3.1. General considerations when defining a CIF data item

BY B. MCMAHON

3.1.1. Introduction

Much of the power and usefulness of the Crystallographic Information File (CIF) arises from the existence of a comprehensive set of data dictionaries that define all data items commonly used in the field. These are the dictionaries that are presented in Part 4 of this volume. The information contained in a CIF is expressed in terms of these data items. A data item consists of a value associated with a data name, or tag. The tag may appear immediately before a single data value or in the heading of a looped list where the values form a column. In either construction, the data value is identified by the tag and this unique character string is the key to the definition of the data value in the dictionary.

A data definition may include information such as a text description of the quantity, its physical units, the range within which valid values must lie, the names of other data items that are related by inheritance or derivation to the data item and so on. Placing this information in a dictionary file, rather than in the data file itself, has a number of important advantages. First, it encourages the standardization of unique tags for data items, which is an essential step towards the seamless and unambiguous exchange of information. Dictionaries also facilitate a globally accepted understanding of what each data item is, and thus ensure that different data files using the same tags have a consistent interpretation.

The existence of global dictionaries does not in any way restrict the expressive power of CIF. A CIF may contain items not in the standard dictionaries, as well as items in local dictionaries with quite idiosyncratic definitions. The choice of which items to include in a CIF depends on the capabilities of the applications that are intended to use the data in the file. It is also influenced by the extent to which the author of the file wishes the data to be retrievable without ambiguity in the future. Of course, the same applies to data in XML (Bray *et al.*, 1998; W3C, 2004) or other data languages. In the adoption and application of CIF as a specific exchange mechanism, the crystallographic community has imposed on itself a particular discipline: the strict definition of its data with carefully maintained dictionaries. This is not to be seen as a restriction but as a means to unambiguous and effective communication.

As mentioned above, data with local definitions are easily accommodated in a CIF. However, for a CIF to be an effective exchange medium, data definitions need to be accessible to the community of users. This is most efficient when commonly used data items are collected into a dictionary or dictionaries that are readily obtainable and centrally coordinated. This is why the CIF dictionaries, containing the definitions of standard data names and their attributes, are published and maintained by a technical committee of the International Union of Crystallography (IUCr): the Committee for the Maintenance of the CIF Standard (COMCIFS). The dictionaries employ a dictionary definition language or DDL (see Chapters 2.5 and 2.6) to describe relevant attributes of CIF data items.

Affiliation: BRIAN MCMAHON, International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, England.

This chapter will discuss the general concepts behind defining data items in CIF dictionaries. It will describe how standard dictionaries may be constructed and disseminated, and also how local extensions may be built and used in ways that do not conflict with the need for community standards. Some necessary details about the administration of standard dictionaries are also provided.

3.1.1.1. Authorship of data dictionaries

A difficulty in developing a standard for information exchange across the field of crystallography is the breadth of the subject area and the many subdisciplines it includes. One feature of the construction of data dictionaries for CIF is the delegation of responsibility for identifying and defining the data items important within a research area to experts in that field. In consequence, a richer compilation of definitions results than would be possible from a single author or small group of authors. However, each subdiscipline will have its own emphases and requirements, and it becomes a challenge to accommodate the needs of each individual subdiscipline within the framework of the general body of definitions covering the entire subject area. COMCIFS deals with this challenge by initiating and ratifying dictionaries written by IUCr Commissions or other specialist groups.

3.1.1.2. Certification for community use

A further responsibility of COMCIFS is to try to harmonize the treatment of similar data requirements in different dictionaries and to maintain maximum compatibility between data files originating from different subdisciplines. To achieve this, COMCIFS can officially approve dictionaries submitted to and reviewed by it. It is these 'official' dictionaries that are included in this volume. Provisional dictionaries may also be issued and used within the relevant community before formal approval is given.

3.1.1.3. DDL versions

Ideally, compatibility between the data dictionaries originating from specific subdisciplines would be ensured by the adoption of the same attribute sets for data items. However, at this point in the evolution of the CIF standard, two slightly different attribute sets have become established. These are expressed in two versions of the dictionary definition language, DDL1 and DDL2 (detailed in Chapters 2.5 and 2.6, respectively). The differences arise because some subdisciplines benefit from a strict data model that is not appropriate in other areas. The core data items in crystallography must of course be accessible across the field, and so there are two formulations of the dictionary of core items, one in each DDL version. The existence of two formulations can make full information interchange across all areas of crystallography difficult, so work is under way to bring about a convergence of the two current representations (Hall *et al.*, 2002). It is particularly important for future interchange between crystallography and other related disciplines that a full understanding be reached of the best way to include different data structure models within a common interchange format.

references