International Tables for Crystallography (2006). Vol. F, Section 24.3.2, pp. 663–665.

24.3. The Cambridge Structural Database (CSD)

By F. H. Allen and V. J. Hoy

24.3.1. Introduction and historical perspective

The Cambridge Structural Database (CSD: Allen *et al.*, 1991; Kennard & Allen, 1993; http://www.ccdc.cam.ac.uk) is a fully retrospective computerized archive of bibliographic, chemical and numerical data from X-ray and neutron diffraction studies of small organic and metallo-organic molecules. Here, ‘small’ means an upper limit of about 500 non-H atoms. The CSD was established in 1965, when the number of small-molecule crystal structures published each year was just a few hundred and, for this reason, it was possible to rapidly assimilate the earlier literature. However, the advent of increasingly powerful computers and associated advances in data collection and structure solution techniques has led to an almost exponential increase in the number of crystal structures being reported (Fig. 24.3.1.1).

Since the mid-1980s, there has been an average year-on-year increase in the number of CSD entries of very close to 10%. In 1999, around 18 000 structures were added to the database, and in mid-2000 the total archive contained over 220 000 structure determinations. This makes the CSD one of the largest numerical data resources currently available in chemistry (Table 24.3.1.1). At present, about 48% of the CSD comprises metallo-organic structures, 42% are pure organics, and the remaining 10% are compounds of the main-group elements. The doubling period of the CSD is approximately 7.5 years and, if account is taken of recent advances in diffractometer technology, it is possible to project a total of at least 500 000 database entries by the year 2010.

In contrast with the Protein Data Bank (PDB: Bernstein *et al.*, 1972; Abola *et al.*, 1997; RCSB, 2000), which has always received its data through direct electronic depositions, the CSD reflects the published literature. Until recently, much of the raw input has been re-keyboarded from hard-copy documents. Thus, in the early years, Cambridge Crystallographic Data Centre (CCDC) software development concentrated on data-validation techniques designed to eliminate keyboarding, typographical and scientific errors so as to ensure the accuracy of the master archive. Validation software has recently been upgraded to take advantage of modern computing methods, particularly the rapid developments in high-resolution graphics systems.

Nevertheless, the massive growth of the database has meant that the development of fast and efficient applications software for database search, data retrieval, numerical analysis and visual display has always been a high priority. The first of these software systems became available towards the end of the 1970s and constant updates ensure that the code continues to develop in response to user needs. The CSD system (CSDS), comprising the database and its applications software, is now distributed to about 1000 academic and industrial institutions worldwide.

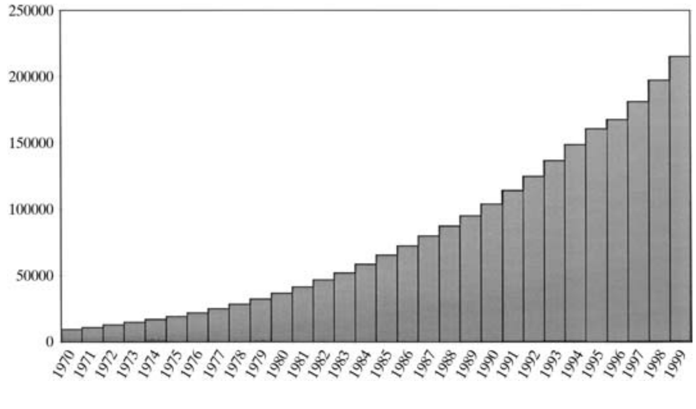


Fig. 24.3.1.1. Growth of the CSD since 1970 expressed in terms of the number of structures published per annum.

Users of the CSD span the scientific spectrum, reflecting the wide range of research applications of the data it contains. Over the past two decades, the CSDS has provided the essential basis for research projects in structural chemistry, structure correlation and the rational design of novel bioactive molecules of pharmaceutical or agrochemical interest. A variety of statistical, numerical and computational methodologies have been applied to the CSD, giving rise to the concept of knowledge acquisition, or data mining, from the ever-increasing reservoir of precise experimental results. To date, nearly 700 papers of this type exist in the literature. This activity has, in turn, raised the possibility of the generation of knowledge-based libraries of structural information from the CSD. IsoStar (Bruno *et al.*, 1997), a library of information on non-bonded interactions which was first released in 1997, is the CCDC’s first knowledge-based product. A companion knowledge base of intramolecular geometry, Mogul, is now under development, and will contain bond-length, valence-angle and torsional-angle information.

24.3.2. Information content of the CSD

24.3.2.1. *Acquisition of information*

Almost all of the information contained in the CSD has been abstracted from the published literature. Over 800 primary literature sources are cited and the earliest reference is from 1930. Much of the data has been re-keyboarded from the original literature and from hard-copy supplementary deposition documents. The CSD now acts as the official depository for some 40 major international journals. Today, an increasing proportion (around 75% in mid-2000) of the numerical information is received directly in electronic form. The switch from hard-copy input to electronic deposition has been catalysed by the development of the exchange format for crystallographic data, the crystallographic information file or CIF (Hall *et al.*, 1991). The CIF has been adopted as the standard for the subject by the International Union of Crystallography, and is now output by nearly all of the major software packages for structure determination and refinement. Development of the CIF has also led to an increase in direct private depositions of structural data to the CSD, data that, for various reasons, are unlikely to be published through formal mechanisms.

24.3.2.2. *Data organization*

Each individual structure in the CSD is referred to as an *entry* and each entry is identified by a *reference code* (*refcode*) containing six alphabetic characters, which characterize a specific chemical compound, and a further two numeric characters which trace the publication history of the structure. The information content of a

Table 24.3.1.1. *CSD statistics (August 2000)*

No. of entries	224 945
No. of compounds	202 669
No. of entries with 3D coordinates	198 136
No. of entries with error-free coordinates	194 784
No. of atoms having 3D coordinates in the CSD	12 906 283
No. of entries in the CSD-Use database	792

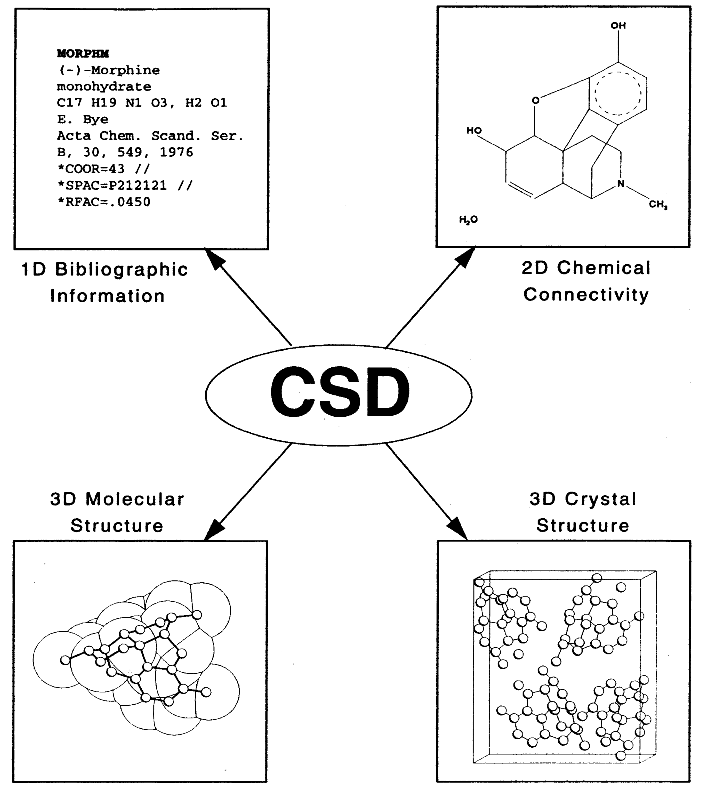


Fig. 24.3.2.1. Information content of the Cambridge Structural Database (CSD).

typical CSD entry is illustrated schematically in Fig. 24.3.2.1. Individual data items can be categorized into three different groupings which are most conveniently described in terms of their dimensionality.

24.3.2.3. *1D bibliographic and chemical data*

The one-dimensional data for each entry comprise chemical and bibliographic text strings, together with certain individual numerical items, *viz* chemical compound name and any common synonym(s), chemical formula, authors names, journal name and literature citation, text comment reflecting any special experimental details (non-room-temperature study, absolute configuration determined, neutron study *etc.*). The cell parameters, crystal data, space group and precision indicators also fall into this category.

24.3.2.4. *2D chemical connectivity data*

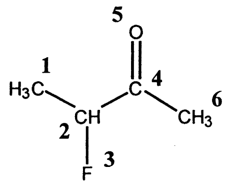
The formal two-dimensional chemical structural diagram for each entry (Fig. 24.3.2.2) is encoded in the form of a compact connection table. Chemical connectivity is recorded in terms of a set of atom and bond properties. The atom properties recorded are: atom number, element type, number of connected non-H atoms, number of terminal H atoms and the formal atomic charge. Bond properties are encoded as a pair of atom numbers and the formal chemical bond type that connects those atoms. Bond types employed in the CSD connectivity descriptions are: single, double, triple, quadruple (metal–metal), aromatic, delocalized double and π bonds. Bond types are (automatically) coded negative if the bond forms part of a cyclic system.

24.3.2.5. *3D crystal structure data*

The three-dimensional data consist of the fractional coordinates and symmetry operators for each entry. This information, together with the cell dimensions, is used to establish a crystallographic connectivity using standard covalent radii. The chemical and crystallographic connectivities are then mapped onto one another, using graph-theoretic algorithms, so that the chemical atom and bond properties are associated with the three-dimensional structure for search purposes. The CSD always records coordinates for complete molecules. Thus, if a molecule adopts a special position in the assigned space group, *i.e.* the asymmetric unit is some fraction of the total number of atoms in the molecule, then the CSD system also records those symmetry-generated atoms that complete the chemical entity. This speeds up the search process and also makes the data more accessible to non-crystallographers.

24.3.2.6. *Derived data and bit-encoded information*

Derived data are calculated directly from the evaluated raw data and stored in the master archive for search purposes. Numerical items such as Z', the number of chemical entities in the asymmetric unit, is a typical (real) numerical data item in this category. However, by far the most useful of the derived data items are a set of 682 individual pieces of yes/no information which are encoded as a bitmap, referred to as the *screen* record. The first 155 of these bits record information about (*a*) the elemental constitution of the compound, (*b*) results of the data-validation procedure and (*c*) summary information about the data content of the entry. These bits can be accessed directly by the user as search keys. The most important parts of the bitmap contain codified yes/no information about the presence/absence of specific features in the complete 2D or 3D structures held in the CSD. When a chemical substructure is entered as a query, its constitution is analysed in the same way to produce a bitmap for the query. Logical comparison of the query bitmap with the bitmap stored for each full CSD entry is computationally rapid, and quickly eliminates those entries that do not contain the requested features. Only those entries that pass this initial screening process need enter the detailed and computationally intensive atom-by-atom, bond-by-bond connectiv-



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

Fig. 24.3.2.2. 2D chemical connectivity data for a simple organic molecule.

ity mapping that finally confirms (or not) the presence of the required query substructure.

24.3.2.7. *Data validation*

All data entering the CSD are subject to stringent check and evaluation procedures. Some of these are visual, but the majority are automated within the CSD program *PreQuest*. The checks ensure that the 1D and 2D information fields abstracted by CCDC staff are accurately encoded, and that the 3D crystallographic coordinates are consistent with both the chemical description of the structure and with the geometrical description supplied by the authors. Most typographical errors in original papers can be corrected by the CCDC but, in the case of serious discrepancies, the original authors are consulted.

24.3.2.8. *The CSD-Use database*

CSD-Use is a database of scientific research papers in which the CSD was used as the principal or sole source of experimental information. The database comprises more than 700 literature citations classified according to the type of systematic study undertaken. Each CSD-Use entry also contains a short summary of the major findings of the research. The database is growing rapidly over time, and is expected to be a valuable resource in the future, since it contains a fully retrospective overview of the data-mining methods and research applications of the CSD.

24.3.3. The CSD software system

24.3.3.1. *Overview*

The CSD is supplied with a suite of fully interactive graphical software modules which provides users with facilities to: (*a*) interrogate all of the 1D, 2D and 3D information fields; (*b*) display entries graphically in a variety of styles; (*c*) retrieve relevant data for search hits, including geometrical parameters derived from the stored coordinates; and (*d*) display the derived numerical information, *e.g.* as histograms, scattergrams *etc.*, generate descriptive statistics and perform more complex numerical analyses. More recently, software has been added that permits users to transform their own in-house structural data to CSD formats for inclusion in these processes. A summary of the overall CSD software system is given in Fig. 24.3.3.1 which shows the functional relationships between the four major applications programs.

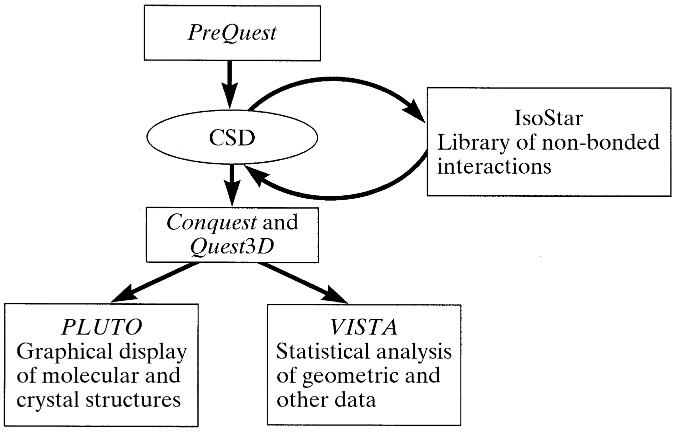

PreQuest
CSD
IsoStar
Library of non-bonded interactions
Conquest and *Quest*3*D*
PLUTO
Graphical display of molecular and crystal structures
VISTA
Statistical analysis of geometric and other data

Fig. 24.3.3.1. Summary of the software components of the Cambridge Structural Database system (CSDS).

24.3.3.2. *PreQuest*

PreQuest is a data-validation and data-conversion program which is used to create high-quality structural data files in CSD format from, *e.g.*, raw input data from a CIF. *PreQuest* is used routinely by CCDC's scientific editors to create and validate entries for inclusion in the master CSD archive, hence the program is constantly being maintained and upgraded. The released version enables users to build a private CSD-format database of their own structures which can then be searched independently of, or in conjunction with, the master CSD files using the database access programs described below.

24.3.3.3. *Searching the CSD: Quest3D and ConQuest*

*Quest*3*D* has been the main search engine and information-retrieval program for the CSD since the late 1980s. Its main features are summarized below. However, since 1997, the CCDC has been developing its successor, the *ConQuest* program, which was first released as part of the CSD system in April 2000. During an interim period, perhaps two years, *ConQuest* and *Quest*3*D* will both form part of the released CSD system on certain computing platforms while the functionality of the new program is being fully developed. Further details of *ConQuest* are provided in Section 24.3.3.5, indicating in particular how it differs from, and improves upon, the facilities available in *Quest*3*D*.

24.3.3.4. *Quest3D*

*Quest*3*D* is the main search engine and information-retrieval program for the CSD. It permits interrogation of all information fields: (*a*) 19 text fields, (*b*) 38 individual numerical fields, (*c*) element symbols and element counts, (*d*) full or partial molecular formulae, (*e*) direct access to over 150 bit screens, (*f*) extensive 2D chemical substructure search capabilities, and (*g*) 3D substructure searching at the molecular level or at the extended crystal-structure level. A search of a specific information field is termed a *test* of that field, and is constructed graphically *via* the menu system; menu components correspond to the categories of searches identified above. A complete *query* is then constructed by combining a number of separate *test* components using Boolean logic.

Substructure searching is the most important and frequently used facility. At the molecular level, the substructure (chemical fragment) query is entered graphically and is defined using the formal covalent bond types present in the 2D chemical connectivity tables of the CSD. The process can be extended to locate non-bonded contacts in the complete crystal structure. Here, the individual atoms or chemical groups involved in the contact must be specified, and a limiting non-bonded contact distance must be provided, along with any other geometrical criteria required to define the contact more precisely.

All substructure searches begin with the user drawing the required chemical unit(s) *via* the BUILD menu. Chemical variability and precision are controlled through (*a*) the PERIODIC TABLE sub-menu, which allows for specification of variable element types at specific atomic sites, (*b*) the 2D-CONSTRAIN menu, which allows further chemical restrictions to be specified, such as cyclicity/acyclicity of bonds, exact hydrogen-atom counts, total coordination numbers for atoms *etc.*, and (*c*) the 3D-CONSTRAIN menu, which permits the user to specify a list of geometrical parameters to be calculated by the program for each instance of the fragment located in the CSD; any of these geometrical parameters may be used as criteria to limit the scope of the search, especially at the intermolecular level. A file of calculated geometrical information is output by *Quest*3*D* and may be read by *Vista*, or by external data analysis software. Other

references