THE CRYSTALLOGRAPHIC INFORMATION FILE (CIF)

***I.D.Brown*[1*] *and B.McMahon*[2]**

[*1] *Brockhouse Institute for Materials Research, McMaster University, Hamilton, Ontario, Canada*
Email: idbrown@mcmaster.ca
[2] *International Union of Crystallography, 5 Abbey Square, Chester, UK*
Email: bm@iucr.org

ABSTRACT

The Crystallographic Information File (CIF), owned by the International Union of Crystallography, is a file structure based on tag–value ASCII pairs with tags defined in machine-readable dictionaries. The crystallographic community publishes and archives large quantities of numeric information generated by crystal structure determinations, and CIF's acceptance was assured by its adoption as the submission format for Acta Crystallographica and by the obvious needs of the community. CIF's strength lies in its dictionaries, which define most of the concepts of crystallography; its weakness is the difficulty of writing software that exploits its full potential.

Keywords: Crystallography, File structure, Crystallographic Information File, Machine readable dictionaries, Archiving

1 OVERVIEW

Crystallographers deal with large quantities of numerical information, ranging from the intensities and positions of the thousands of Bragg reflections they measure in the diffraction patterns of crystals, to the lists of coordinates and displacement parameters of the atoms that give rise to these patterns. It is no surprise that crystallographers were among the first to use computers in the 1950s. By the 1970s a number of suites of crystal-structure-solving programs became available, each defining its own input and output file structures. At this time a proposal was put forward for a universal file structure that would facilitate the transfer of numerical information between programs and allow it to be robustly archived. This proposal (Brown, 1988) used a format based on card images. It was adopted by a small number of programmers, but its use did not become widespread. By the 1990s there was a growing need for a more broadly based structure that could include free text as well as the numeric values. The need was particularly acute for the publishing and archiving of crystal structure determinations. At that time, the computer that was used to refine a crystal structure printed tables of atomic coordinates that were then manually typed into the manuscript. After acceptance of the paper by a structural science journal such as *Acta Crystallographica*, these coordinates were again keyboarded to produce the typeset journal. Once in print the coordinates were keyboarded a third time for entry into a crystal structure database. It was estimated at the time that 10% of all papers reporting crystal structure determinations contained numerical errors that effectively made the numerical results unusable. In an attempt to locate these transcription errors, some editors of *Acta Crystallographica* even keyboarded the atomic coordinates into their own programs to check that they were consistent with the reported bond lengths.

In 1990 the International Union of Crystallography (IUCr) approved the introduction of on-line submission and computer typesetting of papers for its *Acta Crystallographica* family of journals, starting with the standard reports of structure determinations that were published in *Section C*. For this purpose it chose the Crystallographic Information File (CIF) in which each datum is preceded by a data name that serves to identify it. The recognized data names are defined in dictionaries, the first of which was published by Hall, Allen & Brown in 1991. In 1992 the editorial office of *Acta Crystallographica* started accepting reports of crystal structures in CIF, and by 1994 all papers in *Section C* were being typeset directly from the submitted CIFs. This required, as a temporary measure, that papers submitted as hard copy be keyboarded into CIF in the editorial office, but by 1996, only four years after the introduction of CIF, the most common programs could generate such files, and the journal then only accepted electronically submitted papers.

The procedures introduced at that time still form the basis for submission of structure reports to *Acta Crystallographica*. Before formally submitting a paper *via* the web, authors are required to precheck it using the program system *checkcif* available on the IUCr's web site. This checks the consistency and plausibility of the data values reported as well as the syntax of the CIF, returning a list of issues that the author must address before formal submission. In effect, it performs automatically the extensive and laborious checks that had previously been undertaken by particularly conscientious referees. This process has had two benefits: the authors learn quickly from the problems found by *checkcif* and the quality of submitted structure determinations rarely needs to be questioned by the editor. Once the paper is published, the numerical information is made available on the web for downloading in CIF format for use as input into a user application, and the CIF is also passed on to the appropriate crystal structure database. A number of journals besides those belonging to the IUCr now require that any crystal structure report that accompanies a submitted paper must first be run through *checkcif* and be deposited with the appropriate crystal structure database before the paper is sent for refereeing.

2 A DESCRIPTION OF THE STANDARD

CIF is a flexible self-defining file based on the STAR File structure (Hall, 1991). To ensure that a CIF archive remains readable over time, it uses only ASCII characters. Each datum consists of two components: a data name followed by its data value. The data names are defined, along with their attributes (*e.g.*, whether the value is a number or text), in dictionaries that are themselves written as STAR Files. Since CIFs and their dictionaries have the same syntax, the same program can read both; thus the appropriate dictionary can be read in by the program to validate and interpret the CIF. The structure of CIF has similarities to XML, which was developed later, but it differs from XML in important respects. The file structure is simpler so that a raw CIF can easily be read visually and edited using a word processor, a feature that was a major factor in its early acceptance since no CIF editors were available during the first ten years. A second respect in which CIF differs from XML is in the existence of comprehensive dictionaries that are accepted by the whole discipline, a situation that works well in a mature field like crystallography. The existence of CIF dictionaries does not prevent CIFs from containing user-defined data names, and where such data names are found to serve a wider need; they can be moved into the official dictionaries. An important feature of CIF is its ownership by the IUCr who actively discourage the formation of private dialects of CIF. The IUCr has delegated the management of CIF to the Committee for the Maintenance of the CIF Standard (COMCIFS) which approves the dictionaries and establishes the rules under which CIF operates. A full description of the standard appears in *International Tables for Crystallography* Volume G (Hall & McMahon, 2005).

The first CIF dictionary (Hall *et al.*, 1991) defined the items needed to describe the structure of a crystal with a small unit cell. A subsequent major dictionary extended this to crystals containing biological macromolecules. Dictionaries have also been developed to describe powder diffraction, modulated structures, electron densities, symmetry, and images. In addition a number of private dictionaries have been developed, some of which are registered with the IUCr though not formally adopted as CIF dictionaries. All the official dictionaries can be downloaded from the IUCr web site http://www.iucr.org/ in HTML and PDF formats as well as in the definitive machine-readable ASCII format. They can be found on the home page of the CIF project of the *International Union of Crystallography* (n.d.).

3 DEVELOPMENT OF THE STANDARD

There are a number of reasons why CIF was rapidly accepted by the crystallographic community. The long-standing use of computers in crystallography meant that crystallographers were computer literate and most had experienced the frustrations of converting their files from one format to another or of keyboarding tables of atomic coordinates from printed journals. By the late 1980s there were a relatively large number of program systems that could process the same initial diffraction measurements, and the narrow focus of the discipline meant there was already a broad consensus on the quantities that the programs needed to report, but each archived this information in a different and incompatible format.

There were important social factors also. Crystallographers form a tightly knit community that readily accepts the leadership of the IUCr on matters of nomenclature and definition, particularly as promulgated in *International Tables for Crystallography*. The journals of the IUCr enjoy a high reputation within the community, and the adoption of CIF

by these journals was undoubtedly the single most important step in gaining the acceptance of the community. There was also a sense that the increasingly routine nature of many structure determinations required a more streamlined route to the publication of results and their deposition in structural databases. Since the introduction of CIF clearly addressed a number of outstanding and growing problems, there was a strong incentive, particularly among those responsible for the crystallographic software packages, to adopt the new standard.

There were, however, some obstacles to its universal adoption. The free format in which data items were presented challenged program systems designed around fixed-format input/output. In particular, Fortran, which was still the language of choice for small-molecule crystallographic programs in the early 1990s, is not well suited to the parsing and string manipulation operations needed to read CIF data. Software libraries were made available to help programmers to make the transition, and this was certainly important in winning over many software authors. Nevertheless, others were slow to make the necessary changes, perhaps because of the high overhead of reading a very general data description language in applications where only a very specific subset of the available information (*e.g.* lists of atomic positional coordinates) might be required.

Another difficulty arose from the IUCr's efforts to establish the CIF standard as the intellectual property of the crystallographic community by applying for a software patent on the underlying file format. This was done with the best of intentions: to prevent outsiders from developing incompatible dialects of the format that would undermine its reliability as a universal standard. However, the attitude of programmers towards legalistic protection measures was changing at this time, perhaps encouraged by the attempts of software companies to levy royalties on the use of patented methods in widespread graphics file formats, and certainly helped by the dramatic increase in open-source projects. There was evidence of some suspicion about the IUCr's motives, particularly from the biological macromolecular crystallography community, where there was an established tradition of open-source software development and of government-sponsored open access to major crystal structure databases. Latterly the IUCr has released CIF software that it sponsors directly under an open-source license to help promote the understanding of CIF as a truly open standard. Its complete documentation in *International Tables for Crystallography, Vol G* (Hall & McMahon, 2005) should also encourage that understanding.

The IUCr retains copyright on the dictionaries of standard data names; this has proved effective in maintaining its authority over the standard. However, protocols have also been developed to allow different groups to make specific extensions, when required for special purposes, in an orderly way that does not conflict with the public data definitions.

Part of the success of CIF can undoubtedly be attributed to characteristics of the crystallographic community that are not necessarily shared by other disciplines – a close community, a mature discipline, large amounts of numerical information, and a scientific union accustomed to exercising leadership. But there are some features that could be applicable to other situations. The simple structure that allowed files to be edited on a word processor and, most importantly, the adoption of CIF for the submission of papers to the leading journals in the field and its commitment to serve the community by making this information available on the web in machine-readable form are policies that could be adopted by other groups. The one serious weakness in our project remains the relatively small number of software tools. In this respect XML, which has its origins in the computer science community, is better served than CIF. But CIF, with its origins in the crystallographic community, is further ahead in the development of dictionaries and of scientific applications that make direct use of the crystallographic data in all its richness. The construction of dictionaries has been one of COMCIFS major occupations. COMCIFS has discouraged the development of incompatible CIF-like dictionaries and file structures by offering to assist any interested group in producing a fully CIF compliant dictionary. COMCIFS has also encouraged the writers of widely used crystallographic software to make CIF input and output available, but in the absence of a wide range of tools for manipulating CIFs, most of this software has failed to take full advantage of CIF's flexibility.

4 FUTURE DIRECTIONS

CIF is now a widely accepted standard among crystallographers, though in some areas there is a need to exchange data with other disciplines and for these cases CIF to XML conversions are available. Most of the widely used crystallographic program suites are capable of producing CIF output, and an increasing number can also read CIF. The discipline is now well provided with dictionaries whose construction by specialists in the field has been both

challenging and rewarding, since it has required the organization of crystallographic concepts into a coherent system. However, CIF is less well supplied with software, particularly software that makes use of CIF's advanced features though there are signs that the new generation of programmers is interested in accepting this challenge. One such feature of CIF is its ability to read a file using the dictionary that was used to create it. One virtue of this feature is that it allows dictionaries to evolve to meet the changing demands of both crystallography and information technology without the need to write new software. Most of the popular crystallographic program suites have the original set of data names written right into their source code, which requires a new version of the program every time a new version of the dictionary appears. Only in the last year have a couple of browser-editors appeared that provide the user with the opportunity to create or edit a CIF using any desired CIF dictionary. No software is yet available that exploits the linkages between the tables that CIF provides, and even further from general acceptance is the provision in advanced versions of CIF dictionaries that allows an item to be calculated on the fly from other items in the CIF so that information can be retrieved that is only implicitly present in the CIF.

COMCIFS has found that it needs a large reach. It has to moderate the excitement of those on the cutting edge of information technology who are impatient to see new ideas incorporated into CIF, while at the same time stimulate those still trapped in flat-file land who are frustrated by what appears to be the unnecessary complexity of CIF. Change occurs too slowly for some and much too fast for others.

5 ACKNOWLEDGEMENTS

We are grateful to Syd Hall and Frank Allen for their clarity of vision, insight, and energy in the development of the CIF standard, to our colleagues in the crystallographic community who have developed software and assisted with data definitions over many years, and to the IUCr for its support of the CIF project over almost two decades.

6 REFERENCES

Brown, I. D. (1988) Standard Crystallographic File Structure-87. *Acta Cryst.* A44, 232.

Hall, S.R. (1991) The STAR file: a new format for electronic data transfer and archiving. *J. Chem. Inf. Comput. Sci.* 31, 326–333.

Hall, S.R., Allen, F.H., & Brown, I.D. (1991) The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Cryst.* A47, 655–685.

Hall, S.R. & McMahon, B. (Eds.) (2005) *International Tables for Crystallography* Vol. G, *Definition and Exchange of Crystallographic Data*, Berlin: Springer.

International Union of Crystallography (n.d.) Home page of the CIF project of the International Union of Crystallography. Available from: (http://www.iucr.org/iucr-top/cif). Retrieved from the World Wide Web, September 20, 2006.