International Tables for Crystallography (2006). Vol. G, Section 3.1.3, pp. 74–75.

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 (http://www.iucr.org/iucr-top/cif) should be consulted for current practice. However, a short overview of the existing procedures is helpful in describing how the community can participate in extending the standard.

3.1.3.1. Dictionary maintenance groups

Each published dictionary authorized by COMCIFS has a group of specialists appointed or invited to extend and maintain the dictionary to serve the changing needs of the subdiscipline that sponsors the dictionary. Members of these dictionary maintenance groups (DMGs) may suggest extensions or corrigenda on their own initiative or may pass on requests for extensions from individual crystallographers. A DMG will typically debate and review any suggested amendments and produce a draft revised dictionary for approval by COMCIFS.

3.1.3.2. mmCIF review cycle

The macromolecular CIF dictionary covers a very broad and active field, and a more formal procedure exists for the submission and review of proposed extensions. Possible new definitions are submitted using *pro forma* dictionary templates to a member of an editorial board appointed by the mmCIF dictionary maintenance group. Accepted proposals are approved by the DMG and released for general community review in provisional extension dictionaries as circumstances require. The extension dictionary is revised as necessary and is finally incorporated within the parent mmCIF dictionary after COMCIFS approval has been granted.

3.1.3.3. New dictionaries

A completely new dictionary to cover a subdiscipline not otherwise catered for may be commissioned by COMCIFS or may arise from community action, occasionally sponsored by an IUCr Commission. A working group is appointed to create the dictionary and relevant example files or software. The working group is expected to test the new dictionary extensively within its own community before submitting it to COMCIFS for initial approval. It is the responsibility of COMCIFS to check the dictionary for technical consistency and for compatibility with related dictionaries. COMCIFS may refer the dictionary back to the working group for further revisions. When the dictionary finally receives formal COMCIFS approval and is published, a dictionary maintenance group is formed to promote its further development (Section 3.1.3.1). The DMG usually includes one or more members of the initial working group and at least one voting member of COMCIFS.

3.1.4. Choice of data model

The following sections of this chapter describe the technical considerations in defining data items within a dictionary. Fundamental to this is the *data model* on which the dictionary is based. The STAR File upon which CIF is based is a very versatile data format and can accommodate a variety of data models. However, the use within CIF of a single level of looping enforces a rather flat data structure and a typical CIF maps most easily onto a relational database model. This is implicit in DDL1, which assigns different attributes to data items depending on whether they appear in data loops or not. Generally speaking, one may consider a list header and its associated data values as the head and body of a table of data values. The list header (or equivalently the table head) identifies the data items ranged by column within the table. For the dictionary entries relating to the data names in the list header, the `_category` attribute collects together data items which may be looped together in the same table, and the `_list_reference`, `_list_mandatory` and `_list_uniqueness` attributes work together to indicate the data items that must be present and collectively have a unique value to identify a specific row in a table of values.

Example 3.1.4.1. *Core dictionary definitions for the atom-site labels and bond distances in a CIF table of molecular geometry.*

```
data_geom_bond_atom_site_label_
    loop_ _name
                       '_geom_bond_atom_site_label_1'
                       '_geom_bond_atom_site_label_2'
    _category                    geom_bond
    _type                        char
    _list                        yes
    _list_mandatory              yes
    _list_link_parent          '_atom_site_label'
    _definition
;     The labels of two atom sites that form a bond.
      These must match labels specified as
      _atom_site_label in the atom list.
;

data_geom_bond_distance
    _name                       '_geom_bond_distance'
    _category                    geom_bond
    _type                        numb
    _type_conditions             esd
    _list                        yes
    _list_reference    '_geom_bond_atom_site_label_'
    _enumeration_range           0.0:
    _units                       A
    _units_detail               'angstroms'
    _definition
;     The intramolecular bond distance in angstroms.
;
```

For example, the following example from the core CIF dictionary (Chapter 4.1) shows a table of bond distances. The dictionary definitions are given in Example 3.1.4.1.

```
loop_
_geom_bond_atom_site_label_1
_geom_bond_atom_site_label_2
_geom_bond_distance
  O1  C2   1.342(4)
  O1  C5   1.439(3)
  C2  C3   1.512(4)
  C2  O21  1.199(4)
  C3  N4   1.465(3)
  C3  C31  1.537(4)
  C3  H3   1.00(3)
  N4  C5   1.472(3)
```

Within the dictionary, entries for all of `_geom_bond_distance`, `_geom_bond_atom_site_label_1` and `_geom_bond_atom_site_label_2` share the same `_category` attribute, namely 'geom_bond'. (In the rest of this chapter, as elsewhere in the volume, we refer to categories by the upper-case form of their category attribute values; here, therefore, we are referring to the GEOM_BOND category.) The entry for `_geom_bond_distance` has a `_list_reference` value of `'_geom_bond_atom_site_label_'` indicating the data names that may be used to identify this particular table. The trailing underscore in this example indicates that all matching data names must be considered as components of a compound identifier; for this case the matching data names are `'_geom_bond_atom_site_label_1'` and `'_geom_bond_atom_site_label_2'`. The dictionary entry for `_geom_bond_atom_site_label_` has a `_list_mandatory` value of `yes`, indicating that these data items *must* be present within the table. In this way, the attributes specify the unique key within a database table (in this case, the key has multiple components: the labels of both contributing atom sites).

However, the mapping onto a relational database is not exact. In some cases CIFs may present data from a single category across