The CIF dictionaries: how they work

James R. Hester
August 31st, 2019
Why should we care about dictionaries?

**Programmers:** Dictionaries are the main link between data files and scientific knowledge

**Structure reporters:** CheckCIF reports in terms of data names defined in dictionaries

**Scientists:**

- Dictionaries standardise the knowledge of a field unambiguously
- New concepts can be added to dictionaries if you (the expert) help define them
Data files contain values, each of which is assigned to a data name.

A dictionary provides definitions of those data names so both you and the computer understand what the values mean.

Definitions are:

- Human-readable (for programmers, dictionary developers, bed-time reading)
- Machine-readable (for validation, transformation, calculation)

Fun fact: Nothing in the above depends on a particular format!

- The CIF format is a good standard exchange / archiving format
Anatomy of a dictionary

A collection of data name definitions - order doesn’t matter

Lines 3-6: Some header material about the dictionary

Lines 7-12: Category definition (see later)

Lines 13-23: A data name definition

```cif
#\CIF_2.0
#
# Comment

data_SimpleDic
_dictionary.date     2019-08-31
_dictionary.name     SimpleDic
# more dictionary information here...
save_sample
_definition.id       sample
_definition.class    Set
_definition.scope    Category
_definition.date     2019-08-31
save_
save_sample.size
_definition.id       '_sample.size'
_definition.text
;
   The size of the crystal from which data were measured.
;
_name.category_id   sample
_name.object_id     size
_units.code         mm
save_
# many more definitions follow here...
```
Values are assigned to “attributes” (order of appearance doesn’t matter)

An “attribute” is just a data name used for dictionary definitions

A set of attributes forms a “Dictionary Definition Language” (DDL)
Guide to CIF DDLs

• DDL1 (1993): used in “core” CIF and related dictionaries
• DDL2 (1998): used in macromolecular “mmCIF” and related dictionaries
  • Curated by the wwPDB
• DDLm (2012): developed to harmonise DDL1 and DDL2: all DDL1 dictionaries moving to DDLm
  • But the data names mean the same thing, so there is no effect on data files
Data names that can be tabulated together belong to the same category.

A category name is like a name for a table ("loop" in CIF-speak)

```
loop_
_atom_site.label
_atom_site.fract_x
_atom_site.fract_y
_atom_site.fract_z
_atom_site.U_iso_or_equiv
_atom_site.adp_type
_atom_site.occupancy

 o1 .5505(5) .6374(5) .1605(11) .035(3) Uani 1.00000
 o2 .4009(5) .5162(5) .2290(11) .033(3) Uiso 1.00000
 o3 .2501(5) .5707(5) .6014(13) .043(4) Uani 1.00000
 c1 .4170(7) .6930(8) .4954(15) .029(4) Uani 1.00000
 c2 .3145(7) .6704(8) .6425(16) .031(5) Uani 1.00000
 c3 .2789(8) .7488(8) .8378(17) .040(5) Uani 1.00000
 c4 .3417(9) .8529(8) .8859(18) .045(6) Uani 1.00000
 c5 .4445(9) .8778(9) .7425(18) .045(6) Uani 1.00000
 c6 .4797(8) .7975(8) .5487(17) .038(5) Uani 1.00000
 c7 .4549(7) .6092(7) .2873(16) .029(4) Uani 1.00000
```
On dots and underscores

Data names in DDL1 dictionaries were constructed out of words separated by underscores:

_atom_site_label

Data names in DDL2 dictionaries are constructed using the category name first, then the rest after a dot:

_atom_site.label

Data names defined in DDLm dictionaries....use the dotted form <category>.<object>

All legacy data names in non-mmCIF dictionaries have two equivalent forms!

The dots are a convention (carry no formal meaning).
The IUPAC International Chemical Identifier (InChI) is a textual identifier for chemical substances, designed to provide a standard and human-readable way to encode molecular information and to facilitate the search for such information in databases and on the web.

Types

- real, integer, complex number, arbitrary character string
- vector, matrix, list, table
  - contents
  - dimensions
- A small set of values

Units

```plaintext
save_diffrn.ambient_pressure
_definition.id 'diffrn.ambient_pressure'
_description.text
;
    Mean hydrostatic pressure at which intensities were measured.
;
# edited out ...
_type.container Single
_type.contents Real
_enumeration.range 0.0:
_units.code kilopascals
save_
```
Types

- real, integer, complex number, arbitrary character string
- vector, matrix, list, table
  - contents
  - dimensions
- A small set of values

Units

```r
save_space_group_symop.R

_definition.id '_space_group_symop.R'
_description.text
;

A matrix containing the symmetry rotation operations of a space group

| r11  r12  r13 |
R = | r21  r22  r23 |
    | r31  r32  r33 |

# edited out...

_type.container Matrix
_type.contents Real
_type.dimension '[3,3]'
```

Types

- real, integer, complex number, arbitrary character string
- vector, matrix, list, table
  - contents
  - dimensions
- A small set of values

Units

```c
save_diffrn_source.device

_definition.id 'diffrn_source.device'
_description.text

; Enumerated code for the device providing the source of radiation.
;
# edited out ...
_type.container Single
_type.contents Text
loop_
 Enumeration_set.state
 Enumeration_set.detail
tube 'sealed X-ray tube'
nuclear 'nuclear reactor'
spallation 'spallation source'
elect-micro 'electron microscope'
rot_anode 'rotating-anode X-ray tube'
synch 'synchrotron'
```
Data names with *identical* meaning

Historic names, including old underscore-only data names

```plaintext
save_refln.F_meas_su
_definition.id       '_refln.F_meas_su'
loop_
   _alias.definition_id
      '_refln_F_sigma'
      '_refln.F_meas_sigma'
      '_refln.F_meas_su'
_description.text
;  The standard uncertainty of the
   measured structure factor amplitude.
;
# more attributes here...
save_
```
Reading a definition: Relationships

- Which category the data name belongs to
- Are values drawn from values of another data name?
- Is only a single value allowed in a data block?
- Is this the standard uncertainty for a different data name?

```
save_refl.F_meas_su

_definition.id         '_refln.F_meas_su'
_description.text     ;
                      The standard uncertainty of the measured
                      structure factor amplitude.

;                            refln
_name.category_id         refln
_name.object_id           F_meas_su
_name.linked_item_id      '_refln.F_meas'
_type.purpose             SU
_type.source              Related
```

14
Reading a definition: Checking the value

- Allowed range
- Provision of standard uncertainty
- Provenance

```
save_refln.symmetry_multiplicity

_definition.id                "refln.symmetry_multiplicity"
_description.text
;
   The number of reflections symmetry-equivalent under the Laue
   symmetry to the present reflection.
;
_type.purpose       Number
_type.source        Assigned
_type.container     Single
_type.contents      Index
 Enumeration.range   1:48
```

save_
Category definitions

- Overall information about contents of the category
- Examples of complete category loops
- Category keys: data name(s) whose combined values can be used to find a unique row

```plaintext
save_CITATION

_definition.id                    CITATION
_definition.scope                Category
_definition.class                Loop
_description.text

Data items in the CITATION category record details about the literature cited as being relevant to the contents of the data block.

_name.category_id
_name.object_id
loop_
  _category_key.name
    '_citation.id'
save_
```
Creating your own definition

• Data name for local use: prepend \_\[local\]\_ or include the string \[local\]\_ after the period

• Data name that may escape your computer: register a prefix at
  http://www.iucr.org/iucr-top/cif/spec/reserved.html

• If it might be useful outside your lab, engage with the wwPDB (macromolecular) or COMCIFS (everything else)
Creating data names: Some considerations

• Data names that encode software parameters or outputs become meaningless over time, and hide scientific information
  • Instead, describe the meaning of the parameter independent of any software
• Data names that encode instrument positions are largely useless unless those positions can be related to geometry
• Data names that encode instrument settings are largely useless unless those settings can be related to commonly-understood meanings
Linking the data file to the dictionary

• Data file can indicate dictionary conformance using _audit_conform.dict_name tag
• A new tag (DDLm only): _audit.schema - if not missing and not “Default”, consult the specs
• Often software instead just checks for specific data names
COMCIFS (DDL1 and DDLm) guarantee uniqueness of data names

- DDLm:
  - Allows expansion of existing category keys, flagged using `_audit.schema`
  - Final dictionary notionally built by “importing” dictionaries upon which it depends
Finding further information on DDL attributes

DDL attributes are defined in ... their own DDL dictionaries!
Use these to:

• Check actual definition of attribute
• Find lists of possible values

```plaintext
save_type.purpose

_definition.id       _type.purpose'
_definition.class  Attribute
_description.text

; The primary purpose or function the defined data item serves in a
dictionary or a specific data instance.

; _name.category_id  type
_name.object_id     purpose
_type.purpose       State
_type.source        Assigned
_type.container     Single
_type.contents      Code
loop_
_enumeration_set.state
_enumeration_set.detail

# continued on next page...
Finding further information on DDL attributes

Describe
; Used to type items with values that are descriptive
text intended for human interpretation.
;
Encode
; Used to type items with values that are text or codes
that are formatted to be machine parsable.
;
State
; Used to type items with values that are restricted to
codes present in their "enumeration_set.state" lists.
;
Key
; Used to type an item with a value that is unique within
the looped list of these items, and may be used as a
reference "key" to identify a specific packet of items
within the category.
;
Link
;
Used to type an item that acts as a foreign key
between two categories.
;
# And so on (edited out)...

_enumeration.default

Describe

save_
DDLm allows executable code to be included in a definition, describing how to derive values of the defined data name from values of other data names.

```plaintext
save_exptl_crystal.density_diffrn

_definition.id '
_description.text ;
   Crystal density calculated from crystal unit cell and atomic content.
;
_name.category_id exptl_crystal
_name.object_id density_diffrn
#Edited out ...
loop_
   _method.purpose Evaluation
   _method.expression
     _exptl_crystal.density_diffrn = 1.6605 * _cell.atomic_mass / _cell.volume
;
save_
```
Creating a new dictionary

• Macromolecular: Liaise with the wwPDB
• Otherwise: create a group, liaise with COMCIFS and/or the relevant IUCr commission
  • Bring as many stakeholders to the table as possible
• Recent dictionaries:
  • Magnetism - driven by IUCr commission
  • Topology - driven by a small group, accepted after consultation with wider community
Development of DDLm dictionaries:
https://github.com/COMCIFS/cif_core

Core dictionary maintenance group:
https://www.iucr.org/__data/iucr/lists/coredmg/

DDL gateway:
https://www.iucr.org/resources/cif/ddl

Are you a programmer? Join the cif-developers mailing list!
https://www.iucr.org/__data/iucr/lists/cif-developers/