19. X–1 THE CURRENT STATUS OF CRYSTALLOGRAPHIC DATABASES. By Frank H. Allen, Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England.

Since the inception of Strukturbericht structural crystallography has had an enviable history of self documentation. The developmental dependence of the subject on advances in computer technology led to an early interest in machine-readable information sources. Early enough, just, to cope with the computer-controlled explosion of results produced in the past 10-15 years.

There are now six major databases covering the subject, with areas of interest which can be delineated on crystallographic or chemical criteria :

* Powder Diffraction File (PDF) : JCPDS International Center for Diffraction Data, Swarthmore, Pa., USA : Powder patterns, cell parameters etc.
* Crystal Data File (CDF) : National Bureau of Standards, Washington D.C., USA : Crystal data from powder or single-crystal studies.

These two sources are chemically comprehensive, but have limited structural information; the additional atomic coordinate data are available in four chemical divisions.

* Cambridge Structural Database (CSD) : CCDC, Univ. of Cambridge, UK : Organics, organometallics, complexes.
* Inorganic Crystal Structure Database (ICSD) : Univs. of Bonn, FRG / McMaster, Canada : Inorganics.
* Metal Data File (MDF) : National Research Council of Canada, Ottawa : metals, inter-metallics.
* Protein Data Bank (PDB) : Brookhaven National Lab., Upton, NY, USA : Macromolecules.

The majority of information is abstracted from primary literature sources and all databases carry a citation to the original, together with some chemical information : typically compound names and molecular formulae. These descriptors are sometimes extended, e.g. the sequence data of PDB, the chemical connectivity tables of CSD, or the assigned structures of MDF. The bibliographic and chemical material, together with some suitable numeric items form search terms for data access.

Each database includes material from journal or public depositories, and in some cases they act as direct depositories in their own right. The PDB has always operated in this way and in recent years CSD and ICSD have received unpublished coordinates from specific journals.

Each Centre has made a considerable investment in database building software to ensure, as far as possible, an accurate product. More than 15% of published reports contain at least one numeric error; a large proportion of these can be eradicated via data evaluation methods. Statistics will be presented which indicate that, taken together, the six databases represent the most comprehensive machine-readable numeric data resource available in any sub-branch of science.

Whilst database building techniques have conceptually similar aims, methods for their dissemination are less coherent. There are a number of hard-copy products which were either forerunners or spin-offs from the machine files, e.g. The PDF index cards, Crystal Data (CDF) and Molecular Structures and Dimensions (CSD). Each Centre, except PDB, now has software available for database interrogation, and each has its own policy for the distribution of machine files. This policy is often dictated by funding needs or by the requirements of the supporting agencies,and the present situation will be summarized. Differences in file structures and software will also be discussed, in terms of differences in the chemical nature of the data (ionic vs molecular vs macromolecular), and of fundamental differences in the use of the data (rapid identification or reference retrieval vs long term studies based on numeric data for large numbers of related compounds).

19. X–2 OBTAINING STRUCTURAL DATA FROM COMPUTER DATABASES. Jenny P. Glusker and Peter Murray-Rust, The Institute for Cancer Research, The Fox Chase Cancer Center, Philadelphia, Pennsylvania 19111, U.S.A.

The Cambridge Crystallographic Data File contains a wealth of information on small and medium-sized organic structures that can be used to survey for averaged coordinates of chemical groups and for the geometry of the interatomic interactions between such groups in different molecules. These will be illustrated by studies of the bonding in thioesters,[1] the conformations of citrates,[2] metal coordinating and hydrogen bonding potential of the C-F bond,[3] the hydrogen bonding in acridines and the surroundings of functional oxygen atoms such as ketones, ethers, and epoxides.[4] Methods of analysis and results will both be discussed in detail.

1. D.E. Zacharias, P. Murray-Rust, R.K. Preston and J.P. Glusker. Arch. Biochem. Biophys. 222, 22 (1983).

2. J.P. Glusker. Accounts of Chemical Research 13, 345 (1980).

3. P. Murray-Rust, W.C. Stallings, C. Monti, R.K. Preston and J.P. Glusker. J. Amer. Chem. Soc. 105, 3206 (1983).

4. P. Murray-Rust and J.P. Glusker. J. Amer. Chem. Soc. 106, in press (1984).

Work supported by NIH grant CA-10925 and American Cancer Society grant BC-242.

19. X–3 ACCESSING CRYSTALLOGRAPHIC DATABASES VIA PUBLIC NETWORKS. By Pella Machin, SERC Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England

The Crystallographic databases are extremely valuable scientific resources because of the wealth of data which they contain on 3-dimensional structure. Scientists from a range of disciplines require access to these databases for a variety of different applications, from structure comparisons to compound identification. Computer networking has the advantage of allowing multi-user access to an easily maintained central database. In addition to the network it is essential that flexible retrieval, analysis and display software be available. This software is often interactive and will therefore rely upon indexing or inverted file techniques.

The six major crystallographic databases (CSD, ICSD, MDF, PDB, PDF, CDIF) are described in a preceding paper.

For each of these databases there exists some form of central service, ranging from a minimum of magnetic tape distribution to, at the other extreme, comprehensive interactive search and retrieval software.

In a number of respects these databases are ideal for distributed network access and for CSD in particular a wide range of networked interactive systems are currently available. There is provision for searching based on chemical structure as well as text and numeric data retrieval and there are facilities for structure display and datafile transfer.

The availability and use of databases via networks is increasing rapidly and will certainly grow further with the advent of systems such as EURONET and with advances in computer technology, network protocols and file transfer techniques.