

## 1.1. GENESIS OF THE CRYSTALLOGRAPHIC INFORMATION FILE

By the early 1990s, TCP/IP and the Internet dominated international networking. The practices of open exchange of information were developed through a number of initiatives. *Gopher* (Anklesaria *et al.*, 1993) provided a general mechanism to access material categorized and published from a computerized information store. *WAIS* (Kahle, 1991), a wide-area information server application designed to service queries conforming to the Z39.50 information retrieval protocol (ANSI/NISO, 1995), provided an effective distributed search engine. The rapid proliferation of new techniques for searching and retrieving information from the Internet was capped in the mid-1990s by the rapid growth in sites implementing hypertext servers (Berners-Lee, 1989). The World Wide Web had become a reality.

The increasing access to global network facilities during the 1980s led to a growing interest among crystallographers in submitting manuscripts to journals electronically, especially for small-molecule structure studies. The Australian delegation at the 1987 General Assembly of the XIVth IUCr Congress in Perth proposed that IUCr journals (specifically *Acta Crystallographica*) should be able to accept manuscripts submitted electronically. It was argued that this would reduce effort on the part of the authors and the journal office in preparation and transcription of manuscripts, and as a consequence reduce costs and transcription errors and simplify data-validation approaches. The acceptance of this General Assembly resolution led to the creation of a Working Party on Crystallographic Information (WPCI), which had as its mandate the investigation of possible approaches to enable the electronic submission of crystallographic research publications.

### 1.1.6. The Working Party on Crystallographic Information (WPCI)

The WPCI first convened at the 1988 ECM11 conference in Vienna. In the discussions leading up to this meeting, it was widely appreciated that electronic submissions to journals and databases involved data types (*e.g.* manuscript texts, graphical diagrams, the full suite of crystallographic data) that were beyond those accommodated within the SCFS format promoted by the IUCr Data and Computing Commissions. Consequently, it was suggested at the Vienna meeting that a general and extensible universal file approach, similar to the recently developed Self-defining Text Archive and Retrieval (STAR) File format (Hall, 1991; Hall & Spadaccini, 1994), might also be suitable for crystallographic data applications.

At this meeting, it was decided that a WPCI working group, led by Syd Hall, should investigate the development of a universal file protocol that would be suitable for crystallographic data needs. Other universal formats existed, such as ASN.1 (ISO, 2002), which was used for data communications, JCAMP-DX (McDonald & Wilks, 1988), which was used for archiving infrared spectra, and the Standard Molecular Data (SMD) format (Barnard, 1990), which was used for the global exchange of chemical structure data. These were considered relatively inefficient for expressing the repetitive data lists commonly used in crystallography. The working group eventually proposed a Crystallographic Information File (CIF) format which had a syntax similar to, but simpler than, the STAR File. Of particular importance because of the rapid changes taking place with data types, the CIF approach provided a very flexible and extensible file structure in which any type of text or numerical data could be arranged in any order. The typical data structure of a CIF is illustrated in Fig. 1.1.6.1, using the same data as presented in the PDB file of Fig. 1.1.3.1. Similarly, Fig. 1.1.6.2 shows the data in the BCCAB file of Fig. 1.1.3.2 in CIF format.

```

data_crmbin
_entry.id                                1CRN

_audit.creation_date                      1993-04-21
_audit.creation_method                    'manual editing of PDB entry'
_audit.update_record

; 1993-04-21 Original PDB entry history recorded here for completeness.
 30-apr-81 deposition.
 28-jul-81 lcrn 0
 03-dec-81 correct residue number on strand 1 of sheet s1.
 30-sep-83 insert revdat records
 04-mar-85 insert new publication as reference and renumber
 16-apr-87 change deposition date from 31-apr-81 to 30-apr-81.
;
loop_
_struct.entry_id
_struct.title
 1CRN 'Crmbin from Abyssinian cabbage (Crambe abyssinica) seed'
loop_
_citation.id
_citation.year
_citation.journal_abbrev
_citation.journal_volume
_citation.page_first
_citation_journal_id_ASTM
_citation_journal_id_ISSN
_citation_title
primary 1984 Biochemistry 23 6796
? 0006-2960
; Raman spectroscopy of homologous plant toxins: crmbin and alpha 1- and
beta-purothionin secondary structures, disulfide conformation, and
tyrosine environment
;
1 1984 Proc.Nat.Acad.Sci.USA 81 6014
pnasa6 0027-8424
; Water structure of a hydrophobic protein at atomic resolution. Pentagon
rings of water molecules in crystals of crmbin
;
2 1981 Nature 280 107
natuas 0028-0836
; Structure of the hydrophobic protein crmbin determined directly
from the anomalous scattering of sulphur
;
loop_
_citation_author.citation_id
_citation_author.name
primary 'Williams, R.W.' primary 'Teeter, M.M.'
1 'Teeter, M.M.'
2 'Hendrickson, W.A.' 2 'Teeter, M.M.'
loop_
_entity.id
_entity.type
_entity.details
1 polymer 'Protein chain: *'
2 non-polymer 'het group EOH'
loop_
_entity_poly_seq.entity_id
_entity_poly_seq.num
_entity_poly_seq.mon_id
1 1 THR 1 2 THR 1 3 CYS 1 4 CYS 1 5 PRO
1 6 SER 1 7 ILE 1 8 VAL 1 9 ALA 1 10 ARG
1 11 SER 1 12 ASN 1 13 PHE 1 14 ASN 1 15 VAL
#.....sequence data omitted for brevity

_cell.length_a 40.960
_cell.length_b 18.650
_cell.length_c 22.520
_cell.angle_alpha 90.00
_cell.angle_beta 90.77
_cell.angle_gamma 90.00
_symmetry.space_group_name_H-M 'P 1 21 1'

loop_
_atom_type.symbol
_atom_type.description
_atom_type.number_in_cell
C carbon 404 N nitrogen 112 O oxygen 128 S sulfur 12 H hydrogen 7

loop_
_atom_site.label_seq_id
_atom_site.type_symbol
_atom_site.label_atom_id
_atom_site.label_comp_id
_atom_site.label_asym_id
_atom_site.auth_seq_id
_atom_site.label_alt_id
_atom_site.Cartn_x
_atom_site.Cartn_y
_atom_site.Cartn_z
_atom_site.occupancy
_atom_site.B_iso_or_equiv
_atom_site.label_entity_id
_atom_site.id
1 N N THR * 1 . 17.047 14.099 3.625 1.00 13.79 1 1
1 C CA THR * 1 . 16.967 12.784 4.338 1.00 10.80 1 2
1 C C THR * 1 . 15.685 12.755 5.133 1.00 9.19 1 3
1 O O THR * 1 . 15.268 13.825 5.594 1.00 9.85 1 4
1 C CB THR * 1 . 18.170 12.703 5.337 1.00 13.02 1 5
1 O OG1 THR * 1 . 19.334 12.829 4.463 1.00 15.06 1 6
1 C CG2 THR * 1 . 18.150 11.546 6.304 1.00 14.23 1 7
#.....atom-site data omitted for brevity

```

Fig. 1.1.6.1. Example 1 of a CIF (using the same data as shown in Fig. 1.1.3.1).