International Tables for Crystallography (2006). Vol. G, Section 3.1.1.3, pp. 73–74.

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.

In this chapter, there will be some discussion of the differences in practice between the DDL versions DDL1 and DDL2, as these will strongly influence the choice of formalism for a dictionary relevant to a subdiscipline not yet represented.

3.1.2. Informal definition procedures

Before considering the techniques for defining data items in standard globally adopted dictionaries, it is important to discuss the techniques for including information that is only of local interest in a way that does not conflict with public data names.

An author of a CIF is free to include data names for local use (*i.e.* names not intended for common use across the community). However, such local data names *must not* conflict with those defined in public dictionaries, since the data name alone identifies the meaning that one must attach to an associated data value. Some protocols and conventions exist to prevent conflict in data names when the local data name is invented or subsequently, when later public dictionaries are released.

An author may also define local data names in some completely informal manner; that is, there is no obligation to construct an attribute table in an external file that conforms to the style of the public dictionaries. Nevertheless, there are clear advantages to doing so: the author will benefit from standard software tools that validate data against dictionaries and the data names are more easily exported to the public domain if they subsequently become relevant to a wider community. In the following, it is assumed that the author of a new data name wishes to define fully its attributes in an appropriate standard dictionary formalism.

3.1.2.1. The _[local]_ prefix

The string `_[local]_` is *reserved* as a prefix to identify data names that do not appear in any public dictionary. (The left and right square brackets are included in this label.) Hence an author may construct private data names according to one of the following models, secure in the knowledge that the name will not appear in any global dictionary. With DDL1, a private data name will always have the form `_[local]_private_data_name`, while with DDL2 the forms `_[local]_new_category_name.private_data_name` and `_existing_category_name.[local]_private_data_name` may be used. The first DDL2 form is used for private data names in a category not already defined by a public dictionary; the second form permits the addition of local data names to an existing category. Note that the initial underscore character is dropped in the second DDL2 form.

While this convention guarantees that the new data name will not conflict with a public one, it cannot guarantee that it will not conflict with a local data name created by another author. Therefore these data names are appropriate only for testing purposes and not for release in data files that may be used by others.

3.1.2.2. Reserved prefixes

To guarantee that locally devised data names may be placed without name conflict in interchange data files, authors may register a reserved character string for their sole use. As with the special prefix `_[local]_` discussed in Section 3.1.2.1, the author's reserved prefix is simply an underscore-bounded string within the data name (*i.e.* it may not itself include an underscore character). For DDL1 applications it must be the first component of the data name; for DDL2 applications it forms the first component of the data name if describing data names in a category not defined in the official dictionaries; or the first component after the full stop (category delimiter) if the local data name is an extension to an existing category.

Table 3.1.2.1. *Reserved prefixes for private CIF data names*

String	Reserved for the use of
anbf	Australian National Beamline Facility
asd	Active Site Database
B+S	Software developers Bernstein + Sons
ccdc	Cambridge Crystallographic Data Centre
CCP4	*CCP*4 program system
cgraph	Oxford Cryosystems *Crystallographica* package
cifdic	Register of CIF dictionaries
crystmol	*CrystMol* package
csd	Cambridge Structural Database
ebi	European Bioinformatics Institute
edchem	Edinburgh University Chemistry Department
gsas	*GSAS* powder refinement system
gsk	Glaxo Smith Kline
iims	EBI project on integration of information about macromolecular structure
iucr	IUCr journal use
mdb	Model Database (Glaxo)
msd	EBI Molecular Structure Database Group
ndb	Nucleic Acids Database Project, Rutgers University
oxford	*CRYSTALS* package, University of Oxford
parvati	Validation and statistical summaries from *PARVATI* validation server
pdb	Protein Data Bank
pdbx	Protein Data Bank exchange dictionary
pdb2cif	Additions to mmCIF used by program *pdb2cif*
rcsb	Research Collaboratory for Structural Bioinformatics
shelx	*SHELXL* solution and refinement programs
vrf	Validation reply form (IUCr/*Acta Crystallographica* use)
wdc	Entries in the World Directory of Crystallographers
xtal	*Xtal* program system

Prefixes may be registered online through a web form at http://www.iucr.org/iucr-top/cif/spec/reserved.html. Table 3.1.2.1 gives a list of prefixes registered as of March 2005; this list will of course go out of date, but a current list will be maintained on the web at the address above.

An example of a data name incorporating a reserved prefix is the listing of a protein amino-acid sequence recorded temporarily by the Protein Data Bank before a protein structure is released, `_pdbx_prerelease_seq.seq_one_letter_code`.

3.1.2.3. Name spaces

The allocation of special prefixes as in Sections 3.1.2.1 and 3.1.2.2 above is a basic form of name-space allocation, because it gives authors the freedom to reproduce portions of otherwise standard data names within their own private constructions. This raises the wider question of whether a complete formalism for name-space allocation is needed. That is, the same data name might appear with different meanings in different files, provided it was clear which of the alternative definitions must be used in each case. For now, the decision has been taken not to permit the use of the same data names with different meanings in different contexts. This is to enforce uniformity of definition across the whole field of crystallography as far as is possible. This policy might be reviewed in the future if similar formalisms to CIF are created in related disciplines.

3.1.3. Formal definition process

This section describes the formal system for creating public dictionaries or appending to them. It includes information on the review and approval cycles currently required by COMCIFS, which could change if these procedures are modified. The IUCr web page