Acta Crystallographica Section B
Structural
Science

ISSN 0108-7681

The Cambridge Structural Database: a quarter of a million crystal structures and rising

Frank H. Allen

Cambridge Crystallographic Data Centre (CCDC), 12 Union Road, Cambridge CB2 1EZ, England

Correspondence e-mail: allen@ccdc.cam.ac.uk

Received 11 February 2002
Accepted 28 February 2002

The Cambridge Structural Database (CSD) now contains data for more than a quarter of a million small-molecule crystal structures. The information content of the CSD, together with methods for data acquisition, processing and validation, are summarized, with particular emphasis on the chemical information added by CSD editors. Nearly 80% of new structural data arrives electronically, mostly in CIF format, and the CCDC acts as the official crystal structure data depository for 51 major journals. The CCDC now maintains both a CIF archive (more than 73000 CIFs dating from 1996), as well as the distributed binary CSD archive; the availability of data in both archives is discussed. A statistical survey of the CSD is also presented and projections concerning future accession rates indicate that the CSD will contain at least 500000 crystal structures by the year 2010.

1. Introduction

In October 2001, the CCDC passed a major milestone by archiving the 250000th small-molecule crystal structure to the Cambridge Structural Database (CSD; Allen *et al.*, 1979, 1991; Allen & Kennard, 1993). The ongoing creation and maintenance of the CSD has been the core activity of the CCDC since its inception in 1965, and the CSD system – the database and its associated access software (Bruno *et al.*, 2002) – is now used in 109 industrial companies and by 826 academic institutions in 58 countries worldwide.

The CCDC was established at the Department of Chemistry, Cambridge University, to compile a database containing comprehensive information on small-molecule crystal structures, *i.e.* organics and metallo-organic compounds containing up to 500 non-H atoms, the structures of which had been determined by X-ray or neutron diffraction. A specific aim was to store the key numerical results of each analysis, namely the cell parameters, space group and atomic coordinates, making the CSD one of the first numerical scientific databases in the world, and the first to store experimental three-dimensional chemical structure information. The CCDC was established by Dr Olga Kennard as part of the organic crystallography group, and with just two group members and some part-time scientific and clerical assistance assigned to the project, under a grant from the (then) UK Office for Scientific and Technical Information. In the late 1960s just a few hundred structures were published each year, and about 2000 structures published before 1965 were gradually incorporated into the developing database, often using printed compendia,

Table 1
Summary of information content of the Cambridge Structural Database.

Bibliographic and chemical text
Compound name(s), systematic and trivial
Amino acid sequence for peptides
Chemical formula
Authors' names
Journal name and literature citation
Text indicating special experimental conditions or results (*e.g.* neutron study, powder study, non-ambient temperature or pressure, absolute configuration determined *etc.*)
Chemical class (*e.g.* alkaloid, steroid *etc.*)
Text comment concerning disorder, errors located during validation and special structural features
Chemical connection table (see text and Fig. 2)
Formal two-dimensional chemical structure diagram in terms of atom and bond properties
Bond types used in CSD connection tables are: single, double, triple, quadruple (metal–metal), aromatic, delocalized double and π-bonds
Crystal structure data
Cell dimensions and s.u.'s
Space group and symmetry operators
Atomic coordinates and s.u.'s for the crystal chemical unit (see text)
Derived information
Bit-encoded screen records (see text)
Matching of two-dimensional and three-dimensional connectivity representations (see text)
Reduced cell parameters
Z', the number of chemical entities per asymmetric unit
Calculated density

such as the IUCr's *Structure Reports* (1939–1985) volumes, to locate original literature references.

Early software development centred on systems for validating and storing the accumulated information (see *e.g.* Allen *et al.*, 1979). However, systems for search, retrieval, analysis and visualization of CSD information began to be developed in the late 1970s, and were considerably enhanced during the 1980s (Allen *et al.*, 1991) to include full two-dimensional and three-dimensional substructure search capability and the ability to locate intermolecular nonbonded contacts (Allen & Kennard, 1993). The CSD system continues to be enhanced; the latest software developments are described by Bruno *et al.* (2002).

During this period also, the CSD began to be used extensively as a basis for fundamental research (Bürgi & Dunitz, 1983; Allen *et al.*, 1983; Bürgi & Dunitz, 1994), variously denoted as 'structure correlation' or 'knowledge acquisition', forerunners of the modern, and semantically questionable, term 'data mining'. The uptake of the CSD as a research tool in academia, and the advent of computational chemistry methods in many major pharmaceutical and agrochemicals companies, led to a rapid increase in CSD subscriptions during the 1980s. In 1989, the CCDC, then with about 20 staff, became an independent self-financing non-profit institution and was granted UK charitable status.

The modern CCDC now has 45 full-time staff. In addition to the Executive Director, Dr David Hartley, and the Scientific, Development and Business Directors, a total of 15 editorial staff are responsible for the CSD itself, nine work on the development of new software products, five are responsible for the computing infrastructure, release preparation and software for database creation, four work on research projects, four are responsible for customer support and marketing operations, and there are four business, administrative and secretarial staff. The Executive and Scientific Directors are responsible to an International Board of Governors comprising seven distinguished scientists and a financial expert. The CCDC retains close links with Cambridge University, and is recognized by the University as an institution qualified to train postgraduate students. The Centre hosts visiting scientists and also collaborates widely with universities and industrial organizations, both within the UK and internationally.

Structural crystallography has, of course, changed out of all recognition since the mid-1960s. Improvements in data collection, structure solution and refinement techniques have gone hand in hand with dramatic increases in computing power. As a result, more than 24500 structures were archived to the CSD in 2001, representing a near 40-fold increase in worldwide crystallographic productivity compared with 1965. This paper summarizes the current status of the CSD and uses statistics of database growth, together with an analysis of current trends in the subject, to make some observations about future trends. Other papers in this special issue of *Acta Crystallographica* review the more recent scientific applications of the CSD in organic chemistry and crystal chemistry (Allen & Motherwell, 2002), molecular inorganic chemistry (Orpen, 2002), and the life sciences (Taylor, 2002).

2. Information content of the CSD

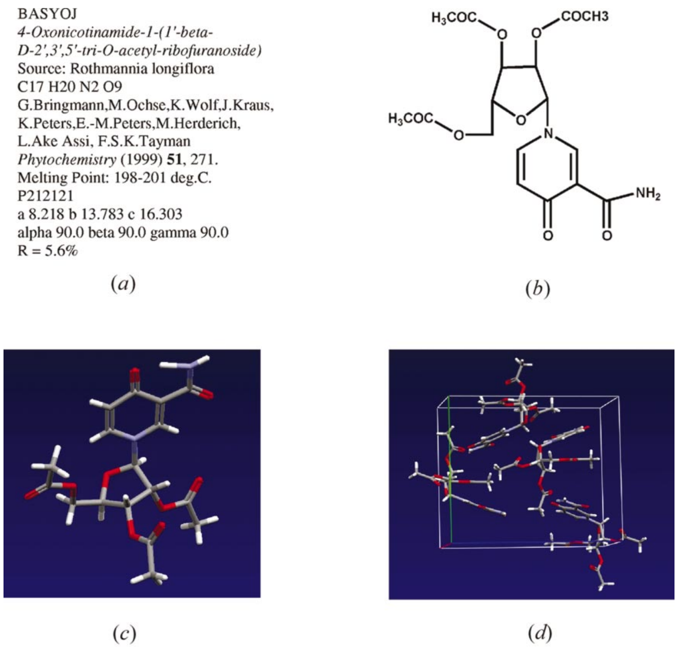
Each individual crystal structure determination forms an entry in the CSD, which is identified by a reference code: six letters identify the chemical compound and two supplementary digits identify additional determinations of the same structure, *e.g.* an improved refinement, studies by different scientists, studies under different experimental conditions *etc.* The information content of each entry is illustrated in Fig. 1 and is summarized in Table 1. The most important information item added by CCDC staff is the two-dimensional chemical structure representation (Fig. 2). The atom and bond properties are converted into a compact connectivity table for CSD storage, and form the basis for substructure searching (Bruno *et al.*, 2002) at the molecular and supramolecular levels.

Each connection table is analysed to assign cyclic/acyclic flags to chemical bonds and to generate a bitmap or 'screen' record. This contains codified 'yes/no' information concerning the presence/absence of specific substructural features in each chemical diagram, *e.g.* atoms with specific connectivity patterns, common functional groups, rings of specific sizes *etc.* The screens are used as heuristics to speed substructure searching: screens generated from a query substructure must all be matched in a candidate CSD entry before that entry is further analysed using CPU-intensive atom-by-atom, bond-by-bond matching. Bit-screens are also employed to encode information about (*a*) elemental constitution, (*b*) letter sequences in author and compound names, (*c*) summary information about data content of the entry, and (*d*) results of the data validation process. These bitmaps are used to speed up specific searches or are available to users as secondary search criteria.

It is important to draw a distinction between the formal connection table and the chemical diagram displayed by CSD system software. While the connection tables held in the CSD describe all atoms and bonds, the displayed diagrams often contain group-symbol abbreviations, *e.g.* Ph, Me, OAc *etc.*, to reduce graphical overlap and improve visual perception (see Fig. 3*a*). Over time, however, the structures characterized by diffraction methods have increased in size and complexity. This is particularly true of metal coordination complexes, an area of chemistry where crystallography has always been *the* vital analytical tool. Many of these structures are inherently three-dimensional and it is sometimes difficult to generate useful two-dimensional visual representations (see *e.g.* Fig. 3*b*). Since 2001, CCDC scientific editors have had the option to generate and store automatically the linear ligand-based formulations exemplified in Fig. 3(*c*). These representations then replace the two-dimensional diagram in CCDC display software [such as *ConQuest* (Bruno *et al.*, 2002)].

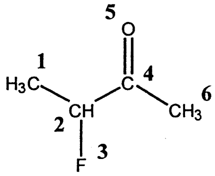
Importantly also, CSD system software such as *PreQuest* (Motherwell *et al.*, 2002), *ConQuest* and *Mercury* (Bruno *et al.*, 2002) can also display chemical bond types in their three-dimensional structure representations (Fig. 4), to aid the chemical interpretation of CSD structures. This is possible because the two-dimensional and three-dimensional structure representations are linked by a graph-theoretical atom-by-atom and bond-by-bond matching, which maps the chemical information of the two-dimensional diagram onto the atoms and bonds of the three-dimensional crystal structure.

Coordinate sets entering the CSD validation process refer, of course, to a crystallographic asymmetric unit, although published coordinate lists may not always comprise atoms from the same asymmetric unit. Once validation is completed successfully, the coordinate data finally stored in the CSD will always describe unique bonded network(s) plus any single-atom species. This assembly of bonded networks and ions is referred to as the 'crystal chemical unit' (c.c.u.). Often, the c.c.u. is synonymous with the asymmetric unit, but when molecular symmetry coincides with crystallographic space-group symmetry, the asymmetric unit is some fraction of a complete molecule. In this situation, the atoms of the asymmetric unit, plus the symmetry-generated atoms which complete the chemical molecule, collectively form the c.c.u., and will be recorded in the CSD. Each atom that is symmetry-generated from the coordinates of the asymmetric unit contains a tag which identifies the symmetry operator that was applied to generate the coordinates.



BASYOJ
4-Oxonicotinamide-1-(1'-beta-D-2',3',5'-tri-O-acetyl-ribofuranoside)
Source: Rothmannia longiflora
C17 H20 N2 O9
G.Bringmann,M.Ochse,K.Wolf,J.Kraus, K.Peters,E.-M.Peters,M.Herderich, L.Ake Assi, F.S.K.Tayman
Phytochemistry (1999) **51**, 271.
Melting Point: 198-201 deg.C.
P212121
a 8.218 b 13.783 c 16.303
alpha 90.0 beta 90.0 gamma 90.0
R = 5.6%

Figure 1
Schematic view of the information content of the Cambridge Structural Database.



Atom Properties

Atom Number	1	2	3	4	5	6
Element Number	C	C	F	C	O	C
No. Connected Non-hydrogen Atoms	1	3	1	3	1	1
No. Terminal Hydrogen Atoms	3	1	0	4	0	3
Net Charge	0	0	0	0	0	0

Bond Properties

Atom 1 of Bond	2	2	2	4	4
Atom 2 of Bond	1	3	4	5	6
Bond Type	1	1	1	2	1

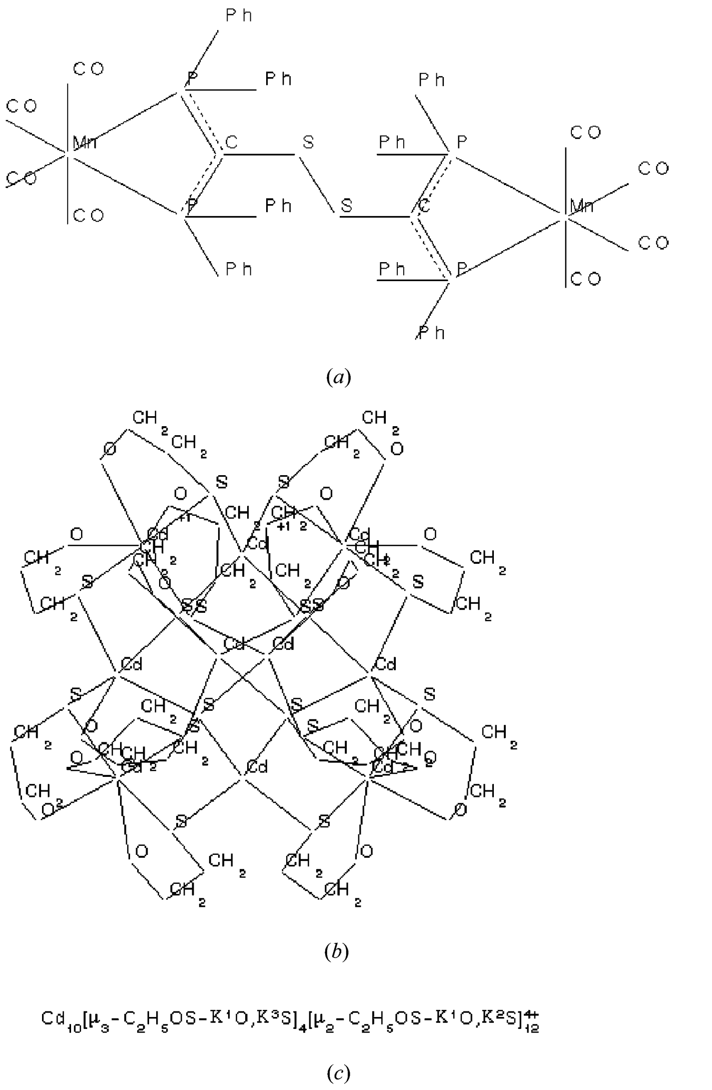
Figure 2
Two-dimensional chemical connectivity representation for a simple organic molecule. Reproduced with permission from Allen & Hoy (2001). Copyright (2001) International Union of Crystallography.

The standard uncertainties (s.u.'s) of cell parameters and atomic coordinates are included in the CSD for entries published since *ca* 1985. Work is now in hand to make these available to database users in CIF files output by the *ConQuest* program (Bruno *et al.*, 2002), and to make use of these data in reporting geometrical parameters calculated within the CSD system software.

Disorder has always presented special problems within the CSD. Until the late 1980s it was CCDC policy to delete the coordinates of minor occupancy sites or, in the case of exact twofold disorder, to select one set of sites for retention. Occasionally, where disorder was very complex, usually affecting all atoms of a molecule, then no coordinates were retained. Since the late 1980s, *i.e.* for the vast majority of current CSD entries, CCDC staff have been able simply to 'suppress' atomic sites that would otherwise have been deleted. The coordinates of these sites have been retained within the CSD, and software is now being planned to improve the analysis and representation of disordered structures. Indeed, ongoing work to improve the information content of the CSD itself is closely related to the recent availability of structural data in electronic form *via* the CIF format.

3. Data acquisition

The universal acceptance of the CIF format (Hall *et al.*, 1991; Brown & McMahon, 2002), adopted as an international standard by the IUCr and rapidly incorporated as an output format by the major crystallographic software packages, has changed the CCDC's data acquisition methods dramatically in the past 5 years. Led by *Acta Crystallographica*, for which electronic CIF-based submission moved rapidly from being 'preferred' to mandatory, the majority of other journals that carry significant crystallographic content now 'advise', 'urge strongly' or 'require' supplementary data to be submitted



(*a*)

(*b*)

$Cd_{10}[\mu_3\text{-}C_2H_5OS\text{-}K^1O,K^3S]_4[\mu_2\text{-}C_2H_5OS\text{-}K^1O,K^2S]_{12}^{4+}$

(*c*)

Figure 3
Two-dimensional structure representations for metallo-organic structures. (*a*) Use of 'group symbols' (*e.g.* Ph) minimizes atomic overlap problems. (*b*) Atomic overlap in a two-dimensional representation (counterions and solvent molecules omitted). (*c*) Linear ligand-based representation of the two-dimensional structure in (*b*).

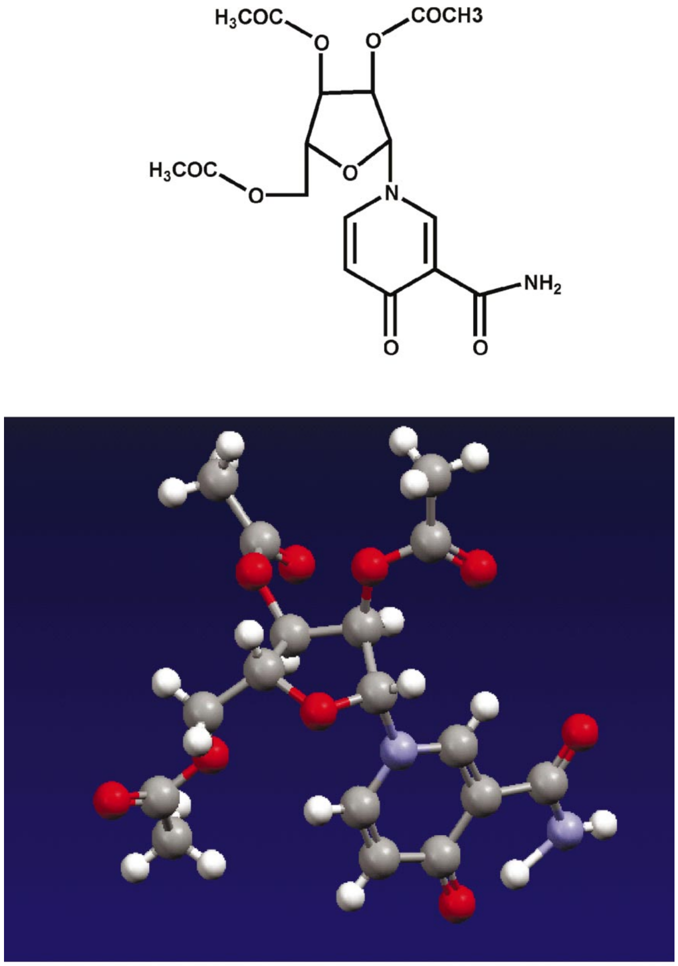


Figure 4
Chemical bond types imposed onto a three-dimensional structure representation in the *Mercury* visualiser, using the two-dimensional:three-dimensional connectivity matching stored in the CSD (see text).

electronically in CIF format. The days when CCDC staff needed to re-keyboard crystal structure information from printed manuscripts, or worse from deposition documents of variable quality and sometimes haphazard organization (Bergerhoff *et al.*, 1986), have receded. Indeed, the crystallographic community has itself played a significant role in speeding the acceptance of electronic depositions by major journals. Even when hard-copy depositions are still a requirement, the community has helped enormously by their willingness to provide electronic copies to the CCDC and other crystallographic databases.

In numerical terms, the percentages of published structures for which electronic data have been received in each year since 1997 are: 30.2% (1997); 47.4% (1998); 61.3% (1999); 72.9% (2000); 80.1% (2001). The residual material arises from hard-copy depositions lodged with journal offices, particularly for journals with limited crystallographic content, or from hard-copy depositions (some in CIF format!) which have been scanned to form downloadable pdf documents. The CCDC is working to form improved relationships with all journals, so that they may be added to a list of more than 51 high-yielding international journals for which the CCDC acts as the official data depository.

For journals which are within the official scheme, authors are requested to send data to the CCDC just before their manuscript is submitted to the journal. The CCDC then resolves any format problems in the CIF, and returns a CCDC Deposition Number for inclusion in the manuscript and in the published paper. Since the data and text are linked by the CCDC Deposition Number, relevant published papers can readily be identified during the CCDC's regular journal scanning activities. The CCDC's direct in-house scanning covers 81 journals and is backed up by automated searches of *Chemical Abstracts* to locate crystal structure publications that appear in less-common primary sources. Further details of the CCDC's pre-publication archive are given later in the paper.

While electronic input has helped the CCDC enormously, particularly in eliminating typographical or keyboarding errors, it has introduced some new validation problems. In order to prepare a CIF for journal deposition, authors must edit the file generated by a crystallographic software package to add, *e.g.* their names and addresses, data items such as compound names, space-group symbol, crystal system, crystal colour, habit, melting point *etc*., that are unknown to their software, together with text fields describing special features of the structure determination. CIF format requirements are quite strict and manual editing is prone to generate format violations. These problems, which affect about 40% of incoming CIFs, are located and corrected by CCDC staff before the CIF enters the pre- or post-publication archives (described below in more detail). The CCDC has now written CIF editing software, the *enCIFer* program (Smith & Johnson, 2002), which will be made available for free download from the CCDC Website and also through the software systems of a number of diffractometer suppliers. The program not only identifies format violations, but also allows interactive editing of new or existing CIF data items and prompts for additional data (see *e.g.* the list above) that will improve the information richness of the resulting CIF archive and CSD entry.

The vast majority (>99%) of CSD entries arise from published work spread over 966 cited journals. However, since 1976 the CCDC has encouraged *Personal Communications* of crystal structure data that would otherwise be lost to the scientific community. For these, each CSD entry records the name(s) and institutional address(es) of the depositor(s). Only 262 such entries were deposited in the period 1976–1995, but since 1996 a further 859 structures have been deposited *via* this route, or 77% of the current total of 1121 *Personal Communications*.

4. Data processing and validation

The addition of chemical information forms an important part of data processing activities: the generation and encoding of chemical connectivity information (Fig. 2), the construction of a two-dimensional chemical diagram or linear formula (Fig. 3), and the checking or addition of systematic chemical compound names and any common synonym name(s). Scientific editors also add any CSD-required information that may be in the printed paper, but has not been included in the CIF deposition.

The CCDC database-building program, *PreQuest* (Motherwell *et al.*, 2002), provides two principal routes for two-dimensional connectivity and diagram generation:

(i) by direct sketching on a Cartesian or hexagonal grid, with facilities to use or modify pre-drawn templates, or apply appropriate symmetry operators to an asymmetric section of a molecule, which is a particularly valuable feature when representing many metal complexes, and

(ii) by software generation of a two-dimensional diagram directly from the three-dimensional crystal structure information.

Diagrams generated using this latter method almost always require subsequent editing, since the software generation of bond types from element types and geometry is not always reliable, and the visual layout and orientation of the diagrams may be imperfect. The use of chemical group symbols, or the presentation of the structure in the linear format of Fig. 3, are under editorial control within the *PreQuest* software.

PreQuest is a successor to the CCDC's original *UNIMOL* program (Allen *et al.*, 1974), in which a central feature was the re-computation of molecular geometry and its comparison with published values. This enables errors in the reported atomic coordinates, cell dimensions and space-group symmetry to be detected, the pattern of geometry discrepancies often being indicative of the source(s) of the error(s). Since the vast majority of such errors were typographical in origin, the editorial work arising from these sources has reduced very significantly over the past few years as the quantity and quality of electronic input has increased. Within the *PreQuest* implementation, the location and correction of such errors has, in any case, been speeded up by increased use of graphical techniques.

Table 2
CSD overview statistics, 30 October 2001.

(*a*) General statistics

Number of structures	251 515
Number of chemical compounds	227 662
Number of atoms with three-dimensional coordinates	15 123 772
Number of different literature sources	966

(*b*) Crystal system statistics (percentage values in parentheses)

Space group completely defined	247 966
Space group is centrosymmetric	188 167 (75.9)
Space group is non-centrosymmetric	59 799 (24.1)
Triclinic	55 277 (22.3)
Monoclinic	132 490 (53.4)
Orthorhombic	50 548 (20.4)
Tetragonal	5914 (2.4)
Trigonal/rhombohedral	2615 (1.1)
Hexagonal	2917 (1.2)
Cubic	1285 (0.5)

The graphical windows of *PreQuest* also permit extensive chemical checking, *inter alia*:

(i) the checking of crystal data against chemical constitution, and

(ii) the cross-checking of the connectivity indicated in the two-dimensional chemical diagram against the connectivity computed from the coordinate data using a set of standard (but adjustable) covalent radii.

This ensures that the correct crystal structure data are associated with the correct chemical compound in (the many) multi-structure papers and that the bonding implied by the chemical diagram is reflected in the published crystal structure. This latter check is particularly important in novel organometallic and metal complex structures, where adjustments to standard radii may be required to obtain a two-dimensional/three-dimensional match, or the assignment of certain bonds (and their types) may be unclear and interaction with the author(s) is required.

The most important crystallographic operation now is the description of disordered structures, which may not be as clear as it might be in many incoming data sets. As indicated above, present policy is to retain a single set of disordered site(s) that corresponds to a complete chemical entity, the other sites being retained as 'suppressed' atoms in the master CSD archive. Improvements in disorder handling are currently in hand.

A feature of database building over many years has been the collegial interactions with the crystallographic community. Often, use of the *PreQuest* program can suggest corrections to data errors or inconsistencies. The majority of these are relatively simple, and CSD editors will implement these routinely and include a text record to indicate the nature of the error and its solution. However, for serious inconsistencies or error indications, CSD editors will always refer the problem back to the original author(s) for clarification and resolution, or for confirmation of their suggested solution to the problem.

5. Archiving, availability and distribution

The CCDC now maintains three data archives, as follows:

(i) The secure pre-publication archive of CIFs deposited prior to submission to a journal with which the CCDC has a formal agreement. Some pre-publication depositions are also received which are intended for journals outside the formal scheme. The depositor is issued with a CCDC Deposition Number (DepNum) to include in their manuscript and which will appear in the published article in 'agreement' journals, a scheme modelled on the Protein Data Bank ID code mechanism (Berman *et al.*, 2002). Such numbers may also appear in other journals at the request of the authors. Data in this archive are held securely on trust and are only provided to staff of the journal concerned and to its *bona fide* referees. If a paper is rejected by any journal and resubmitted elsewhere, then the DepNum remains associated with the data, even if data are revised and re-communicated to the CCDC by the authors or the journal. Thus, CCDC staff can use the printed DepNum to link data to any eventual publication, wherever it appears.

(ii) The post-publication CIF archive comprises CIFs moved across from (i) after publication, together with other CIFs arriving after publication of an article in any other journal. Data from this archive enters the CSD processing and validation system and is then filed to the CSD binary archive (iii) described below. The CCDC will freely supply individual CIF datasets from the post-publication archive to any scientist who requests them, whether they are CSD subscribers or not. These CIFs contain all the data supplied by the authors to satisfy the requirements of the journal concerned, including data such as atomic displacement parameters *etc.*, which are not currently available in the distributed binary ASER file (iii) below. Supply of individual data sets is automated using a Web-based request form available *via* the CCDC Web site.[1] The post-publication archive currently (30 October 2001) contains CIFs for 49908 structures, with the majority corresponding to the period from 1998 onwards.

(iii) The distributed CSD archive, held in the CCDC's binary ASER format. This file contains all value-added data items, *e.g.* chemical connection tables, coordinates transformed to the c.c.u. basis, processing flags and text *etc.*, and structured for search, analysis and display using CCDC software (Bruno *et al.*, 2002). The complete and growing ASER file is supplied twice yearly (April and October) on CD-ROM to all CSD subscribers. At the time of writing (30 October 2001), the CCDC is piloting the Web-based availability of inter-release CSD entries, so that subscribers may have ASER files that are as up to date as possible. As the CIF archive (ii) contains only relatively recent data, the CCDC will also freely supply earlier individual data sets to non-subscribers. These data sets are supplied as a CIF containing those data items available in ASER which would normally have been deposited

[1] The CCDC Web site, http://www.ccdc.cam.ac.uk/, contains full current information about the availability of the Cambridge Structural Database system. It also provides e-mail addresses for the deposition of structural data, the CCDC help desk and for administrative enquiries.

with a journal prior to publication. Full details of data availability and the (developing) systems for handling individual requests can be found on the CCDC Web site.

Table 3
Comparative CSD entry information statistics.

	Inclusive of 1983		Inclusive of 1990		30 October 2001	
	Structures	%	Structures	%	Structures	%
Total structures	52 363	100.0	104 380	100.0	251 515	100.0
Organic structures	28 995	55.4	52 450	50.3	112 113	44.6
Transition metal present	20 439	38.9	45 588	43.7	120 638	48.0
Li–Fr or Be–Ra present	2887	5.5	5299	5.1	13 471	5.4
Main group metal present	2206	4.2	5024	4.8	16 171	6.4
Three-dimensional coordinates present	37 318	71.0	83 884	80.4	223 920	89.0
Error-free coordinates	35 032	93.9†	80 372	95.8†	219 864	98.2†
Error records added	5629	10.7	11 225	10.8	16 371	6.5
Neutron studies	567	1.08	786	0.8	1062	0.4
Low/high temperature studies	3275	6.2	9943	9.5	55 752	22.2
Absolute configuration determination	1330	2.5	2344	2.2	4924	2.0
Disorder present in structure	4943	9.4	13 594	13.0	45 728	18.2
Polymorphic structures	3231	6.1	4618	4.4	7892	3.1
R factor < 0.100	37 190	70.8	85 389	81.8	227 181	90.3
R factor < 0.075	29 937	57.0	73 424	70.3	202 848	80.7
R factor < 0.050	15 974	30.4	42 996	41.2	125 112	49.7
R factor < 0.030	2231	4.2	7150	6.8	22 346	8.9
n(atoms)/structure‡	44	–	53	–	76	–
Mbyte data added in year§	14	–	29	–	74	–

† Taken as a percentage of structures for which coordinates are present in the CSD. ‡ Average number of atoms per structure in the year cited in the column heading. The figure for 1970 was 27. § Number of Mbytes of data added to the CSD in the year cited in the table heading; the value in the 30 October 2001 column is for 2000 (the last complete year). The figure for 1970 was 2 Mbyte.

Table 4
Journal publication statistics for 1999 and 2000.

(*a*) Top 20 Journals by number of CSD structures published (*N*) and percentage of all structures published in 1999 and 2000 (%)†

	N	%
Organometallics	3100	9.02
J. Chem. Soc. Dalton Trans.	2919	8.49
Inorg. Chem.	2889	8.40
Acta Cryst. Sect. C	2328	6.77
J. Am. Chem. Soc.	2009	5.84
J. Organometal. Chem.	1929	5.61
Inorg. Chim. Acta	1394	4.06
Chem. Commun.	1159	3.37
Eur. J. Inorg. Chem.	1129	3.28
Polyhedron	1126	3.27
J. Org. Chem.	1062	3.08
Z. Anorg. Allg. Chem.	1053	3.06
Angew. Chem. Int. Ed. Engl.	986	2.87
Chem.-A Eur. J.	838	2.43
Tetrahedron	696	2.02
Tetrahedron Lett.	604	1.76
Z. Naturforsch. B	529	1.54
CSD Personal Communications	514	1.50
Eur. J. Org. Chem.	490	1.43
Z. Krist. New Crystal Struct.	488	1.42
Total	27 242	79.24

(*b*) Top five journal publishers by numbers of crystal structures published (*N*) and percentage of all structures published in 1999 and 2000 (%)†

	N	%
American Chemical Society	9611	27.96
Elsevier	6413	18.65
UK Royal Society of Chemistry	5353	15.57
Wiley–VCH	4496	13.08
IUCr	2760	8.03
Total	28 633	83.29

† Data for 1999 cover 17 867 structures. Data for 2000 cover 16 510 structures and do not yet include all structures from a small number of less-common journals.

6. Statistics

Tables 2 and 3 present a statistical overview of the quarter of a million structures in the CSD on 30 October 2001. Table 3 is more indicative of trends over time, since it compares current statistics for a variety of CSD information fields with those from two earlier snapshots of the database, taken at the ends of 1983 and 1990. These two year-ends were chosen since the CSD contained *ca* 50 000 and *ca* 100 000 structures, respectively, at those times. Table 4 presents statistics for those journals that have published the greatest numbers of structures during 1999 and 2000.

All of the trends of Table 3 derive from the radical improvements in structure solution and refinement methodologies, the scientific and technical enhancements in data collection equipment, and dramatic increases in computing resources that have taken place over the past 35 years. The most obvious and significant trend is the rapidly accelerating world productivity of small-molecule crystal structures, discussed later, coupled with a similar increase in complexity of the structures being reported: from an average of 27 atoms per structure in 1970, to a current average of 76. There has also been a continuing change in the types of compounds being characterized by crystallographic methods. Thus, the percentage of organic compounds studied has fallen from 55.4% to 44.6% of the CSD in 2001, with a corresponding increase in the proportion of organometallics and metal complexes. Novel compounds in this latter class are now routinely studied by X-ray analysis to obtain the most reliable determination of connectivity and three-dimensional structure. It is no accident, therefore, that seven of the top-ten journals of Table 4 are wholly devoted to this area of chemistry, and together yield about 42% of current CSD input. It is from this area that the increased structural complexity of the current CSD largely derives. The proportion of compounds of (i) the Groups 1 and 2 metals, and (ii) the main group metals have remained almost constant over time, each contributing about 5% of CSD content.

Table 3 shows clearly that the availability of coordinate data associated with published reports of crystal structures has also increased significantly over time. The current overall figure is 89%, but is 95% for the period from 1991 onwards. Many of the CSD entries which lack coordinates arose from preliminary publications in the days before coordinate depositories were operated by journals. Nowadays, all major journals make

Table 5
Expansion of the CSD 1976–2000.

Year	Total structures at end of year	Number added in year	Increase on previous year (%)	Average annual increase over 5 year period (%)	CSD doubling period in years
1975	14 066	3088	27.2		
1980	24 532	4690	19.1	22.9	3.6
1985	57 612	6879	13.6	16.0	4.7
1990	95 754	8008	9.1	11.1	6.7
1995	149 758	11 572	8.3	9.2	7.9
2000	231 866	17 866	8.3	9.1	8.0

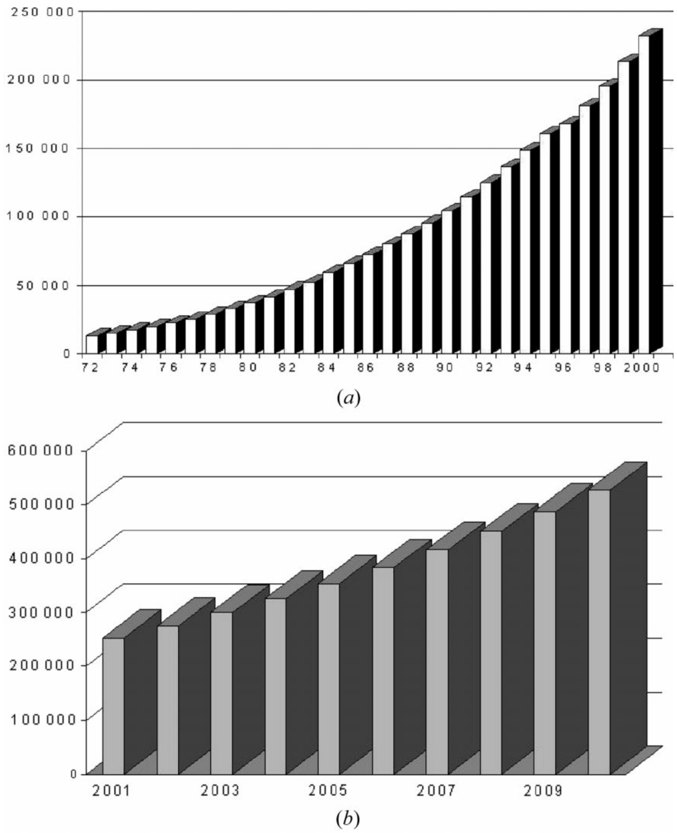
appropriate arrangements to preserve supplementary data and, in the case of crystal structure data, they have formal arrangements with the crystallographic databases. Today, data deposition is principally electronic, facilitated by the CIF format. The value of the electronic delivery in preserving the integrity of numerical crystal structure data is also shown by the significant reduction in the proportion of structures having ERROR records incorporated into their CSD entries during the validation process. The current overall proportion of 6.5% of entries (Table 3) is composed of two very distinct phases: a rate of more than 11% which existed prior to 1991, and a rate of less than 4% from 1991 to date.

Other indicators of note in Table 3 reflect significant improvements in experimental equipment and in overall data quality. Thus, (i) the proportion of non-room-temperature (principally low-temperature) studies has increased to 22.2% overall, and 31% for the past decade, while (ii) the ability to resolve disorder situations has risen steadily to an overall figure of 18.2%. Conversely, the number of neutron studies published annually has remained almost static and their overall proportion in the CSD has decreased to 0.4%. The number of structures determined from powder diffraction data is now 370, a number which is surely set to rise. The most direct indicators of improved data quality are the R-factor statistics over time, as shown by the percentages of structures having $R < 0.075$ (increased from 57.0 to 80.7%) and $R < 0.050$ (up from 30.4 to 49.7%) in moving from the 1983 CSD snapshot to the present day (Table 3). A more detailed study of structural precision based on CSD data has been presented by Allen *et al.* (1995*a*,*b*).

7. CSD growth: past, present and future

Fig. 5(*a*) shows the cumulative rate of growth of the CSD for the period from 1970. It is more instructive to study growth over 5-year periods (Table 5) and to express the data in terms of structures added in each period or year. These accession-year statistics differ from the publication-year data that have been used to assemble Tables 2, 3 and 4. Table 5 shows that while the number of structures archived per annum to the CSD has increased by a factor of 4, the doubling period for the database has itself more than doubled since 1980, from 3.6 to 8.0 years. However, data from 1990 onwards would indicate that expansion rates are settling down, and it is possible to use the data of Table 5 to project the numbers of structures that will need to be processed each year over the next decade. The projected data indicate accession rates of *ca* 28000 structures in 2005 and 40000 structures in the year 2010, generating a CSD that contains well over half a million structures by the end of that year, as shown in Fig. 5(*b*).

These projections are based, of course, on a continuation of current methods of placing crystal structure data into the public domain. They do not take account of any significant surge in publication rates caused by the increased application of CCD detector technology, nor can they take account of any changes that may occur in publication strategies over time. At present, the rate-determining step in the publication process appears to be the human task of writing a paper, but even this barrier is now being reduced by the advent of electronic-only journals. For crystallography, *Acta Crystallographica, Section E: Structure Reports Online* (http://www.iucr.org/) has been created to cater specifically for the rapid publication of electronic reports of crystal structure data, with peer review, coverage by major abstracting services, and rapid entry into the CSD. It is to be hoped that this initiative will attract an increasing proportion



(*a*)

(*b*)

Figure 5
CSD growth statistics. (*a*) Growth of the CSD 1970–2000. (*b*) Projected growth of the CSD 2001–2010.

of the known reservoir of unpublished crystal structures into the public domain. Direct deposition of crystal structure data into the CSD remains an option though and the number of *Personal Communications* places this source at 18th in the 1999/2000 journal statistics presented in Table 4.

Clearly, the CCDC will process at least as many structures in the next nine years as it has in the first 36 years of its existence. If we take account of the gradually increasing size of structures over time, then the amount of information that will enter the CSD by 2010 will nearly triple its size in megabyte terms. However, it is likely that these figures are minima, especially in view of the ongoing development of rapid electronic publication and data deposition routes. Completeness is an important criterion in judging the value of any database, and the CCDC continues to work with the crystallographic and publishing communities to maximize the information content of the CSD for the future benefit of the scientific community.

The CCDC would like to thank the many scientists who have contributed to the CSD, and for the cordial and constructive relationships that have been built up with that community over many years. It is also a pleasure to acknowledge the work of all CCDC staff, past and present, who have contributed so much to the creation and maintenance of the database. The current CCDC Database Group comprises Stella Barrett, Jenny Field, Kathleen Foreman, Jan Vincent, Gill Heale, Stephen Holgate, Aubrey Prout, Tracy Allgood, Gwenda Kyd and Matthew Lightfoot (Scientific Editors), Teresa Linstead, Julie Lister, Lorna Rouse and Jodie Flack (Editorial Assistants), and Cheryl Cook, Sarah Martin and Sue King (Data Preparation). They use software currently written and maintained by Owen Johnson, Greg Shields, Sam Motherwell, Clare Macrae and Barry Smith.

References

Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T. W. A., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. & Watson, D. G. (1979). *Acta Cryst.* B**35**, 2331–2339.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*a*). *Acta Cryst.* A**51**, 95–111.

Allen, F. H., Cole, J. C. & Howard, J. A. K. (1995*b*). *Acta Cryst.* A**51**, 112–121.

Allen, F. H., Davies, J. E., Galloy, J. J., Johnson, O., Kennard, O., Macrae, C. F., Mitchell, E. M., Mitchell, G. F., Smith, J. M. & Watson, D. G. (1991). *J. Chem. Inf. Comput. Sci.* **31**, 187–204.

Allen, F. H. & Hoy, V. J. (2001). *The Cambridge Structural Database (CSD)*. International Tables for Crystallography, Volume F, *Crystallography of Biological Macromolecules*, edited by M. G. Rossmann and E. Arnold, ch. 24.3. Dordrecht/Boston/London: Kluwer Academic Publishers.

Allen, F. H. & Kennard, O. (1993). *Chem. Des. Autom. News*, **8**, 1, 31–37.

Allen, F. H., Kennard, O., Motherwell, W. D. S., Town, W. G., Watson, D. G., Scott, T. J. & Larson, A. C. (1974). *J. Appl. Cryst.* **7**, 73–78.

Allen, F. H., Kennard, O. & Taylor, R. (1983). *Acc. Chem. Res.* **16**, 146–153.

Allen, F. H. & Motherwell, W. D. S. (2002). *Acta Cryst.* B**58**, 407–422.

Bergerhoff, G., Allen, F. H., Bellard, S. A. & Lucas, C. V. (1986). *Acta Cryst.* C**42**, 1671–1675.

Berman, H. M., Battistuz, T., Bhat, T. N., Bluhm, W. F., Bourne, P. E., Burkhardt, K., Feng, Z., Gilliland, G. L., Iype, L., Jain, S., Fagan, P., Marvin, J., Ravichanran, V., Schneider, B., Thanki, N., Padilla, D., Weissig, H., Westbrook, J. D. & Zardecki, C. (2002). *Acta Cryst.* B**58**, 899–907.

Brown, I. D. & McMahon, B. (2002). *Acta Cryst.* B**58**, 317–324.

Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.

Bürgi, H.-B. & Dunitz, J. D. (1983). *Acc. Chem. Res.* **16**, 153–161.

Bürgi, H.-B. & Dunitz, J. D. (1994). *Structure Correlation.* Weinheim: VCH.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Motherwell, W. D. S., Macrae, C. F. & Shields, G. P. (2002). *PreQuest. A Program for Crystal Structure Validation*. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.

Orpen, A. G. (2002). *Acta Cryst.* B**58**, 398–406.

Smith, B. & Johnson, O. (2002). *enCIFer: A Program for Checking and Editing CIF Files*. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, England.

Structure Reports (1939–1985). Published for the International Union of Crystallography by D. Reidel, Dordrecht, The Netherlands.

Taylor, R. (2002). *Acta Cryst.* B**58**, 879–888.