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.

Copyright © 2006 International Union of Crystallography

references