

1.1. GENESIS OF THE CRYSTALLOGRAPHIC INFORMATION FILE

scientific results. However, the sheer volume of diffraction data needed to repeat a crystallographic study precludes these from publication, and has led in the past to relatively *ad hoc* procedures for depositing supplementary data in local or centralized archives. Typically in the past, only the crystal and structure model parameters were published in the refereed paper and the underpinning diffraction information had to be archived elsewhere. Because the archived data were usually stored as paper in various unregulated formats, considerable information about the experiment and structure-refinement parameters was never retained. Moreover, the archiving of supplementary data *via* postal services was very slow and labour-intensive; equally, the recovery of deposited data was difficult, with information supplied as either a photocopy of the original deposition or an image taken from a microfiche.

Prior to 1970, when less than 9000 structures were deposited with the Cambridge Crystallographic Data Centre, data sets were still small enough to make these deposition and retrieval approaches feasible, albeit tedious. Even so, records show that very few archived data sets were ever retrieved for later use. The rationale of data storage changed radically in the 1980s. The increasing role of computers, automatic diffractometers and phase-solving direct methods in crystallographic studies led to a rapid acceleration in the number and size of structures determined and published. This was the period when fast minicomputers became affordable for laboratories, and the consequent demand for the electronic storage and exchange of information grew exponentially. Typical data-archival practices changed from using paper to magnetic tapes, as these now became the least expensive and most efficient means of storing data.

1.1.3. Card-image formats

Although the interchange of scientific information depends implicitly on an agreed data format, it remains independent of whether the transmission medium is paper tape, punched card, magnetic tape, computer chip or the Internet. Crystallography has employed countless data-exchange approaches and formats over the past 60 years. Prior to the advent of computers, the standard approach involved the exchange of typed tables of coordinates and structure factors with descriptive headers. In the 1950s and 1960s, as computers became the dominant generators of data, the transfer of data between laboratories was still relatively uncommon. When it was necessary, the Hollerith card formats of commonly used programs, such as *ORFLS* (Busing *et al.*, 1962) and *XRAY* (Stewart, 1963), usually sufficed. Even when magnetic tape drives became common and were standardized (mainly to the 1/2-inch 2400-foot reel), the 80-column 'card-image' formats of these programs remained the most popular data exchange and deposition approach.

As the storage and transporting of electronic data became easier and cheaper, structural information was increasingly deposited directly in databases such as the Cambridge Structural Database (CSD; Allen, 2002) and the Protein Data Bank (PDB; Bernstein *et al.*, 1977). The CSD and PDB simplified these depositions by using standard layouts such as the *ASER*, *BCCAB* and *PDB* formats. Both the PDB and CSD used, and indeed still use as a backup deposition mode, fixed formats with 80-character records and identifier codes. Examples of these format styles are shown in Figs. 1.1.3.1 and 1.1.3.2.

The card-image approach, involving a rigid preordained syntax, survived for more than two decades because it was simple, and the suite of data types used to describe crystal structures remained relatively static.

```

HEADER  PLANT SEED PROTEIN 30-APR-81 1CRN 1CRND 1
COMPND  CRAMBIN
SOURCE  ABYSSINIAN CABBAGE (CRAMBE ABYSSINICA) SEED
AUTHOR  W.A.HENDRICKSON,M.M.TEETER
REVDAT  5 16-APR-87 1CRND 1  HEADER 1CRND 2
REVDAT  4 04-MAR-85 1CRNC 1  REMARK 1CRNC 3
REVDAT  3 30-SEP-83 1CRNB 1  REVDAT 1CRNB 1
REVDAT  2 03-DEC-81 1CRNA 1  SHEET 1CRNB 2
REVDAT  1 28-JUL-81 1CRN 0 1CRNB 3
REMARK 1 1CRN 7
REMARK 1 REFERENCE 1 1CRNC 2
REMARK 1 AUTH M.M.TEETER 1CRNC 3
REMARK 1 TITL WATER STRUCTURE OF A HYDROPHOBIC PROTEIN AT ATOMIC 1CRNC 4
REMARK 1 TITL 2 RESOLUTION. PENTAGON RINGS OF WATER MOLECULES IN 1CRNC 5
REMARK 1 TITL 3 CRYSTALS OF CRAMBIN 1CRNC 6
REMARK 1 REF PROC.NAT.ACAD.SCI.USA V. 81 6014 1984 1CRNC 7
REMARK 1 REFPN ASTM PNAS6 US ISSN 0027-8424 040 1CRNC 8
REMARK 1 REFERENCE 2 1CRNC 9
REMARK 1 AUTH W.A.HENDRICKSON,M.M.TEETER 1CRN 9
REMARK 1 TITL STRUCTURE OF THE HYDROPHOBIC PROTEIN CRAMBIN 1CRN 10
REMARK 1 TITL 2 DETERMINED DIRECTLY FROM THE ANOMALOUS SCATTERING 1CRN 11
REMARK 1 TITL 3 OF SULPHUR 1CRN 12
REMARK 1 REF NATURE V. 290 107 1981 1CRN 13
REMARK 1 REFPN ASTM NATUAS UK ISSN 0028-0836 006 1CRN 14
REMARK 1 REFERENCE 3 1CRN 10
REMARK 1 AUTH M.M.TEETER,W.A.HENDRICKSON 1CRN 16
REMARK 1 TITL HIGHLY ORDERED CRYSTALS OF THE PLANT SEED PROTEIN 1CRN 17
REMARK 1 TITL 2 CRAMBIN 1CRN 18
REMARK 1 REF J.MOL.BIOL. V. 127 219 1979 1CRN 19
REMARK 1 REFPN ASTM JMOBOK UK ISSN 0022-2836 070 1CRN 20
SEQRES 1 46 THR THR CYS CYS PRO SER ILE VAL ALA ARG SER ASN PHE 1CRN 51
SEQRES 2 46 ASN VAL CYS ARG LEU PRO GLY THR PRO GLU ALA ILE CYS 1CRN 52
SEQRES 3 46 ALA THR TYR THR GLY CYS ILE ILE ILE PRO GLY ALA THR 1CRN 53
SEQRES 4 46 CYS PRO GLY ASP TYR ALA ASN 1CRN 54
HELIX 1 H1 ILE 7 PRO 19 1 3/10 CONFORMATION RES 17,19 1CRN 55
HELIX 2 H2 GLU 23 THR 30 1 DISTORTED 3/10 AT RES 30 1CRN 56
SHEET 1 S1 2 THR 1 CYS 4 0 1CRNA 4
SHEET 2 S1 2 CYS 32 ILE 35 -1 1CRN 58
TURN 1 T1 PRO 41 TYR 44 1CRN 59
SSBOND 1 CYS 3 CYS 40 1CRN 60
SSBOND 2 CYS 4 CYS 32 1CRN 61
SSBOND 3 CYS 16 CYS 26 1CRN 62
CRYST1 40.960 18.650 22.520 90.00 90.77 90.00 P 21 2 1CRN 63
ATOM 1 N THR 1 17.047 14.099 3.625 1.00 13.79 1CRN 70
ATOM 2 CA THR 1 16.967 12.784 4.338 1.00 10.80 1CRN 71
ATOM 3 C THR 1 15.685 12.755 5.133 1.00 9.19 1CRN 72
ATOM 4 O THR 1 15.268 13.825 5.594 1.00 9.85 1CRN 73
ATOM 5 CB THR 1 18.170 12.703 5.337 1.00 13.02 1CRN 74
ATOM 6 OG1 THR 1 19.334 12.829 4.463 1.00 15.06 1CRN 75
ATOM 7 CG2 THR 1 18.150 11.546 6.304 1.00 14.23 1CRN 76
ATOM 8 N THR 2 15.115 11.555 5.265 1.00 7.81 1CRN 77
ATOM 9 CA THR 2 13.856 11.469 6.066 1.00 8.31 1CRN 78
ATOM 10 C THR 2 14.164 10.785 7.379 1.00 5.80 1CRN 79
CONNECT 20 19 282 1CRN 398
CONNECT 26 25 229 1CRN 399
CONNECT 116 115 188 1CRN 400
CONNECT 188 116 187 1CRN 401
CONNECT 229 26 228 1CRN 402
CONNECT 282 20 281 1CRN 403
END 1CRN 405

```

Fig. 1.1.3.1. An abbreviated example of a PDB format file.

1.1.4. The Standard Crystallographic File Structure (SCFS)

By the 1980s, the many different fixed formats used to exchange data electronically had become a significant complication for journals and databases. Because of this, the IUCr Commissions for Crystallographic Data and Computing formed a joint Working Party which was asked to recommend a standard format for the exchange and retention of crystallographic data. They proposed a partially fixed format in which key words on each line identified blocks of data containing items in a specific order. This format was the Standard Crystallographic File Structure (Brown, 1988). An example of an SCFS file is shown in Fig. 1.1.4.1.

The effectiveness of the SCFS format approach was curtailed because its release coincided with the arrival of powerful minicomputers, such as the VAX780, in crystallographic laboratories. This led to a period of enormous change in crystallographic computing, in which new data types and file formats proliferated. It was also a time when automatic diffractometers became standard equipment in laboratories and the development of new crystallographic software packages flourished. The fixed-format design of the SCFS was unable to adapt easily to these continually changing data requirements, and this eventually led to a proliferation of SCFS versions.