International Tables for Crystallography (2006). Vol. G, Section 1.1.2, pp. 2–3.

1.1. Genesis of the Crystallographic Information File

By S. R. Hall and B. McMahon

1.1.1. Prologue

Progress in science depends crucially on the ability to find and share theories, observations and the results of experiments. The efficient exchange of information within and across scientific disciplines is therefore of fundamental importance. The rapid growth in the use of computers and networks over the past half century and in the use of the World Wide Web over the past decade have brought remarkable improvements in communications among scientists. Such communications are most effective when there is a common language. The English language has become the standard means of expression of ideas and theories; increasingly, the exchange of data requires the computer equivalent of such a *lingua franca*. It was with considerable foresight that the International Union of Crystallography (IUCr) in 1990 adopted a data-handling approach based on universal file concepts. At that time this was considered to be a radical idea. The approach adopted by the IUCr is known as the Crystallographic Information File (CIF). This volume of *International Tables for Crystallography* describes the CIF approach, the associated definition of CIF data items within dictionaries, and handling procedures, applications and software.

In this opening chapter, we give a historical perspective on the reasons why the CIF approach was adopted and how, over the past decade, CIF applications have evolved.

CIF is the most fully developed and mature of the various universal file approaches available today. It combines flexibility and simplicity of expression with a lean syntax. It has an unsurpassed ability to express 'hard' scientific data unambiguously using extensive dictionaries (ontologies) of relevant terms. It has proved to be remarkably well suited to the publication and archiving of small-unit-cell crystallographic structures. What was a radical idea in 1990 has today become the dominant mode of expression of scientific data in this domain.

The CIF data model provided the key to the internal restructuring of data managed by the Protein Data Bank in its transition from an archive to a database. The CIF approach is being tested in an increasing number of domains. In some cases, it may well become as successful as it has been for small-molecule crystallography. In other cases, the syntax will be unsuitable, but yet the conceptual discipline of agreed ontologies will still be required. Here, the experience of developing the CIF dictionaries may be carried across into different file formats and modes of expression.

Nowadays, informatics is a rapidly evolving field, in which everything is obsolete almost as soon as it is created. Yet there is a responsibility on today's scientists to preserve data and pass them on to the next generation. CIF was developed not only as a data-exchange mechanism, but also as an archival format, and considerable care has been taken over the past decade and more to keep it a stable and smoothly evolving approach. Some points of detail have been modified or superseded in practice. Other changes will necessarily occur as the approach evolves to meet the changing demands of an evolving science. Readers should therefore be aware of the need to consult the IUCr website (http://www.iucr.org/iucr-top/cif) for the latest versions of, or successors to, the data dictionaries and the software packages described in this volume. However, the basic concepts have already been shown to be remarkably effective and durable. This volume should therefore provide an invaluable reference for those working with CIF and related universal file approaches.

The success of any data-exchange approach depends on its efficiency and flexibility. It must cope with the increasing volume and complexity of data generated by the computing 'information explosion'. This growth challenges conventional criteria for measuring exchange and storage efficiency based on high data-compression factors. Today's fast, cheap magnetic and chip technologies make bulk volume a secondary consideration compared with extensibility and portability of data-management processes. Most importantly, improvements in computing technology continue to generate new approaches to harnessing semantic information contained within data collections and to promoting new strategies for knowledge management.

The basis for an efficient information-exchange process is mutually agreed rules for the supplier and the receiver, *i.e.* the establishment of an exchange protocol. This protocol needs to be established at several levels. At the first level, there must be predetermined ways that data (*i.e.* numbers, characters or text) are arranged in the storage medium. These are the organizational rules that define the syntax or the format of data. There must also be a clear understanding of the meaning of individual data items so that they can be correctly identified, accessed and reused by others. At an even higher level, a protocol may also provide rules for expressing the relationships between the data, as this can lead to automatic processes for validating and applying the data values.

These higher levels provide the *semantic* knowledge needed for the rigorous identification and validation of transmitted and stored data. One may consider the analogy of reading this paragraph in English. To do this we must first be able to recognize the individual words, comprehend their meaning (if necessary with the use of an English dictionary) and understand all of this within the context of sentence construction. As with data, the arrangement of component words is based on a predetermined grammatical syntax, and their individual meaning (as defined in a dictionary or elsewhere), coupled with their contextual function (as nouns, verbs, adjectives *etc.*), leads to the full comprehension of a word sequence as semantic information.

1.1.2. Past approaches to data exchange

The crystallographic community, along with many other scientific disciplines, has long adhered to the philosophy that experimental data and results should be routinely archived to facilitate long-term knowledge retention and access. An early approach to this, recommended by IUCr and other journals, was for authors to deposit data as hard copy (*i.e.* ink on paper) with the British Library Lending Division. Retaining good records is fundamental to reproducing

Affiliations: Sydney R. Hall, School of Biomedical and Chemical Sciences, University of Western Australia, Crawley, Perth, WA 6009, Australia; Brian McMahon, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

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.

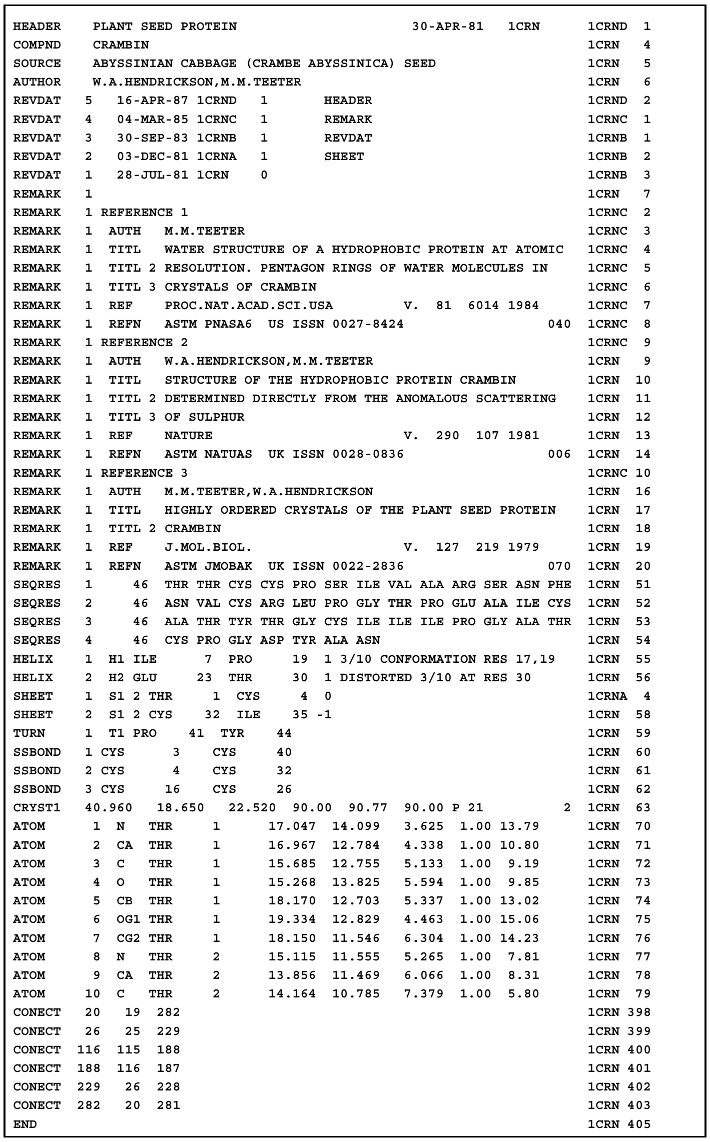
Prior to 1970, when less than 9000 structures were deposited with the Cambridge Crystallographic Data Centre, data sets were still small enough to make these deposition and retrieval approaches feasible, albeit tedious. Even so, records show that very few archived data sets were ever retrieved for later 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                                                       1CRN   4
SOURCE    ABYSSINIAN CABBAGE (CRAMBE ABYSSINICA) SEED                   1CRN   5
AUTHOR    W.A.HENDRICKSON,M.M.TEETER                                    1CRN   6
REVDAT   5   16-APR-87 1CRND   1       HEADER                           1CRND  2
REVDAT   4   04-MAR-85 1CRNC   1       REMARK                           1CRNC  1
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  REFN   ASTM PNASA6  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  REFN   ASTM NATUAS  UK ISSN 0028-0836                  006  1CRN  14
REMARK   1 REFERENCE 3                                                  1CRNC 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  REFN   ASTM JMOBAK  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
CONECT   20   19  282                                                   1CRN 398
CONECT   26   25  229                                                   1CRN 399
CONECT  116  115  188                                                   1CRN 400
CONECT  188  116  187                                                   1CRN 401
CONECT  229   26  228                                                   1CRN 402
CONECT  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.