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 18000 structures were added to the database, and in mid-2000 the total archive contained over 220000 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 500000 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.

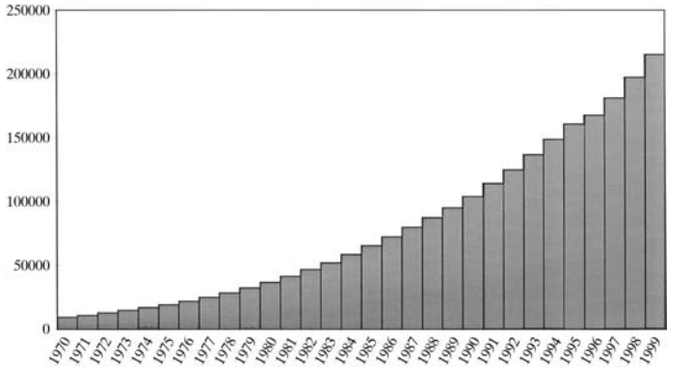

Fig. 24.3.1.1. Growth of the CSD since 1970 expressed in terms of the number of structures published per annum.

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	224945
No. of compounds	202669
No. of entries with 3D coordinates	198136
No. of entries with error-free coordinates	194784
No. of atoms having 3D coordinates in the CSD	12906283
No. of entries in the CSD-Use database	792

references