Cheminformatics 1 (2004) 37–67
IOS Press

Chapter 2

The United Kingdom Chemical Database Service: CDS

Bob McMeeking and Dave Fletcher
CCLRC, Daresbury Laboratory, UK

1. Introduction

The Chemical Database Service (CDS) [1] is a national service that provides access to a wide range of high-quality chemical data systems to the UK academic community. It is centrally funded by the Chemistry Programme of the Engineering and Physical Sciences Research Council (EPSRC) [2] and access to its services and facilities are provided free of charge to all eligible users. The CDS has its origins in the 1970s, but its nature has changed considerably since that time. Originally CDS stood for Chemical Databank System, reflecting the fact that the retrieval system evolved from a core of related software codes as the scope of the data searchable by the CDS search software progressively increased. In the earlier days most effort was expended on software development enhancing the search functionality of the core codes, but major modifications were also needed to accommodate new types of data. The change of name to 'Service' reflects the change in its nature as it has evolved over the years. The breadth of the community interests catered for by the Service has expanded as have the numbers who access it (currently over 3600 registered users).

2. History

2.1. Origins

The core CDS program codes have their origins in the early 1970s. Most of the pioneering work associated with the CDS in the UK was done by Mike Elder and Pella Machin, who were originally based at the computational crystallography support group at the Science Research Council (SRC) [2] Atlas Laboratory in Oxfordshire. Owen Mills, at the University of Manchester's Department of Chemistry, also played an influential role in establishing the early system. In fact there were a number of joint postdoctoral appointees (some based at Manchester) who worked on various aspects of system development up until the mid 1980s.

In 1974, at the request of the SRC Data Compilation Committee, Elder and Machin performed an assessment of an information retrieval program to provide interactive access to the Cambridge Crystallographic Data Centre's (CCDC) [3] databank of organic and organometallic crystal structures, the Cambridge Structural Database (CSD) [4]. At that stage this database consisted of less than 20,000 entries. Each entry had a bibliographic element, a connectivity element, and a data element, but they were only searchable by the various batch-oriented computer programs then available from CCDC. The CSD has grown enormously since then with currently approximately 300,000 structures and a range of sophisticated search and analysis software (see Chapter 6 in this issue). At that point, however, truly interactive search systems of any kind, especially available over a wide area network, were very much in their infancy.

The original codes used by the UK team to make the CSD searchable by a new, interactive software system were provided by Richard Feldmann from the National Institutes of Health (NIH) at Maryland, USA. They were based on the highly innovative ideas [5] developed in collaboration with Steve Heller and Bill Milne with whom the programs [6] were co-written. They had developed the NIH nested tree substructure search system, which operated on a connection table database and could be run on a general purpose time-shared computer. This was to become the Structure and Nomenclature Search System (SANSS) [7], which incorporated these ideas. The initial motivation of the American group was to develop a search system for what became the NIST/EPA/NIH Mass Spectral Database [8].

The NIH system was soon to be augmented with the addition of a whole range of chemical databases. This became the NIH/EPA Chemical Information System (CIS) [9] operated by Fein-Marquart Associates. For a period the CIS system offered subscription-based access to all these systems internationally, usually via telephone links. However, the system that evolved in the UK was completely separate from this initiative. The CSD file was processed and indexed by a suite of 'inversion' programs adapted from what was provided by the NIH. The original information retrieval program was the Crystal Structure Search and Retrieval (CSSR) [10] program. This was also from the NIH, but subsequent evolution happened independently of any NIH involvement. The CSSR program was modified and developed into the general 'classic' CDS program capable of handling not only crystal structures but also data from other chemical databases. There were a number of major innovations, including a better procedure for complex ring searching. Here processing involved novel substructure perception and ring decomposition procedures. There was also an improved capability for rendering the chemical diagrams. Later a system to allow general substructure searching for the non-expert or occasional user was also introduced. The systems remained, however, essentially command line driven and the various new features drew heavily on the original NIH ideas [10].

Various databases were eventually made available for retrieval by derivatives of the classic CDS program. One of the first to be investigated was CNMR [11], a 13C NMR spectral database, developed in part by the Mills group at Manchester with

close interactions with the CDS. The chemical connectivity information for CNMR was abstracted using the Wiswesser Line Notation (WLN) [12]. It was necessary to convert these WLN strings to connection tables, in a reliable and automated way, so that the data could be handled by the inversion process. A package, DARING, was written by Elder to do this. The Fine Chemicals Database (FCD) [13] produced by Fraser Williams (Scientific Systems) Ltd [14] also had structures inputted as WLN strings, and a year or so later was also converted to be searchable by the CDS software. About the same time the REWARD package was written by Steve Hull, a Manchester postdoc. REWARD generated chemical structure diagrams from information available in connectivity tables. It was based on Tektronix graphics rather than the character-based system included in the original NIH code.

Both DARING and REWARD were for a time licensed to and made commercially available by Fraser Williams. There was a good relationship with Fraser Williams with the prospect of the sale of software providing useful funds. Also there was thought to be the prospect of collaborations with Pergamon-InfoLine [15] that might lead to exchange of software for databanks or funds for future software improvements. In addition, licenses to use the CSSR retrieval system were sold to various industrial users and overseas academics.

In the mid 1970s networked systems for the general academic chemist only offered 'promise' and gaining access within a university environment was by no means routine [16]. Use of the software that had been developed so far was restricted to a select group of mainly crystallographers who had access to the limited network facilities available at the time. Support for the CSSR system was perceived as a logical addition to the function of the SRC Crystallography Support Group. Thus the group was renamed Applications Group and moved to another major SRC facility at Daresbury [17] in Cheshire during 1977. Funding was eventually obtained to provide more formal arrangements for access by the UK academic community in general, and the prototype CDS was well established by the late 1970s. Other databases of interest to research workers at Daresbury in the areas of nuclear and atomic physics were also examined by the group, as were alternative database management systems. However the work involving chemistry-based systems and building on developments of the original NIH-derived software were to prove the most enduring.

2.2. Early service

In 1978 a preliminary remote service had been established with the retrieval program mounted on a DEC System-10 at the Edinburgh Regional Computing Centre (ERCC). This was part of the SRC's Interactive Computing Initiative. Users of the Service gained access via Post Office dial-up facilities (pre-privatization). The Service at that stage provided free access to 25 registered academic users and 5 UK chemical companies, who paid a levy based on access time. Advances in networking made this possible but, of course, this was limited by today's standards and even by those of the early to mid 1980s.

The closure of the ERCC DEC-10 facility was announced in the early 1980s. This left the CDS in a difficult position, and the ideal solution was to acquire a dedicated server for database use, situated at Daresbury. A case for funding was drafted in 1983 and was eventually successful, but this outcome was by no means certain. It involved a fight against powerful skeptical elements within the UK chemical community who doubted the value of computer-based chemical information systems, given the costs involved (and in some cases irrespective of cost).

The new service was started in 1984 and its launch coincided with that of JANET (Joint Academic NETwork), which provided good quality connectivity between the UK Institutes of Higher Education. The CDS was hosted by its own VAX 11/750 server (actually a machine clone produced by SYSTIME) running the VMS operating system, which was considered to be a very powerful microcomputer system at the time. It had a main processor clock speed of around 6 MHz! Savings were made through the ability to share peripherals such as the line printer and tape drive with a data acquisition VAX system already on site, but the total cost was still approaching £100,000. The CDS was the 2nd VAX system on site and was named DLVB. DLVB came with 2 MB of memory and 134 MB of fixed disk space and two 67 MB exchangeable disk drives. All retrieval codes had to be converted to run on the new system. This was by no means a trivial task, involving modifications to take account, amongst other things, of the change from the 36-bit architecture on the DEC-10 to 32 bits for the VAX. Database inversion processing was required relatively infrequently, but was computationally demanding. The codes involved in this also required conversion but now ran on the Daresbury mainframe system (a NAS AS7000 system) under the IBM MVS operating system.

In the original case for funding involving DLVB it was anticipated that the full power of the system would not be required to support chemical database activities in the first instance, and it would be possible to charge third parties for use of the spare machine capacity. The idea was that this would defray the cost of machine maintenance, etc. In fact provision of the Brookhaven Protein Databank (PDB) [18] originally fell into that category. DLVB was also made available as a front end server to an Evans and Sutherland PS3000 graphics system, running the protein structure modeling and electron density fitting programs, FRODO, GOLLUM and HYDRA [19]. These packages were used by visiting protein crystallographers while performing structure determinations using the Synchrotron Radiation Source (SRS) [17], a major research facility also housed at Daresbury. Another initiative being explored, also of interest to the molecular biology community, was provision of protein and DNA/RNA sequence databases. Other members of the SRS community paid for the Chem-X package [20], which was also mounted on DLVB. There was, in fact, a good deal of synergy between many of these adjunct activities and the main service. For instance, although it was mainly intended for local use, general CDS users could also access Chem-X, and many chemists also had an interest in the protein structural data.

2.3. Established service

With the new funding the Service also felt able to recruit extra permanent staff. Elder was to leave Applications Group and take on a higher management role within Daresbury. Machin took over as head and also had strong involvement with protein crystallography code development, including the CCP4 project [21], but was still able to commit about 50% of her time to the CDS project. In addition two Ph. D. level recruits were appointed (Bob McMeeking and Ian Kirkman) specifically for CDS work together with a graduate (Ian Clifton) who allocated 50% of his time to CDS related work. Also at this point there was the happy situation where the main CDS computer resources were under the strategic control of the CDS group, but day-to-day system management effort was available from expert, centrally-funded support staff who handled the nuts and bolts of hardware and system maintenance.

Initiatives to enhance the coverage of crystallographic data began when three new data sources became available. The databases involved were the Inorganic Crystal Structure Database (ICSD) [22,23], CrystMet (Metals Crystallographic Data File, MDF) [24,25] and the Crystal Data Identification File (CDIF) [26]. This was a highly-productive period and not only for crystallographic systems. Developments also continued with the CNMR and FCD databases. Building on experience with CNMR there were collaborative interactions with Preston Scientific Ltd [27], who were compiling a new 19F NMR database system. This new FNMR database was also made available to the CDS community via a retrieval system adapted from the standard CDS codes.

Another focus was the provision of tools to add extra useful functionality, mainly in the field of structural analysis. For instance, an auxiliary program, CRAD (Calculation of RAdial Distribution), evaluated distances about any selected atom within a chosen structure. It had great use in association with EXAFS (Extended X-Ray Absorption Fine Structure) data analysis. EXAFS techniques were becoming established as a very powerful experimental tool for structure analysis on the SRS. The VIEW program, as its name implies, allowed display of the structures from the 3D coordinate data. It used Tektronix 4010/4014 graphics and even in the mid 1980s could be used by a relatively large section of the community who did not have access to the high-end graphics hardware that was required by some of the more sophisticated commercial graphics program systems that were becoming available at that time. Another program developed around that time was CDA (Crystal Data Analyser) [28]. CDA was to prove a bit of a dead end as it relied on the use of the single-user PERQ minicomputer. The PERQ [29] was a highly-innovative graphics workstation, which was being promoted heavily in UK academic circles, and, like many such systems, did not achieve long-term commercial success.

The organic chemistry community was beginning to make increasing use of the Service, in particular the FCD and the NMR databases. There was also interest in gaining access to other databases related to organic synthesis. At that time (mid 1980s) there were two major systems that might be contenders for inclusion in the

Service. These were the REaction ACCess System (REACCS), a commercial system produced by MDL [30] and Organic Reactions Accessed by Computer (ORAC) [31]. ORAC was being developed by the Wolfson Unit for Computer-Aided Design of Organic Synthesis at the University of Leeds. A spin-off company, ORAC Ltd, had been founded but there continued to be strong academic connections and the Service was able to get free access to the ORAC software in the summer of 1986 through the cooperation of Dr. Peter Johnson, who was the director of the Wolfson Unit. Access to LHASA (Logistic and Heuristic Applied to Synthetic Analysis; see also Chapter 4) was also obtained through the Leeds connection. LHASA was set up by the Corey group at Harvard University, but LHASA UK [32] was to be based at Leeds. ORAC came with a database of some 10,000 entries. ORAC, REACCS, and LHASA all had become commercial products and came with their own retrieval and knowledge base management systems. They used relatively advanced display features for the time, and also required more expensive graphics hardware. Initially this created difficulty for a number of users but cheaper solutions, usually involving terminal emulation on PCs, were to become available.

Tragically, Machin and Elder died in a mountaineering accident in March 1987, but details of the Service at that time were presented in a review article [10] written by Machin shortly before her death. The article includes some details of the inversion procedures and retrieval facilities available at that time. It also stated the philosophy underpinning the system, which included the following points.

- This is an information retrieval system, not a database management system (DBMS), and it is therefore concerned with providing rapid interactive retrieval rather than the more general facilities for registration, storage, and manipulation of structures. Database producers require a DBMS to organize their data, but once the database has been compiled and distributed the emphasis is entirely on information retrieval.
- The software is interactive and relies upon indexed and inverted file techniques. This method will be described in some detail in a later section of this paper.
- A strength of the system (compared with many commercial systems) is that it allows retrieval by chemical structure as well as retrieval by text and numeric values. A structure may be entered as a diagram, matched against structures in the database, and the resulting hits may be displayed as structure diagrams.
- The system is aimed at chemical users who are able to ask questions in chemical terms, consider the implications of their results, and then refine their queries accordingly. Less interactive systems that handle searches more automatically may be simpler to operate but are less flexible and probably give less informative results in the long run.
- The system may be used with minimal equipment, such as a VDU, over wide-area networks. The majority of users are UK university scientists and, although some have access to powerful graphics terminals, many do not. All access is from a distance over networks and telephone lines.

2.4. Transitional period

At that point the Service had made a further grant application and the SERC [2] committed funds to support the Service for a further four years beyond April 1987. Clearly the funding committee was persuaded of the value of a central database facility, such as provided by the CDS, but was only able to provide support equivalent to a staffing level of 1.5 persons as opposed to the 2.5 requested. The expectation was that external sources would cover the deficit. One avenue, which was apparently being explored by Machin at that stage, was to attempt to place the provision of the molecular biology databases on a more formal basis. Funding would be sought from the SERC biological sciences community. The DLVB system would be enhanced with extra disk space. Support for the new facilities could be absorbed by current personnel but additional funding, possibly equivalent to half a person, acquired. Despite new initiatives, such as those outlined above, the main emphasis of the Service was to provide tools for rapid and convenient searching of chemical databases using the sort of networking and equipment which was becoming readily available to the average chemist within his or her department.

Applications Group was split up and responsibility for the database part passed to Howard Sherman who was at that time head of the central Computer Operations and Support Department and thus had responsibility for the computer helpdesk (User Interface Group; UIG) section. Sherman was able to take on many of the management tasks involved with the Service and brought his enthusiasm and optimism. An additional benefit was to speed up the process (already started) whereby the UIG took over functions such as user registration, front line fielding of user queries, maintenance and distribution of forms, manuals, and newsletters, tabulation of usage statistics, etc. These were jobs which would become increasingly important with the expansion of the user community and relieved pressure on the core CDS personnel.

Sherman's background was in physics but he had no previous direct involvement with chemical database systems. The remaining members of the CDS team were at that point McMeeking and Kirkman. The loss of Elder and Machin was a particular blow because they had unique knowledge of many of the locally written codes. The original NIH programs incorporated some excellent features which they had adapted and built upon. The main retrieval code, at some 27,000 lines, was relatively small by modern standards, but it had grown in a relatively unstructured manner, becoming complex and difficult to maintain by anyone without an intimate knowledge of its history and development. The database inversion codes for the newly-added database systems were designed from scratch to run on a VMS platform, but those for the older systems, most importantly CSSR, still ran on the old AS7000 mainframe. These codes involved many interlocking components, which were quite complex and highly machine-dependent.

In theory the inversion process was largely automated but inevitably required a good deal of manual intervention and troubleshooting. The AS7000 was soon to be replaced by a new Convex central facility, which ran under the UNIX operating

system. In fact the loss of the AS7000 was to present the Service with a major problem.

2.5. New directions

A number of initiatives that had been started by Machin before her death were now developed further. One of these was in the area of electrochemistry and involved a collaborative venture with Professor Victor M.M. Lobo from the University of Coimbra, Portugal, who had compiled a collection of electrolyte solution data [33]. Kirkman produced software for data entry and validation, together with retrieval software (again based on the standard CDS codes). A small NATO travel grant was obtained, which contributed towards some of the expenses relating to the collaboration, and an initial release of the Electrolyte Solutions Database (ELYS) was made available to CDS users in 1987.

In 1986 Bednorz and Müller had discovered the first examples of a new class of high-temperature superconducting materials [34], for which they subsequently won the Nobel Prize. This created huge excitement, especially among members of the solid state physics community, many of whom were keen to take part in the race to come up with a satisfactory theoretical explanation (a quest which continues until the present day). This novel phenomenon was seen to have huge commercial potential, conjuring visions of magnetically levitated high-speed trains and dirt-cheap electrical power. There were requests from theoreticians at Daresbury and a small prototype database was set up using the ideas developed for the ELYS project. The aim was to provide timely access for the UK community to the very latest literature in published and preprint form. At an early stage it incorporated the articles collected by the Iowa State University High-Tc Superconductivity Information Center [35]. The Center was founded in April 1987 as a short-term solution for a crisis situation: the need to communicate frequent (almost daily) breakthroughs in the field. Over the years it became a successful model for how research results can be very effectively communicated, and it continued through to December 2003. In addition, there were more formal arrangements involving the UK Department of Trade and Industry (DTI) and the database gained access to data abstracted by the GEC Hirst Research Centre. This data tended to reflect the interests of GEC and shifted the focus towards the more technological aspects of the field.

Around the same time, molecular biology databases were becoming a particularly hot topic and the Human Genome Mapping Project [36] was soon to be established. The Biological Sciences Directorate of the SERC was keen to be involved in the action and initiated an open tender to universities and research council sites to provide services. Daresbury submitted a proposal based largely on Machin's original scheme. In retrospect it is difficult to see why such a sparsely-provisioned bid should have been taken very seriously, given competition from other groups who were experts in the field and who, in certain instances, already provided well-established services to their local communities within, for instance, the Medical Research Council (MRC) and

the Agricultural and Food Research Council (AFRC) [2]. The reasons were probably mainly political. The SERC directorate was keen to establish a high profile within this emerging field and favored a submission from one of its own establishments. Also, within the UK molecular biology community there was keen rivalry between the various factions and Daresbury may have been seen as a temporary stop gap that offered no particular threat to others. The original bid was considered to be insufficient, but Daresbury was invited to resubmit a modified bid. There was a strong steer from the Biological Sciences Directorate and a good deal of effort was devoted to this exercise, resulting in a more realistic submission. It included a provision to support a dedicated service manager and pay for its own dedicated MicroVax, more modern technologically and with considerably more power than DLVB. The new service was called SEQNET [37] and Alan Bleasby was recruited as service manager, starting mid 1989, and his unique qualities ensured its continuation at Daresbury for the next decade There were further group members recruited to SEQNET over the years and a new Collaborative Computer Project (CCP11 [38], also based at Daresbury) was initiated in this area.

2.6. Crisis in the core provision

SEQNET and HITC were high-profile systems that were both timely and that fulfilled extremely valuable functions. The group demonstrated that it was capable of moving very quickly, responding to the needs of new communities. Indeed, Daresbury did receive a certain amount of kudos from these initiatives. The HITC project attracted extra funding equivalent to half a person over 2 years, specifically dedicated to associated database maintenance. These were, however, departures from the CDS core function, and there was a feeling of unease in those immediately involved in supporting the system as a whole. They felt that their efforts had to some degree been hijacked and user support had been diverted to meet the needs of a different community. Indeed, Kirkman decided to leave the group in 1988 but was replaced some time afterwards by Lesley Curwen who remained with the Service until 1992.

As indicated in Section 2.4, the group was unable to provide sufficient effort to cope with the withdrawal of the AS7000 computer system in 1988. One immediate effect of this was that both the CSSR and FCD components were not updated for a period approaching 18 months. In 1990 there was a crisis of funding for the whole of the SERC. Projections of funding from software sales and royalties made by the Service also turned out to be very optimistic. Opportunities had been reduced because commercial vendors were now able to fill the gaps once available to academically-driven systems such as those provided by the CDS. In addition, the Service was not able to properly develop its provision of reaction databases since these databases had now become fully commercialized. Indeed, key facilities had to be withdrawn because of insufficient funds. These included the ORAC, LHASA, and CNMR database systems.

Not surprisingly there was disquiet amongst key users. A more general discontent reflected the recognition that the CDS services had, over the intervening years, become a vital part of the research infrastructure for a growing segment of the UK academic chemistry community. The EPSRC recognized the importance of restoring the primary focus of the Service to serving the needs of the chemistry community, and appointed a Management Advisory Panel (MAP) to help in this regard. However, the Service was unable to meet mounting user expectations for enhanced systems and new data without a substantial increase in grant funding.

2.7. Evolution of the current service

In 1992 the SERC set up a committee, initially under the chairmanship of Professor Jennings from Warwick University, to investigate the information needs of the UK academic chemistry community. The committee recommended database provision in four areas: crystallography; synthetic organic chemistry; spectroscopy; and physical chemistry. A competitive tender was issued to run the Service, which was won by Daresbury. The new service was fully funded with 3 full time staff (Bob McMeeking, Dave Fletcher, and Don Parkin) plus support from the central services at Daresbury (UIG, systems and networks), and a significant expansion in the databases available took place right from the start.

- Crystallography, the core of the original Service, remained largely unchanged apart from the introduction of the Quest graphical interface from the CCDC.
- Synthetic organic chemistry received a huge boost, with the introduction of the REACCS software from MDL with a collection of reaction databases, the main component of which was the Current Synthetic Methodology (CSM) [39] system compiled by FIZ Chemie, Berlin. In addition, the Available Chemicals Directory (ACD) was made available, also using the REACCS system.
- Spectroscopy was enhanced with the SpecInfo package from Chemical Concepts [40], with a large 13C NMR dataset, and smaller hetero-atom NMR and IR datasets. The data in SpecInfo immediately superseded that in CNMR, and this was removed from the Service. FNMR, however, was retained for some time.
- Physical chemistry was the only area not to receive an immediate boost and for some time was only represented by the ELYS database.

The focus of the Service moved away from local code development and with the additional manpower it was possible to provide enhanced support and training to the user community. Staff members gave frequent training courses and were also able to devote time to developing on-line support, mostly centered on the Service's Web site, which was started in 1994.

The Service was now managed through a Service Level Agreement (SLA) with monitored monthly performance indicators. This became a more important issue when the funding body (EPSRC) was separated as an organization from Daresbury

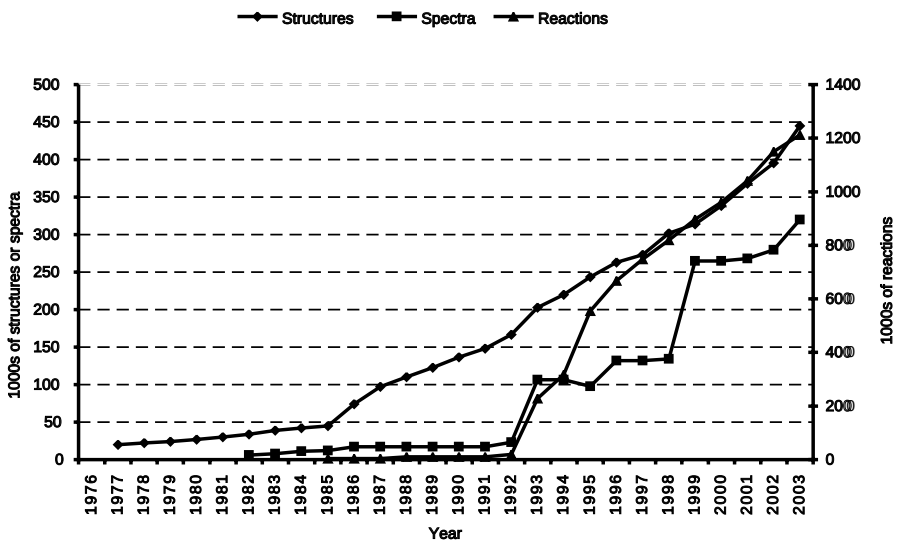

Fig. 1. Growth of data available from the CDS.

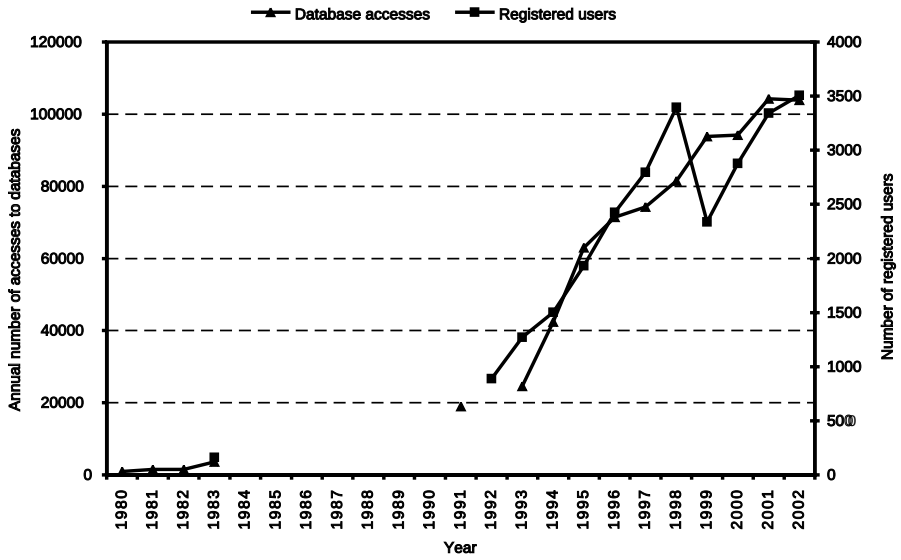

Fig. 2. Growth of CDS usage.

Laboratory (CCLRC) [2]. The grant to fund the Service was reviewed and renewed by the EPSRC on average every two years. The first such renewal, in 1995, led to the inclusion of a significant element of money for new, but as yet unspecified, databases.

This allowed the Service to acquire several new databases, which were of value to the user community, without having to expend further time and effort applying for additional funding. For example a number of important specialist databases from Synopsis (now part of Accelrys [41]), the chiral chromatography separation database, ChirBase [42], and additional spectral data (particularly assigned 1H NMR spectra) were obtained in the last decade. Importantly, this 'contingency' funding also gave the Service the flexibility to find the resources to run the 1996 trial of the Beilstein/Crossfire system [43] for UK academics (see Section 2.8).

It is clear that post-1993 there was a rapid expansion in both the quantity of data available and the size and activity of the user base. This is readily apparent from the graphs shown in Figs 1 and 2. 1994 also saw the start of the transition of the Service onto UNIX-based computers. In 1996 a paper was written describing and reviewing the Service [44]. It gave users a reference to cite in their published work and this gave a mechanism to help track the benefits of the Service. To date the paper has received well over 500 citations, but in reality this only represents a small fraction of the work which has benefited from use of the Service. The main features of the current Service are recognizable in the description given in the 1996 paper. An updated outline is given in Section 4 but, with the exception of the major addition of the Detherm system, many of the changes are more a matter of detail.

With internal reorganizations at Daresbury in the late 1990s it became less cost-effective to use the services of the UIG group so in 2001 these functions (front line user support, registrations etc.) were returned to the core CDS group and the savings used to hire a fourth member of the group (David Osa-Edoh). In reality, with the advent of more intuitive interfaces, on-line training material, and automated registration procedures, these tasks had become far less onerous and time-consuming than they were a decade earlier, despite the substantial growth in the user base. Thus a certain amount of time was again available for development work (Bedlam, CrystalWeb – see Section 3.3) and to keep up with the continuing trend towards introducing new interfaces for existing databases (ISIS, Chemscape, ConQuest, Oracle, ICSD-WWW – see Table 1), some of which required considerable effort to setup.

2.8. The Beilstein experience

By 1995 the Beilstein Informations Systeme GmbH had converted the extensive data in the authoritative Beilstein Handbook of Organic Chemistry into electronic form [43] and this was searchable via their CrossFire client/server software. Also around the same time, IBM in Europe had developed linking software between the MDL ISIS system and the CrossFire search engine. It was this particular configuration that sparked the initial interest of members of the CDS team.

Negotiations were started with Beilstein and it became clear that a CDS user trial, but with the native CrossFire search system, might be a possibility. The print version of the Beilstein Handbook had a very high reputation and was traditionally considered the definitive reference source by organic chemists. It also held a unique collection

of physical properties data in addition to reactions and related material. It was not cheap, and this, together with the fact that it required training and a knowledge of German to be used effectively, meant that its use within the UK academic community had declined considerably. In fact the John Rylands University Library (JRUL) in Manchester was the only UK university library with a current subscription at that time. The electronic version is in English and is easy to use by the research chemist. It was a comprehensive source (with some 7 million chemical substances and 10 million reactions at that time), which was being maintained and actively updated. However, it was agreed that an evaluation trial lasting up to 6 months would be necessary to re-educate the UK community as to the value of the data in the Beilstein Handbook.

The CDS was clearly the most appropriate place to host a UK national academic Beilstein service but there were doubts as to whether the EPSRC would be able to make available finance to support the system using its standard 'free to end user' model. The EPSRC and our MAP were, however, very supportive and encouraged the Service to divert available resources to set up and run a UK trial. In particular a new computer system, which was about to be commissioned, was pressed into service for CrossFire use and necessary additional disks and peripherals were acquired.

It turned out that Beilstein was also being considered by a JISC (Joint Information Systems Committee [45]) advisory committee and there was the real possibility that funding could be sorted out via that route. It was decided to join forces and the formal negotiations were taken on by the CHEST (Combined Higher Education Software Team) acting on behalf of JISC. This was a case where all parties involved (including Beilstein) had a strong commitment to arrive at a successful outcome. There was powerful lobbying within the JISC committee structure (particularly by Diana Leitch of JRUL), but it all hinged on the success of the CDS trial. The effectiveness of the trial was considerably enhanced by the help provided by Malcolm Bruce, a chemist also at Manchester. Bruce continued to be an active user of the Beilstein Handbook for his own research. He was a powerful and articulate advocate who was in a position to lobby key, powerful elements within the UK chemistry hierarchy.

A deal was eventually signed (this also included the Gmelin Handbook Component [46]) in late 1996. Arrangements involved rates affordable to UK institutions via a site subscription mechanism. This was unique worldwide, and at that point UK academics had a worldwide lead in a key aspect of database provision. This in many ways mirrored the situation when the CDS MDL reaction database deal was brokered back in 1993. The contract to run the Beilstein service in the longer term was put out to competitive tender, but it transpired that JISC wished to restrict those who were allowed to compete to their own sponsored national data centers (EDINA at Edinburgh, MIMAS at Manchester, and BIDS at Bath). The CDS figured in a subcontracting role in the submission put together by EDINA, but the MIMAS [47] was successful. The CDS was able to hand over the complete Beilstein service to MIMAS by summer 1997 with all major teething problems sorted out. These were mainly associated with users having to install the client/server software. By that

Table 1
Databases available from the CDS

Crystallography databases

Database	Supplier	Current contents	Access program	Dates available	Description
Cambridge Structural Database (CSD) [4]	Cambridge Crystallographic Data Centre (CCDC) [3]	307,000 crystal structures of organic and organometallic molecules	Crystal Structure Search and Retrieval (CSSR)	1976 – present	Command line interface with Tektronix graphical display
			Quest	1994 – present	X-Windows interface
			ConQuest	2000 – present	X-Windows interface
			CrystalWeb	2002 – present	Web interface-bibliographic and cell data searching only
Inorganic Crystal Structure Datafile (ICSD) [22,23]	FIZ Karlsruhe [75]	74,000 crystal structures of inorganic compounds	ICSD	1986 – present	Command line interface
			ICSD-WWW	1999 – present	Web interface
			CrystalWeb	2002 – present	Web interface bibliographic and cell data searching only
CrystMet [24,25]	Toth Information Systems [76]	75,000 crystal structures of metals and alloys	Metals Data File (MDF)	1987 – present	Command line interface
			CrystalWeb	2002 – present	Web interface bibliographic and cell data searching only
Crystal Data Identification File (CDIF) [26]	NIST [77]	Crystal cell data for 238,000 compounds	CDIF	1985 – present	Command line interface
			CrystalWeb	2002 – present	Web interface bibliographic and cell data searching only
Brookhaven Protein Databank (PDB) [18]	Brookhaven National Laboratory [67]	8,000 protein structures (1999)	PDB	1986–1999	Command line interface with X-Windows graphical display

Synthetic Organic Chemistry: molecule databases

Database	Supplier	Current contents	Access program	Dates available	Description
Fine Chemicals Directory (FCD) [13]	Fraser Williams (Scientific Systems) [14]	Supplier details on 67,000 compounds (1992)	FCD	1982–1992	Command line interface with Tektronix graphical display

Table 1, continued

Synthetic Organic Chemistry: molecule databases – continued					
Database	Supplier	Current contents	Access program	Dates available	Description
Available Chemicals Directory (ACD)	MDL Information Systems [30]	Supplier details on 254,000 compounds	REACCS	1993–1996	Tektronix graphical interface
			ISIS	1995 – present	Client/server system
			Chemscape	2001 – present	Web interface
National Cancer Institute (NCI) [64]	MDL Information Systems	214,000 molecules	ISIS	2001 – present	Client/server system
			Chemscape	2001 – present	Web interface
ChirBase	ENSSPICAM [42]	85,000 chiral separations of 28,000 molecules	ISIS	1996 – present	Client/server system
			Chemscape	2001 – present	Web interface
Molecular Diversity Preservation International (MDPI)	Molecular Diversity Preservation International [65]	9,000 compounds	ISIS	1999 – present	Client/server system
			Chemscape	2001 – present	Web interface
Beilstein/ Gmelin [43,46]	MDL Information Systems	Data on some 8 million organic compounds (1997)	CrossFire	1996–1997	Client/server system
Screening compounds	Various [66]	Supplier details on 2,424,000 compounds	ISIS	2003 – present	Client/server system
			Chemscape	2003 – present	Web interface

Synthetic Organic Chemistry: reaction databases					
Database	Supplier	Current contents	Access program	Dates available	Description
ORAC	ORAC [31]	20,000 reactions (1992)	ORAC	1985–1992	Tektronix graphical interface
Reference Library of Synthetic Methodology	MDL Information Systems	210,000 reactions	REACCS	1993–1996	Tektronix graphical interface
			ISIS	1995 – present	Client/server system
			ReactionWeb	2000 – present	Web interface
Current Synthetic Methodology [39]	MDL Information Systems	22,000 reactions (1996)	REACCS	1993–1996	Tektronix graphical interface
ChemInform RX [39]	MDL Information Systems	843,000 reactions	ISIS	1995 – present	Client/server system
			ReactionWeb	2000 – present	Web interface

Table 1, continued

Database	Supplier	Current contents	Access program	Dates available	Description
Synthetic Organic Chemistry: reaction databases – continued					
Protecting Groups	Accelrys [41]	39,000 reactions	REACCS	1993–1996	Tektronix graphical interface
			ISIS	1996 – present	Client/server system
			ReactionWeb	2001 – present	Web interface
Solid Phase Synthesis	Accelrys	23,000 reactions	ISIS	1997 – present	Client/server system
			ReactionWeb	2001 – present	Web interface
BioCatalysis	Accelrys	38,000 reactions	ISIS	1999 – present	Client/server system
			ReactionWeb	2001 – present	Web interface
Derwent Journal of Synthetic Methods (DJSM) [78]	MDL Information Systems	79,000 reactions	REACCS	1993–1996	Tektronix graphical interface
			ISIS	1995 – present	Client/server system
			ReactionWeb	2000 – present	Web interface
Organic Syntheses [79]	MDL Information Systems	6,000 reactions	REACCS	1993–1996	Tektronix graphical interface
			ISIS	1995 – present	Client/server system
			ReactionWeb	2000 – present	Web interface
Spectroscopy databases					
Database	Supplier	Current contents	Access program	Dates available	Description
Carbon Nuclear Magnetic Resonance (CNMR) [11]	Netherlands Information Combine	14,000 13C NMR spectra (1993)	CNMR	1982–1993	Command line interface with Tektronix graphical display
Fluorine Nuclear Magnetic Resonance (FNMR)	Preston Scientific [27]	6,000 19F NMR spectra (2000)	FNMR	1985–2000	Command line interface with Tektronix graphical display
SpecInfo	Chemical Concepts [40]	102,000 13C NMR spectra 2,000 19F NMR spectra 1,000 15N NMR spectra	SpecInfo v 2	1993–1996	Command line interface with Tektronix graphical display
		1,000 17O NMR spectra 2,000 31P NMR spectra	SpecInfo v 3	1996–2002	X-Windows interface

Table 1, continued

Spectroscopy databases – continued					
Database	Supplier	Current contents	Access program	Dates available	Description
		61,000 1 H NMR spectra 21,000 IR spectra 130,000 Mass spectra	SpecInfo v 4	2002–2004	X-Windows interface
			SpecSurf	1999 – present	Web interface

Physical Chemistry databases					
Database	Supplier	Current contents	Access program	Dates available	Description
THERM	Dr Pedley, University of Sussex	Heats of formation of 2700 compounds	THERM	1983	Command line interface
Electrolyte Solutions [33]	Prof. Lobo, Coimbra University	10,000 electrolyte solution datasets	ELYS	1987 – present	Command line interface with Tektronix graphical display
High Temperature Superconductors	Various	28,000 bibliographic references	HITC	1986–1993	Command line interface
Detherm	DECHEMA [51]	4.2 million thermophysical datasets on 21,000 compounds and 101,000 mixture	DETHERM	2001–2002 and 2004 – present	Client/server system

stage it had a substantial and active user base. Since that time MIMAS has done an excellent job in operating the service, which has continued to grow and flourish [48].

The Beilstein/CrossFire system provided a huge new store of data for UK users However, even in areas of overlap (in particular for reaction data), tests performed involving the CDS group [50] indicate that there is a high degree of complementarity between Beilstein and the CDS portfolio.

3. Services

The CDS started as a system whose scope was limited to the CSD. Subsequently other crystallographic databases were added together with 13C NMR data and chemical supplier catalogue information. All these systems evolved from the original core in-house system. Further databases, usually with their own retrieval system developed by third party suppliers, were added at a later stage. At present, the largest components of the Service are the organic chemistry reaction databases, which are currently searched using MDL's ISIS retrieval software [30]. There is also a large

section of spectroscopic data using the Chemical Concepts SpecInfo system [40]. Physical chemistry systems have historically been less well supported by the Service, but this has been recently addressed by the acquisition of the thermophysical properties data included in the Detherm system produced by DECHEMA e.V. [51]. Further details of the development of the CDS database portfolio are given in the sections below and summarized in Table 1.

3.1. Central services

The Chemical Database Service is a central resource for the whole of the UK academic community. In the past the computational disk space requirements of most databases made a central resource the only practicable mechanism for providing widespread access. Today powerful computer systems with gigabytes of storage are readily available and the costs of these are not the major consideration. The expense of the hardware is now eclipsed by that of the database licenses and support staff. Even so, central services may still be a cost-effective solution. National academic site licensing arrangements are frequently much cheaper than the aggregate cost of many individual licenses. Global support costs will be less than the sum over many small installations. In addition the centralized pool of experience gives a valuable resource for troubleshooting, problem solving, and to provide training. Additional benefits include enhancements through the use of auxiliary programs and specially-written linking scripts, which a central service typically can provide.

The EPSRC, as a funding agency, has to balance its commitment to central funding for chemical information provision against the call for direct funding from workers engaged in high-quality research. Within the UK it is currently recognized that the two are totally complementary. In the earlier years, as indicated in Section 2.2, this was not always the case. This is the case at the moment, but the cost-effectiveness of the central provision offered by the CDS will always remain under review. Any national service must demonstrate its value and relevance within the current changing computation and information scene. Currently the bulk of the CDS effort is taken up by providing a reliable and consistent service with a higher emphasis on publicizing the available features and providing training to the users.

3.2. Services access

Most of the major databases provided are now commercial products, which come with their own access software. These packages, of course, did not exist in the early days, and there has been continuing evolution with enhanced interfaces available to the end user. There is a clear trend in the development of most of the database access programs. They have changed from the original text-based command line interfaces, through the addition of graphical displays (usually Tektronix-based), to X-Windows-based applications and client/server systems, to the current crop of Web-based interfaces. With the growth of the World Wide Web it is now generally

expected that everything can be accessed through a Web browser. This was the primary motivation for the recent development of the CrystalWeb system by the Service. CrystalWeb provides simplified Web access to all the crystallographic data that is currently provided by the Service. Virtually all the data the CDS makes available is now accessible via some form of Web interface, though in some cases the functionality is still limited compared to other available mechanisms.

Despite these excellent new systems, some of the CDS legacy codes are still supported and are actively used by a section of the user community. These codes have been fully maintained over the years but it is generally agreed they are reaching the end of their useful life, and retirement beckons. Although the legacy codes are largely superseded by new systems from third party suppliers, there are cases where they do provide unique functionality. For instance, downloading of coordinate data via the old software underpins some of the important functionality of CrystalWeb. Replacing this with a more modern and easily-maintained system is currently a high priority.

3.3. Software development

This is still part of the work of the CDS team but the team no longer has the central role it did in the very early days. Pre-1990, significant code development was needed for each database. Commercial software was generally not used (or not available in most cases). The nature of database support has also changed substantially with the advent of what are notionally 'off the shelf' systems. Computer hardware and operating systems have advanced and changed beyond recognition during the lifetime of the Service. Also, what was the preserve of a central computer service priesthood can now be handled by members of the CDS group. This means, however, that given the number of disparate packages and hardware platforms involved, there is now an even greater support overhead, and this is handled totally within the CDS group itself. Non-commercial software developed and maintained by CDS staff includes:

– **Classic CDS programs.** These include the codes which powered the early Service. As described in Section 2.1, the original CDS package (CSSR) was designed to search the CSD. Over the first two decades of the CDS this code was developed to provide access to most of the data provided by the Service (ICSD, MDF, ELYS, FCD, CNMR, and FNMR were all derived from the CSSR system). Inverted file structure techniques were necessary when the programs were written to produce acceptable search response times. Today bibliographic searches, such as for an author, are essentially instantaneous using these systems. The disadvantage of using inverted data structures is that a great deal of information needs to be preprocessed to produce them. This took a significant period (days of processing time) on the early VMS and AS7000 systems. All these systems retain a basic command line interface, but over time Tektronix graphics and X-Windows display facilities for chemical structures were added. Various simple

search options are available and Boolean combinations of resultant hit sets can be used to answer complex queries. In addition, further software was written in the mid 1980s to access the data available in CDIF. The CDIF interface provides special facilities to perform unit cell searching that is not dependent on the choice of crystallographic setting selected. It is completely different from the other interfaces but is also command line driven. Command line interfaces are now considered to be essentially obsolete for general user programs but the basic CDS retrieval system is used in the background to underpin some of the functionality of the CrystalWeb interface.

– **CrystalWeb.** This is the most recent interface built by the CDS. It provides simplified Web access to all of the crystallographic data. It is not intended as a replacement for ConQuest (see Chapter 6 in this issue) and the other interfaces since it has a limited functionality by comparison. CrystalWeb allows bibliographic, name, formula, cell data, and reduced cell data searches. There are no geometric or structure searching capabilities. The system is built round an Oracle database [52], which contains the bibliographic, name, and formula information together with cell data, and is accessed through a forms-based Web interface and CGI scripts. Currently it uses the older CSSR, MDF, and ICSD programs for structure data retrieval. Structure display and manipulation requires a suitable browser plug-in such as Chime [53] from MDL [30], or Jmol [54], but structure output is available in most major crystallographic and modeling formats. Retrieved references are delivered in printable or Endnote [55] forma and are automatically set up to use the CDS LitLink server [55] (see below) to give single-click access to the available electronic source(s) of the primary literature.
– **Auxiliary programs.** There has always been a need for extra functionality not covered by the main packages, such as support for additional file formats, and the CDS has attempted to fill this software niche. There is also a need for glue-type utilities to link different systems together (though much of this work is done behind the scenes, invisible to the users). Structural chemistry and modeling have traditionally been the main areas of interest for such auxiliary programs. Format conversion has been handled for a long time by the public domain Babel program [57] but additional routines, which correctly handle the finer details of crystallographic symmetry, were also written locally, and these were eventually combined into the Bedlam program. In the beginning crystal structure display was a specialist task and no CDS tools were available, but there were utilities to calculate geometric properties such as atom-atom distances and angles. In particular the CRAD package, mentioned in Section 2.3, has been refined, and is still very widely used by the UK EXAFS community. In the early days fairly primitive but useable molecular graphics were provided by the VIEW program, also discussed in Section 2.3. Later the CCDC provided the Pluto system (see Chapter 6 in this issue), which has continued to be developed and still provides many useful features, particularly for displaying H-bonded systems. In the 90s

public domain utilities using X-Windows graphics, such as RasMol [58] and XMol [59] became available and were combined with locally-written routines allowing users to build up crystal structures and display them interactively on-line. More recently, excellent display facilities have been integrated directly into some of the main retrieval packages discussed in Section 4 (e.g. ConQuest, CrystalWeb, and ICSD-WWW).

– **LitLink** [56] is MDL technology for linking references from secondary sources (such as databases) to electronic forms of the primary literature (on-line journals). LitLink was initially made available in 2000 but only for the MDL-supplied packages. The technology is now integrated with all of the Web interfaces (ReactionWeb, CrystalWeb, ICSD-WWW) and it is also available with ConQuest. In general it allows a single-click transfer from an entry in a particular database to the available source(s) of the corresponding electronic journal reference. Some customization of the LitLink system has been carried out by the Service, mainly the addition of crystallography journals and electronic sources.

4. Packages/databases

Most of the chemistry databases made available today are commercial products, coming with their own access software (see Table 1). The CDS supports four areas of chemistry: crystallography; synthetic organic chemistry; spectroscopy; and physical chemistry. The organic category can be subdivided into molecule and reaction databases. The Service also strayed into the molecular biology area but this function was taken on at Daresbury by the SEQNET service in the late 1980s. In addition, the CDS took a leading role in establishing the current arrangements for provision of Beilstein/CrossFire to the UK academic community (see Section 2.8). Details of the current CDS offerings are always available on the CDS Web site [1].

Figure 1 shows the growth in the amount of data available from CDS with time for three categories of data: crystal structures; organic reactions; and spectra of compounds. In general, growth has been linear with time, the major discontinuities corresponding to the addition of new databases to the CDS portfolio. This view is perhaps somewhat at odds with the popular perception of exponential growth in electronic information resources over the last decade or two.

4.1. Crystallography

As noted earlier, from its inception the CDS provided access to the CSD collection of small molecule organic and organometallic structures, the ICSD collection of inorganic structures without organic carbon, and the CrystMet collection of metals, alloys, and intermetallics. ICSD and CrystMet complemented the range of data available in the CSD, with limited overlap between them. Thus the CDS provides comprehensive coverage of most areas of structure determination.

The CSD, ICSD, and CrystMet databases all include full 3-dimensional coordinate data for most of their entries. In addition, the PDB contains coordinate data. A primary mirror site for this within the UK was SEQNET. A facility for search and display of PDB was introduced into the CSD Quest program in 1995 and remained available until 1999. The CDS also had access to the more frequent PDB update feed provided by SEQNET. When the PDB search via Quest facility was discontinued it was decided to discontinue storing PDB, given that it was readily available via high-quality interfaces available on a number of Web sites worldwide [18].

The Cambridge Crystallographic Data Center (CCDC) team has developed much high-quality software (see Chapter 6 in this issue) and a great deal of this is made available to CDS users. Components include ConQuest, Quest, Vista, Mercury, Pluto, and PreQuest. The bulk of the users have switched from Quest to using the more modern ConQuest interface. We plan to continue to provide Quest access while this remains an option and there continues to be user demand. The display package Pluto has in many ways been superseded by Mercury but still has many features of value to the research crystallographer and will continue to be made available. Another CCDC system available on the system is IsoStar. The IsoStar package gives access to a knowledge base of non-bonded interactions derived from both CSD small molecule data and the PDB. Thus the Service retains elements of protein and related data in addition to the small molecule peptide data within the CSD itself.

As indicated in Section 3.3, access to the three main crystallographic databases is available via the original 'classic' command line interface. A Web interface for the ICSD database, ICSD-WWW [61], was added to the Service in 1999. ICSD-WWW was developed by Alan Hewat at the Institut Laue-Langevin, Grenoble. It continues to be developed by Hewat and has become a very-widely-used component of the Service. CrystalWeb is a Daresbury initiative and provides Web access to all the main crystallographic components under a common interface (see Section 3.3). This system also continues to be under active development

Further crystallographic coverage is provided by CDIF. In contrast to the other databases, CDIF contains only cell dimension data together with the chemical formula and details of the bibliographic reference. It includes entries from the CSD, ICSD, CrystMet, and the Powder Diffraction File [62] together with data from other sources where only cell dimensions are available. The main purpose of CDIF is sample identification using cell dimension information. CDIF provides a way of checking whether a given structure is already present in the literature, and it can also be used to check for structural relationships. For a given cell there may be an arbitrary choice of axes. For this reason all cell data in the file have been reduced to standard form [63]. Briefly this process involves determining that primitive cell which has the shortest possible cell lengths. In a number of cases further restrictive rules have to be applied for a unique choice. A similar process must be carried out during the search process. Our last update for the CDIF data was received back in 1998 and we understand that it will remain in its current frozen state. However, a reduced cell search has been added to the CDS software for the separate CSD, ICSD, and CrystMet files. CrystalWeb

also supports reduced cell searching; in this case duplicate CDIF entries also present in the other data files have been removed.

4.2. Spectroscopy

Spectroscopy is represented by the SpecInfo package produced by Chemical Concepts. This allows search, retrieval, and spectral prediction using an associated collection of high-quality, validated spectra. Currently there are over 300,000 spectra available, with an emphasis on 13C NMR, 1H NMR, and mass spectra. Over time, use has been made of various versions of the SpecInfo software, and these have employed various levels of Tektronix and, at a later, stage X-Windows graphics. Currently access is solely via the SpecSurf Web interface. The range of spectroscopic data has been expanded over the years. The aim has been to acquire correctly-assigned datasets. In recent years there has been a substantial increase in the holding of 1H NMR data; high-quality collections of this type of data were less readily available when the original CDS system was first set up in 1993. Traditionally the predictive capabilities of the software for 1H NMR spectra has been less well developed compared to that for hetero-atom NMR systems. This has improved markedly in recent years. Another focus has been on acquiring specialist datasets such as 13C NMR spectra of natural product compounds. Additional collections of various other hetero-atom spectra will be added in the near future.

4.3. Organic chemistry

The current CDS holding is approximately 1.2 million reactions, and these are all searchable via the Integrated Scientific Information System (ISIS). ISIS is a client/server system, which was acquired by the Service in 1995 and has now taken over from the old REACCS system. Around the time when ISIS was introduced there was also a substantial increase in the size of the core reaction collection. This was possible because we were able to convert our subscription for the CSM database to one for ChemInform RX (CSM is a subset of ChemInform RX [39]). All the reaction data is also accessible via a common Web interface. The interface used is based on ReactionWeb, which is based on ISIS and Chemscape. These technologies are also provided by MDL and have been configured to meet local CDS requirements. The functionality of the Web interface now approaches that of the client/server version of ISIS and has been adopted by a significant proportion of the user community. In addition to the major collections of general reaction data provided by MDL, the Service has acquired various additional third party databases, which have a specialist focus. These include databases compiled by Accelrys [41] covering protecting group chemistry, solid phase synthesis, and biocatalysis (see Table 1).

The Available Chemicals Directory (ACD) has become one of the most widely-used components of the Service. Access is available via both the ISIS client/server and Web interfaces. In addition to the ACD, further molecular databases have been

added, including the collections totaling some 214,000 molecules that are made available free of charge from the USA. by the National Cancer Institute [64]. The Service has for some time provided a facility to search the collection of novel and unusual compounds (currently approximately 9,000 entries) coordinated by Molecular Diversity International [65]. More recently we have embarked on a project to coordinate access to a whole range of screening compounds available from the catalogues of a wide range of suppliers worldwide [66].

The other main database accessible via the ISIS client/server and Web interfaces is ChirBase. ChirBase is a powerful Information System for Chiral Separations of enantiomers by Chromatography [42]. It currently covers some 55,000 chiral separations of 28,000 compounds. ChirBase, in common with all the ISIS databases that refer to the primary literature, has been enabled to use the LitLink system (see Section 3.3).

4.4. Physical chemistry

As indicated above, the Service has not traditionally catered well for the needs of the physical chemistry community. One of the primary motivations for promoting the Beilstein and Gmelin databases within the UK communities (see Section 2.8) was the fact they are a rich source of physical properties data. More recently the Service held extended evaluation trials for the Detherm database system provided by DECHEMA e.V. [51]. This exercise involved outreach not only to the physical chemists but also to other communities such as chemical engineers, who so far have made relatively little use of the service. Funding to provide Detherm for the Service on a longer-term basis has now been secured. Detherm is a major database of thermophysical properties. It contains data for some 122,000 compounds and mixtures covering more than 200 properties. Our holding covers the full scope of Detherm data collections and currently includes approximately 4.2 million datasets. The data resides in an Oracle data management system and is accessed using dedicated client software.

The CDS ELYS collection contained some 4,000 entries when it was first released to users in 1987, with a total of 10,000 by 1990, at which point it has remained frozen. Detherm includes a very wide range of properties [51]. Even within the electrolyte solutions area it contains approximately five times as many data entries as the ELYS package. We plan to withdraw ELYS very shortly when its user community fully adopts the new system.

4.5. Molecular biology

In the mid 1980s the Service started to provide a download service for the PDB [18]. The PDB at that stage contained only about 350 protein crystal datasets, with quarterly updates, and was obtained from the Brookhaven National Laboratory in the USA [67]. Later it seemed appropriate to provide access to further data in the molecular biology area. Interactive programs obtained from the National Biomedical Research

Foundation's Protein Identification Resource (PIR) [68] were mounted on the CDS VAX and made available to SERC-funded groups via JANET. The programs provided were relatively limited but included PSQ (Protein Sequence Query program), which provided interactive access to the Dayhoff databank of protein amino acid sequences [69], and NAQ (Nucleic Acid Sequence Query program) which gave similar access to three nucleic acid sequence databanks: NBRF [70], GENBANK [71], and EMBL [72]. These databases were quite small at that time. There were also a number of associated batch programs for search and comparison of sequences. When Daresbury was awarded funding specifically to provide the SEQNET national service in 1988 (see Section 2.5), this aspect grew greatly in scope and importance but developed separately from the CDS. It included such features as nightly updates from the EMBL feed, which were considered vital in such an important and fast moving field. SEQNET also took over the main PDB distribution role to UK academics, but collaborative links remained between the CDS and SEQNET until 1999, when SEQNET merged with the MRC Human Genome Mapping Resource Centre [37] and moved to Hinxton near Cambridge.

5. Hardware, users and accesses, teaching

5.1. Hardware

Traditionally the CDS has tried to keep any hardware changes to a minimum, and where they have been desirable and necessary we have, where possible, shielded the user community from any associated disruption. Table 2 shows the main hardware platforms used for the Service. Changes have occurred roughly every four years in recent times; by comparison earlier systems were typically in use for double that time. These changes are now in many cases driven by the fact that critical software packages are only supported on specific computer operating systems. Over the last decade, there has been much consolidation and change in the range of platforms supported by the major software vendors. Previously hardware changes were made to improve performance and lower support costs. Today performance is a lesser issue and hardware support costs (and indeed hardware costs) are negligible compared with the cost of databases. The Service has recently moved to a mixed SUN/Solaris and Intel/Linux configuration, again due to changes in software availability (in these cases the dropping of Irix support by MDL and Oracle).

In the mid 1990s a diverse collection of machines was needed to run the Service. However, much of this complexity was hidden from the user using a mix of system programming and file sharing to create a 'virtual computer' [44]. The user experience was to log on to the same single computer and run any of the database access programs from there, even though in reality many of the programs would be running on different machines. With the advent of the Web it is now much simpler to collect these diverse resources together into what is, from the user's perspective, a single, logical point of access.

Table 2
Main hardware used by the CDS

Main service machine(s)	Operating system(s)	Dates in service
DEC-10 (shared)		1976–1984
VAX 11/750 (clone)	VMS	1984–1991
MicroVAX 3600	VMS 5	1991–1995
2 IBM RS/6000s SGI Indy Dec alphastation	AIX 3 and 4 Irix 5.3 OpenVMS 6	1994–1999
2 SGI Origin 200s	Irix 6.5	1998–2004
Sun SunFire 280 2 Dell PowerEdge 2600s	Solaris 8 Linux (RedHat ES 3)	2004 – present

5.2. Users and accesses

CDS use is restricted to the academic community in the UK. Users come from every chemistry department within the UK and from the majority of related departments such as biochemistry or chemical engineering. Figure 2 shows the growth of usage and the user base with time. User growth has been approximately linear over the last decade. The dip in 1999 is due to a change of machine making necessary a re-registration of all users. The constant turnover of users (graduates finishing courses, short term researchers, etc.) necessitates an annual clean up of IDs (currently we delete some 400 per year). All these activities put greater demands on the time of the CDS support personnel.

The number of accesses also shows an approximate linear growth, mostly due to the increase in the size of the user base together with a gradual increase in the number of accesses per individual user.

5.3. Teaching services

Over time a culture with a much greater emphasis on user support and education has developed. On-line support material has become much more effective with the availability of new technologies and is now a key feature of the Service. The CDS Web site has become a major primary source of information about the CDS and other chemistry database and related services. It also includes a wide range of tutorial and related material produced by the current CDS personnel [73]. An important feature here is the recent development of Macromedia Flash-based [74] movie material demonstrating practical use of most of the major components available on the Service.

The funds made available for the expansion of the Service in 1993 included a significant allocation for the training of users. Provision of support and training is one of the more compelling arguments in favor of a centralized database service like CDS as compared with a multitude of individual database installations. Initially, training consisted of CDS staff visiting sites and giving courses on the use of the various database packages. Over time a significant body of on-line training material has been amassed and is now made available on the CDS Web site. In addition, local knowledge within the departments has grown and it was believed this had reduced

the need for external courses given by the CDS personnel. A recent major focus of effort has been the production of effective on-line material and in the presentation of awareness-style lectures for potential new users and those unfamiliar with the full range of the CDS offerings within the UK community. However, there is a clear message through feedback from important segments of the academic community that too great a reliance on on-line material is not the answer and an increased schedule of off-site training visits is being implemented.

6. The future of CDS

The Chemical Database Service has existed for nearly 30 years, developing from a crystallography service to encompassing the whole range of chemistry databases. New data, such as the Detherm database, will increase the attraction of the Service to other disciplines, such as chemical engineering, materials science, and the life sciences. Managing this diversification of the user base is one of the challenges facing the Service today. Another challenge is the consistent and common interface expectation of the users, fuelled by the increasing integration of the products of a given supplier.

The justification for central services, such as CDS, has changed over the years. Hardware costs are now largely irrelevant and the savings/benefits are in economies of scale in licensing and maintenance, and in the stable pool of expert knowledge that builds up in the central service. The various factors influencing the cost-to-benefit ratio will change in the future (for example, interfaces becoming easier to use and therefore needing less or no training) but it seems likely that the financial justification for CDS will remain for some time to come, providing access to now expensive but increasingly vital information resources.

Ideas developed from current and future work on the semantic grid will become increasingly influential [80]. Exactly how and when all this will happen is not yet clear, but it seems inevitable that these new technologies will take on a similar importance in chemistry information exchange to that of the mainstream World Wide Web at present. In this and other areas the advice we receive from our Management Advisory Panel [81] is of great value to us and provides a useful way of keeping in touch with new areas.

References

[1] See http://cds.dl.ac.uk.

[2] UK Government science funding is channeled through research councils. Initially, the Science Research Council (SRC), which was renamed the Science and Engineering Research Council (SERC) in 1981, the Medical Research Council (MRC), and the Agricultural and Food Research Council (AFRC).
In 1994 major changes occurred: SERC was disbanded and two new councils, the Engineering and Physical Sciences Research Council (EPSRC), and the Particle Physics and Astronomy Research

Council (PPARC), were formed to take on most of SERC's areas of responsibility. Responsibility for biological sciences passed from SERC to a third new Research Council, the Biotechnology and Biological Sciences Research Council (BBSRC), which also subsumed the responsibilities of the former AFRC. The MRC remained largely unchanged.
In 1995 the Daresbury Laboratory, together with the Rutherford Appleton Laboratory in Oxfordshire, was removed from EPSRC and formed into the Council for the Central Laboratory of the Research Councils (CCLRC).

[3] Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK; http://www.ccdc.cam.ac.uk.

[4] F.H. Allen, The Cambridge Structural Database: a quarter of a million crystal structures and rising, *Acta Crystallogr* **B58** (2002), 380–388.

[5] R.J. Feldmann, in: *Computer representation and manipulation of chemical information*, W.T. Wipke, S.R. Heller, R.J. Feldmann and E. Hyde, eds, John Wiley, NY, USA, 1974, pp. 55–81.

[6] R.J. Feldmann, G.W.A. Milne, S.R. Heller, A. Fein, J.A. Miller and B. Koch, An Interactive Substructure Search System, *J. Chem. Inf. Comput. Sci.* **17** (1977), 157–163.

[7] G.W.A. Milne, S.R. Heller, A. Fein, E.F. Frees, R.G. Marquart, J.A. McGill, J.A. Miller and D.S. Spiers, The NIH-EPA structure and nomenclature search system, *J. Chem. Inf. Comput. Sci.* **18** (1978), 181–185.

[8] S.R. Heller, The History of the NIST-EPA-NIH Mass Spectral Database, *Today's Chemist at Work* **8**(2) (1999), 45–50.

[9] G.W.A. Milne and S.R. Heller, The NIH/EPA Chemical Information System, *J. Chem. Inf. Comput. Sci.* **20** (1980), 204–211.

[10] P.A. Machin, The Chemical Databank System (CDS), in: *Crystallographic Databases*, F.H. Allen, G. Bergerhoff and R. Sievers, eds, IUCr Bonn/Cambridge/Chester, 1987, pp. 184–197.

[11] CNMR was a 13C NMR spectral database owned and distributed by an international collaboration fronted by the Netherlands Information Combine. It contained both spectral and structural data, growing to 14,000 entries.

[12] W.J. Wiswessr, 107 years of Line-Formula Notations (1861–1968), *J. Chem. Doc.* **8** (1968), 146–150.

[13] FCD contained supplier details on some 60,000 chemicals. Acquired by MDL Information Systems [30], bulk chemicals were added to produce the Available Chemicals Directory (ACD).

[14] Fraser Williams (Scientific Systems) Ltd., no longer exists as a separate entity; http://www.fraser-williams.com.

[15] Pergamon-InfoLine, an on-line information service in the 1980s, part of the Maxwell group of companies.

[16] S.J. Smith and B.T. Sutcliffe, The Development of Computational Chemistry in the United Kingdom, in: *Reviews in Computational Chemistry*, (Vol. 10), K.B. Lipkowitz and D.B. Boyd eds, VCH Publishers, Inc., New York, 1997.

[17] The Daresbury Laboratory, in Cheshire UK, is now part of the CCLRC [2]. It hosts a number of major research facilities including the Synchrotron Radiation Source (SRS) and the HPCx high performance supercomputer. It is home to a larger number of smaller projects, including the CDS. See http://www.dl.ac.uk.

[18] H.M. Berman, T. Battistuz, T.N. Bhat, W.F. Bluhm, P.E. Bourne, K. Burkhardt, Z. Feng, G.L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, V. Ravichanran, B. Schneider, N. Thanki, D. Padilla, H. Weissig, J.D. Westbrook and C. Zardecki, *The Protein Databank, Acta Crystallogr* **D58** (2002), 899–907; see http://www.pdb.org.

[19] T.A. Jones, *J. Appl. Crystallogr.* **11** (1978), 268.

[20] Chem-X, Chemical Design Ltd., now part of Accelrys [41].

[21] Collaborative Computing Project Number 4 (CCP4) in Protein Crystallography, set up in 1979 to support collaboration between researchers working on such software in the UK; http://www.ccp4.ac.uk.

[22] G. Bergerhoff and I.D. Brown, The Inorganic Crystal Structure Database, *J. Chem. Inf. Comput. Sci.* **23** (1983), 66–69.

[23] G. Bergerhoff and I.D. Brown, Inorganic Crystal Structure Database, in: *Cystallographic Databases*, F.H. Allen, G. Bergerhoff and R. Sievers, eds, IUCr, Chester, 1987, pp. 77–95.
[24] J.R. Rodgers and G.H. Wood, NRCC Metals Crystallographic Data File (CRYSTMET), in: *Crystallographic Databases*, F.H. Allen, G. Bergerhoff and R. Sievers, eds, IUCr, Chester, 1987, pp. 96–106.
[25] P.S. White, J.R. Rodgers and Y. Le page, CRYSTMET: a database of the structures and powder patterns of metals and intermetallics, *Acta Crystallogr* **B58** (2002), 343–348.
[26] A.D. Mighell, J.K. Stalich and V.L. Himes, NBS Crystal Data – Database Description and Applications, in: *Crystallographic Databases*, F.H. Allen, G. Bergerhoff and R. Sievers, eds, IUCr, Chester, 1987, pp. 134–143.
[27] Preston Scientific Ltd, 184 Whalley Road, Langho, Blackburn, Lancashire BB6 8AA, UK.
[28] M. Elder, P.A. Machin and S.E. Hull, CDA: an interactive program for the comparative analysis of crystal structure coordinate data, *J. Mol. Graphics* **2** (1984), 70–78.
[29] The PERQ was a single-user graphics workstation sold in the UK by ICL and heavily promoted in academic circles in the 1980s. See http://www.retrobeep.com/computers/icl/iclPERQ1.htm.
[30] MDL Information Systems Inc., 14600 Catalina Street, San Leandro, CA 94577, USA; http://www.mdl.com. Producers of REACCS, ISIS, Chemscape, LitLink, CrossFire and Chime.
[31] ORAC was eventually merged with MDL Information Systems [30] and the data is now available through the MDL software ISIS and Chemscape.
[32] LHASA Limited, Department of Chemistry, University of Leeds, Leeds, LS2 9JT, UK (was LHASA UK); http://www.chem.leeds.ac.uk/luk.
[33] V.M.M. Lobo, Handbook of Electrolyte Solutions Volumes I and II, Elsevier Publishing Company, Amsterdam, 1989.
[34] J.G. Bednorz and K.A. Müller, Possible high Tc superconductivity in the barium-lanthanum-copper-oxygen system, *Z. Phys.* **B64** (1986), 189.
[35] High-Tc Superconductivity Information Center, Department of Astronomy and Physics, Iowa State University, IA 50011 USA; http:// www.iitap.iastate.edu/htcu.
[36] Human Genome Mapping Project; see http://www.ensembl.org/genome/central/ and http://www.sanger.ac.uk/HGP.
[37] SEQNET (SEQuence NETwork) was started in 1989 at Daresbury. In 1999 it was merged with the MRC Human Genome Mapping Resource Centre (now the Rosalind Franklin Centre for Genomics Research) and moved to Hinxton near Cambridge, UK; see http://www.hgmp.mrc.ac.uk/About/SEQNET.
[38] Collaborative Computing Project Number 11 (CCP11), in Biosequence and Structure Analysis, was established to foster bioinformatics in the UK research community; http://www.ccp11.ac.uk.
[39] ChemInform RX and CSM (a subset) reaction databases, produced by Fachinformationszentrum Chemie Berlin, Franklinstraβe 11, D-10587 Berlin, Germany; http://www.fiz-chemie.de.
[40] Chemical Concepts, c/o Wiley-VCH, Boschstrasse 12, D-69469 Weinheim, Germany; http://www.chemicalconcepts.com.
[41] Accelrys Inc., 9685 Scranton Road, San Diego, CA 92121-3752, USA; http://www.accelrys.com.
[42] ENSSPICAM Chirbase project, Avenue Escadrille Normandie-Niemen, 13397 MARSEILLE Cedex 20, France; http://chirbase.u-3mrs.fr.
[43] The Beilstein Database is the world's largest and most comprehensive factual database in the area of organic chemistry. Produced by, Beilstein-Institut zur Förderung der Chemischen Wissenschaften, Trakehner Str. 7–9, D-60487 Frankfurt am Main, Germany; http://www.beilstein-institut.de. The Beilstein/CrossFire system is now available through MDL Information Systems [30].
[44] D.A. Fletcher, R.F. McMeeking and D. Parkin, The United Kingdom Chemical Database Service, *J. Chem. Inf. Comput. Sci.* **36** (1996), 746–749.
[45] The Joint Information Systems Committee (JISC) supports further and higher education in the UK by providing strategic guidance, advice, and opportunities to use information and communications technology. In particular, it provides access to electronic resources through national data centres and the UK's academic network (JANET); http://www.jisc.ac.uk.

[46] The Gmelin Database is the analogue of the Beilstein Database [43] for inorganic and organometallic chemistry. The data is owned by the German Chemical Society (Gesellschaft Deutscher Chemiker e.V.); http://www.gdch.de. The Gmelin/CrossFire system to access the data is available from MDL Information Systems [30].
[47] Manchester Information and Associated Services (MIMAS) is a JISC-supported [49] national data center run by Manchester Computing, at the University of Manchester, to provide the UK higher education, further education, and research community with networked access to key data and information resources. Known as MIDAS until 1999; see http://www.mimas.ac.uk.
[48] P. Meehan and H. Schofield, CrossFire: a structural revolution for chemists, *Online Information Review* **25** (2001), 241–249.
[49] The Joint Information Systems Committee (JISC) supports further and higher education in the UK by providing strategic guidance, advice, and opportunities to use information and communications technology. In particular, it provides access to electronic resources through national data centres and the UK's academic network (JANET); http://www.jisc.ac.uk.
[50] F.A. Parkar and D. Parkin, Comparison of Beilstein CrossFirePlusReactions and the Selective Reaction Databases under ISIS, *J. Chem. Inf. Comput. Sci.* **39** (1999), 281–288.
[51] DECHEMA, Gesellschaft für Chemische Technik und Biotechnologie e.V., Theodor-Heuss-Allee 25, D-60486 Frankfurt am Main, Germany; http://www.dechema.de. See also http://cds.dl.ac.uk /cds/datasets/physchem/detherm/detherm.html.
[52] Oracle Corporation, 500 Oracle Parkway, Redwood Shores, CA 94065; http://www.oracle.com.
[53] Chime plugin for Web browsers – displays molecular structures, produced by MDL Information Systems [30]; see http://www.mdl.com/products/framework/chime/index.jsp.
[54] JMol – open source molecule viewer; http://jmol.sourceforge.net.
[55] EndNote – a popular personal bibliographic database system, produced by ISI ResearchSoft, 2141 Palomar Airport Road, Suite 350, Carlsbad, CA 92009, USA; http://www.endnote.com.
[56] LitLink server – software linking secondary to primary data sources, produced by MDL Information Systems [30]; see http://www.litlink.com/products/knowledge/litlink_ server/index.jsp.
[57] Babel – molecular file format converter, by Pat Walters and Matt Stahl, Department of Chemistry, University of Arizona, Tucson, AZ 85721, USA. Now OpenBabel; http://openbabel.sourceforge.net.
[58] RasMol – molecular visualisation program, originally by Roger Sayle. Current version by Herbert J. Bernstein, Bernstein + Sons, P.O. Box 177, Bellport, NY, USA; http://www.bernstein-plus-sons.com/software/rasmol.
[59] Xmol – X-based molecule viewer and format converter, produced by Minnesota Supercomputer Center, Inc., USA.
[60] I.J. Bruno, J.C. Cole, J.P.M. Lommerse, R.S. Rowland, R. Taylor and M.L. Verdonk, *J. Comput. Aided Mol. Des.* **11** (1997), 525–537.
[61] ICSD-WWW interface, by Alan Hewat, Institut Laue-Langevin, 6, rue Jules Horowitz, BP 156 – 38042 Grenoble Cedex 9 – France; see http://icsd.ill.fr/icsd.
[62] J. Faber and T. Fawcett, The Powder Diffraction File: present and future, *Acta Crystallogr* **B58** (2002), 326–332.
[63] A.D. Mighell and J.R. Rodgers, Lattice symmetry determination, *Acta Crystallogr* **A36** (1980), 321–326.
[64] NCI Database; see http://dtp.nci.nih.gov, ISIS-compatible version made available by MDL Information Systems [30]; see http://www.mdl.com/products/knowledge/cancer_ db/index.jsp.
[65] Molecular Diversity Preservation International, Matthaeusstrasse 11, CH-4057 Basel, Switzerland; http://www.mdpi.org.
[66] The screening compounds database is assembled by CDS from a number of supplier catalogues received in electronic form. Current suppliers include: Sigma-Aldrich Library of Rare compounds; Bionet Research; Maybridge; Asinex Platinum and Gold Collection; ChemBridge Corporation; Chemical Diversity; ChemStar; SPECS; TimTec Corporation; MicroSource Discovery Systems; InterBioScreen; I.F. LAB; and Enamine.
[67] Brookhaven National Laboratory, Upton NY 11973, USA; http://www.bnl.gov/world.

[68] Protein Information Resource, National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Rd., Washington, DC 20057, USA; http://pir.georgetown.edu.
[69] An electronic version of the data collated by M.O. Dayhoff at NBRF [66].
[70] National Biomedical Research Foundation (NBRF) protein sequence database, now part of the PIR-International Protein Sequence Database; see http://pir.georgetown.edu/pirwww/search /textpsd.shtml.
[71] Genetic Sequence Data Bank (GENBANK) database of nucleic acid sequences; see http://www. ncbi.nlm.nih.gov/Genbank/GenbankOverview.html.
[72] European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Data Library; see http://www.ebi.ac.uk/embl/index.html.
[73] The current CDS team consists of Dr. R.F. McMeeking, Dr. D.A. Fletcher, Dr. D. Parkin and Mr. D. Osa-Edoh. Prof. M.F. Guest took over as grant holder in 2004 from Prof. P.J. Durham, who became the grant holder in 1997 following the retirement of Dr. H.J. Sherman.
[74] Macromedia Flash – popular system for delivering multimedia content, Macromedia, Inc., 600, Townsend Street, San Francisco, CA 94103, USA; http://www.macromedia.com.
[75] Fachinformationszentrum Karlsruhe, PO box 2465, D-76012 Karlsruhe, Germany; http://www.fiz-karlsruhe.de.
[76] Toth Informations Systems, Inc., 2045 Quincy Avenue, Ottawa, ON K1J 6B2, Canada; http://www. tothcanada.com.
[77] National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899-2310, USA; http://www.nist.gov.
[78] DJSM – reaction database taken from the Journal of Synthetic Methods, produced by Thomson Derwent; http://www.thomsonderwent.com.
[79] Organic syntheses – an electronic version of the entire series of Organic Syntheses (first published in 1921), by Organic Syntheses, Inc.; http://www.orgsyn.org.
[80] P. Murray-Rust and H.S. Rzepa, Towards the Chemical Semantic Web. An introduction to RSS, Internet J. Chem., 2003, 6, article 4. (http://www.ijc.com) See also The World Wide Molecular Matrix (http://wwmm.ch.cam.ac.uk).
[81] The CDS Management Advisory Panel (MAP) was established in 1990. The MAP has been chaired by Prof. W. Clegg (Newcastle), Prof. P. Willett (Sheffield), Prof. D.W.H. Rankin (Edinburgh) and Dr. K.A. Johnson (Liverpool). For further details about the MAP and its current membership see http://cds.dl.ac.uk/cds/service_ info/map.html.