many points in the Cross Service Report are either factually incorrect or are based on a false interpretation of our submission.

5. Individual arranging for direct access to commercial systems will inevitably be more expensive overall both in terms of direct costs and also support effort. In particular, this is true for the current portfolio of large database systems.
6. Such economies will also be important for many more specialist systems (where the aggregates cost would be substantial). It is a sensible function of national provision to allow central access to these systems.
7. Reliance on emerging new technologies such as GRID-based systems is over optimistic in the short, and arguably, the medium term.
8. Moreover, the existence of a central cheminformatics service will be critical in ensuring that the potential of these e-initiatives are actually delivered to those not expert in IT and chemical information issues.
9. Chemistry information support, advice and training available to the user community will be very much reduced or become non existent with the loss of central service provision
10. The community will lose a value resource which once lost will be very difficult to replace.

APPENDIX. CDS Usage Statistics referred to in this Document

Unique/individual active users of the databases

ISIS+SpecInfo (organic chemistry and spectroscopy)

Apr03-Mar04	951	
Apr04-Mar05	1,009	(a rise of just over 6% from previous year)
Apr05-Mar06	1,138	(a rise of just under 13% from previous year)

Almost 10% rise in users on average every year over the last 2 years

	Unique Active Users	`New' Users in 2nd Year
Over 2 Years (Apr03-Mar05)	1,403	452
Over 2 Years (Apr04-Mar06)	1,507	498
Over 3 Years (Apr03-Mar06)	1,857	

Average number of New Registered Users per year over last 3 years = 717
Therefore, on average, 66% of new users use ISIS/SpecInfo

Almost **80%** of active users are from institutes with 4, 5 or 5* RAE grade chemistry departments.

Crystallography databases

Apr03-Mar04	1,047	
Apr04-Mar05	1,157	(a rise of just under 11% from previous year)
Apr05-Mar06	1,177	(a rise of just under 2% from previous year)

Just over 6% rise in users on average every year over last 2 years

	Unique Active Users	'New' Users in 2nd Year
Over 2 Years (Apr03-Mar05)	1,524	477
Over 2 Years (Apr04-Mar06)	1,576	416
Over 3 Years (Apr03-Mar06)	1,896	

On average, 62% of new users use the Crystallography databases.

Some overlap of users:-

Last Year (Apr05-Mar06)	Cryst+ISIS+SpecInfo = 1,907 Unique Active Users
Over 2 Years (Apr04-Mar06)	Cryst+ISIS+SpecInfo = 2,468 Unique Active Users
Over 3 Years (Apr03-Mar06)	Cryst+ISIS+SpecInfo = 2,976 Unique Active Users

ISIS/SpecInfo users represent **61%** of all active users of CDS.