Introduction to the CIF Ontology

Brian McMahon
International Union of Crystallography
5 Abbey Square, Chester CH1 2HU, UK

This talk introduces the 2023 COMCIFS Dictionary Writing Workshop at the Melbourne Congress. It serves two main functions: (1) to explain the historical development of CIF dictionaries to define in a machine-actionable manner the contents of data files covering various aspects of crystallography and related structural sciences; (2) to demonstrate some of the more complex types of information that can be handled with this approach. A simple controlled vocabulary is adequate to describe individual objects (*e.g.* the volume of a unit cell). However, to describe more complex objects, the relationships between the simple 'building blocks' need to be characterised. The great strength of the CIF approach is that the building blocks and their relationships can be handled within the same formalism – a formalism, moreover, that allows the way the relationships are described to be defined also in the same way. We use the term 'CIF ontology' here to cover the canonical dictionaries maintained by the IUCr and wwPDB. The purpose of the workshop is to help participants to develop these further, but also to create their own extension dictionaries (in structural science or, indeed, any other knowledge domain).

The Crystallographic Information Framework

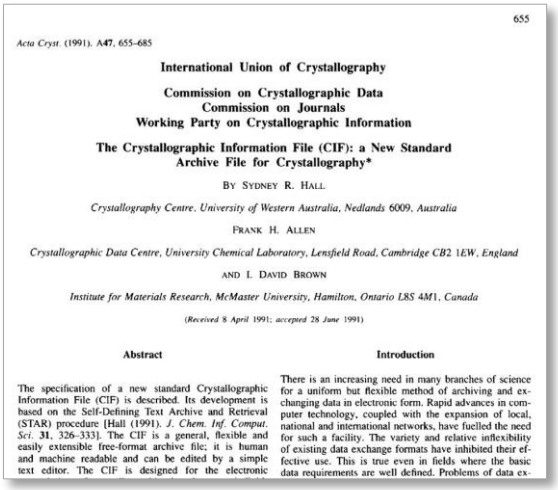
- CIF entered the world in 1991 as:

The crystallographic information file (CIF): a new standard archive file for crystallography

S. R. Hall , F. H. Allen and I. D. Brown

Acta Cryst. (1991). A**47**, 655-685

[doi:10.1107/S010876739101067X]



655

Acta Cryst. (1991). A**47**, 655–685

International Union of Crystallography

Commission on Crystallographic Data
Commission on Journals
Working Party on Crystallographic Information

The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography*

BY SYDNEY R. HALL

Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

FRANK H. ALLEN

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, England

AND I. DAVID BROWN

Institute for Materials Research, McMaster University, Hamilton, Ontario L8S 4M1, Canada

(*Received* 8 *April* 1991; *accepted* 28 *June* 1991)

Abstract

The specification of a new standard Crystallographic Information File (CIF) is described. Its development is based on the Self-Defining Text Archive and Retrieval (STAR) procedure [Hall (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333]. The CIF is a general, flexible and easily extensible free-format archive file; it is human and machine readable and can be edited by a simple text editor. The CIF is designed for the electronic

Introduction

There is an increasing need in many branches of science for a uniform but flexible method of archiving and exchanging data in electronic form. Rapid advances in computer technology, coupled with the expansion of local, national and international networks, have fuelled the need for such a facility. The variety and relative inflexibility of existing data exchange formats have inhibited their effective use. This is true even in fields where the basic data requirements are well defined. Problems of data ex-

The 1991 CIF paper of Hall, Allen and Brown was not the IUCr's first attempt to standardise information interchange. It built on the 'Standard Crystallographic File Structure' (SCFS) of David Brown in the 1980s. Like SCFS, the first iteration of CIF had two components: a specific file format, and a list of identifiers ('tags') that should be used by all compliant software.

Simple tag, value structure

```
data_crystal_structure
 _chemical_formula       'C13 H12 O5'
 _chemical_name
; 3-(2,5-dihydro-4-hydroxy-5-oxo-3-phenyl-2-furyl)propionic acid
;
 _publication_title
; Structure of WF-3681, 3-(2,5-Dihydro-4-hydroxy-5-oxo-3-phenyl-2-furyl)propionic Acid.
;

_cell_a                  18.757(8)
_cell_b                   7.282(2)
_cell_c                  17.511(8)
_cell_alpha              90
_cell_beta               91.20(3)
_cell_gamma              90
_cell_volume             2391(3)

_symmetry_space_group    '-C 2yc'
loop_ _symmetry_pos_in_xyz
 'x,y,z'              '-x,-y,-z'           '-x,y,1/2-z'
 'x,-y,1/2+z'         '1/2+x,1/2+y,z'      '1/2-x,1/2-y,-z'
 '1/2-x,1/2+y,1/2-z' '1/2+x,1/2-y,1/2+z'
```

October 22, 1989

The SCFS file format was very much in the spirit of Fortran, defining specific formatted data types and a certain ordering of presentation. The CIF format was much more fluid, allowing arbitrary whitespace to separate the identifying tags, now called 'data names', from their values. Further, there was no restriction on the order of presentation, except that closely related data values that would naturally be tabulated together were collected in a looped structure. (But even here, table columns could be permuted in any order.)

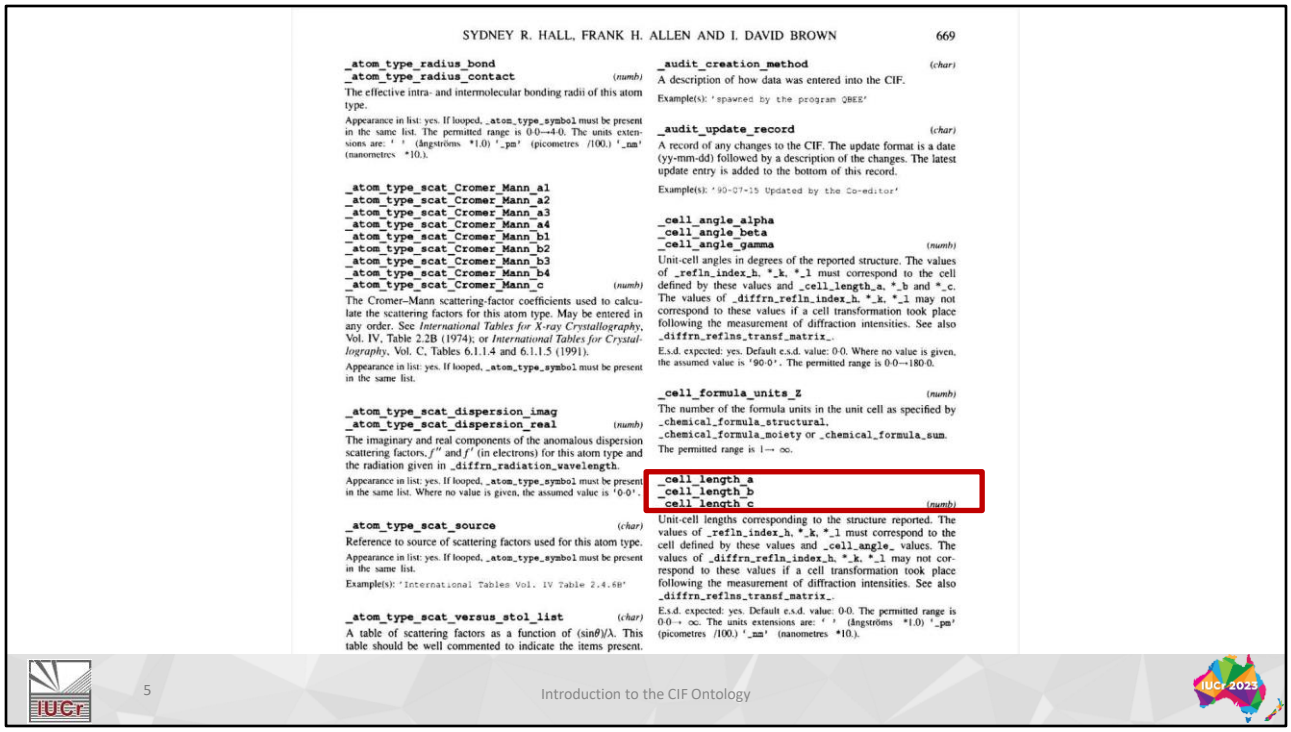
Each data name is recognized because it has a leading underscore. The value follows, separated by white space (the amount of white space does not matter). If the value itself *contains* white space, it should be wrapped in quote marks or surrounded by semicolons in the first column. Looped values are laid out as if in a table, where the data names are listed together after the loop_ keyword and the corresponding values follow in strict rotation.

Simple tag, value structure

```
data_crystal_structure
 _chemical_formula        'C13 H12 O5'
 _chemical_name
; 3-(2,5-dihydro-4-hydroxy-5-oxo-3-phenyl-2-furyl)propionic acid
;
 _publication_title
; Structure of WF-3681, 3-(2,5-Dihydro-4-hydroxy-5-oxo-3-phenyl-2-furyl)propionic Acid.
;

_cell_a                  18.757(8)
_cell_b                   7.282(2)
_cell_c                  17.511(8)
_cell_alpha              90
_cell_beta               91.20(3)
_cell_gamma              90
_cell_volume             2391(3)

_symmetry_space_group    '-C 2yc'
loop_ _symmetry_pos_in_xyz
 'x,y,z'              '-x,-y,-z'           '-x,y,1/2-z'
 'x,-y,1/2+z'         '1/2+x,1/2+y,z'      '1/2-x,1/2-y,-z'
 '1/2-x,1/2+y,1/2-z' '1/2+x,1/2-y,1/2+z'
```

October 22, 1989

Each data name could occur only once in a file (strictly, in a *data block*; files could be partitioned into separate data blocks, each representing a standalone description of, say, a crystal structure or a diffraction experiment).



SYDNEY R. HALL, FRANK H. ALLEN AND I. DAVID BROWN 669

_atom_type_radius_bond
_atom_type_radius_contact (*numb*)
The effective intra- and intermolecular bonding radii of this atom type.
Appearance in list: yes. If looped, _atom_type_symbol must be present in the same list. The permitted range is 0·0→4·0. The units extensions are: ' ' (ångströms *1.0) '_pm' (picometres /100.) '_nm' (nanometres *10.).

_atom_type_scat_Cromer_Mann_a1
_atom_type_scat_Cromer_Mann_a2
_atom_type_scat_Cromer_Mann_a3
_atom_type_scat_Cromer_Mann_a4
_atom_type_scat_Cromer_Mann_b1
_atom_type_scat_Cromer_Mann_b2
_atom_type_scat_Cromer_Mann_b3
_atom_type_scat_Cromer_Mann_b4
_atom_type_scat_Cromer_Mann_c (*numb*)
The Cromer–Mann scattering-factor coefficients used to calculate the scattering factors for this atom type. May be entered in any order. See *International Tables for X-ray Crystallography*, Vol. IV, Table 2.2B (1974); or *International Tables for Crystallography*, Vol. C, Tables 6.1.1.4 and 6.1.1.5 (1991).
Appearance in list: yes. If looped, _atom_type_symbol must be present in the same list.

_atom_type_scat_dispersion_imag
_atom_type_scat_dispersion_real (*numb*)
The imaginary and real components of the anomalous dispersion scattering factors, f'' and f' (in electrons) for this atom type and the radiation given in _diffrn_radiation_wavelength.
Appearance in list: yes. If looped, _atom_type_symbol must be present in the same list. Where no value is given, the assumed value is '0·0'.

_atom_type_scat_source (*char*)
Reference to source of scattering factors used for this atom type.
Appearance in list: yes. If looped, _atom_type_symbol must be present in the same list.
Example(s): 'International Tables Vol. IV Table 2.4.6B'

_atom_type_scat_versus_stol_list (*char*)
A table of scattering factors as a function of $(\sin\theta)/\lambda$. This table should be well commented to indicate the items present.

_audit_creation_method (*char*)
A description of how data was entered into the CIF.
Example(s): 'spawned by the program QBEE'

_audit_update_record (*char*)
A record of any changes to the CIF. The update format is a date (yy-mm-dd) followed by a description of the changes. The latest update entry is added to the bottom of this record.
Example(s): '90-07-15 Updated by the Co-editor'

_cell_angle_alpha
_cell_angle_beta
_cell_angle_gamma (*numb*)
Unit-cell angles in degrees of the reported structure. The values of _refln_index_h, *_k, *_l must correspond to the cell defined by these values and _cell_length_a, *_b and *_c. The values of _diffrn_refln_index_h, *_k, *_l may not correspond to these values if a cell transformation took place following the measurement of diffraction intensities. See also _diffrn_reflns_transf_matrix_.
E.s.d. expected: yes. Default e.s.d. value: 0·0. Where no value is given, the assumed value is '90·0'. The permitted range is 0·0→180·0.

_cell_formula_units_Z (*numb*)
The number of the formula units in the unit cell as specified by _chemical_formula_structural, _chemical_formula_moiety or _chemical_formula_sum.
The permitted range is 1→ ∞.

_cell_length_a
_cell_length_b
_cell_length_c (*numb*)
Unit-cell lengths corresponding to the structure reported. The values of _refln_index_h, *_k, *_l must correspond to the cell defined by these values and _cell_angle_ values. The values of _diffrn_refln_index_h, *_k, *_l may not correspond to these values if a cell transformation took place following the measurement of diffraction intensities. See also _diffrn_reflns_transf_matrix_.
E.s.d. expected: yes. Default e.s.d. value: 0·0. The permitted range is 0·0→ ∞. The units extensions are: ' ' (ångströms *1.0) '_pm' (picometres /100.) '_nm' (nanometres *10.).

In the original paper, each data name was defined, alongside information about the values associated with it (*e.g.* type, numeric range constraints, permitted code terms *etc*.) in a layout similar to that of a conventional linguistic dictionary.

A textual ‘dictionary’ of definitions…

_cell_length_a
_cell_length_b
_cell_length_c *(numb)*

Unit-cell lengths corresponding to the structure reported. The values of **_refln_index_h**, ***_k**, ***_l** must correspond to the cell defined by these values and **_cell_angle_** values. The values of **_diffrn_refln_index_h**, ***_k**, ***_l** may not correspond to these values if a cell transformation took place following the measurement of diffraction intensities. See also **_diffrn_reflns_transf_matrix_**.

E.s.d. expected: yes. Default e.s.d. value: 0·0. The permitted range is 0·0→ ∞. The units extensions are: ‘ ’ (ångströms *1.0) ‘_pm’ (picometres /100.) ‘_nm’ (nanometres *10.).

Notice how, in the manner of a lexicographic dictionary, the definition includes a number of attributes explaining how the term should be used. These attributes were initially somewhat *ad hoc*, but were stored in a CIF-format file that was the precursor of the dictionary paradigm we use today. This informal set of attribute definitions is sometimes called ‘DDL0’.

… became (1997) a data file of machine-readable attributes

```
data_cell_length_
    loop_ _name                 '_cell_length_a'
                                '_cell_length_b'
                                '_cell_length_c'
    _category                     cell
    _type                         numb
    _type_conditions              esd
    _enumeration_range            0.0:
    _units                        A
    _units_detail               'angstroms'
    _definition
;               Unit-cell lengths in angstroms corresponding to the structure
                reported. The values of _refln_index_h, *_k, *_l must
                correspond to the cell defined by these values and _cell_angle_
                values. The values of _diffrn_refln_index_h, *_k, *_l may not
                correspond to these values if a cell transformation took place
                following the measurement of the diffraction intensities. See
                also _diffrn_reflns_transf_matrix_.
;
```


By 1997, the attributes themselves were formally described using the same attribute set, and this was the first 'official' dictionary definition language, DDL1. Some of the early attributes permitted by the original CIF paper (*e.g.* ability to generate variant data names according to the physical units used) were omitted from DDL1. The decision was taken to enforce a single physical unit to be associated with each defined physical quantity. This did not exclude the possibility of defining new data names where the only difference was in the associated physical unit, but the practice was discouraged.

… further refined (2013) as a relational ontology

```
save_cell.length_a

    _definition.id                '_cell.length_a'
    _alias.definition_id          '_cell_length_a'
    _name.category_id             cell
    _name.object_id               length_a



    _definition.update            2014-06-08
    _description.text
;
     The length of each cell axis.
;
    _type.purpose                 Measurand
    _type.source                  Recorded
    _type.container               Single
    _type.contents                Real
    _enumeration.range            1.:
    _units.code                   angstroms

save_
```

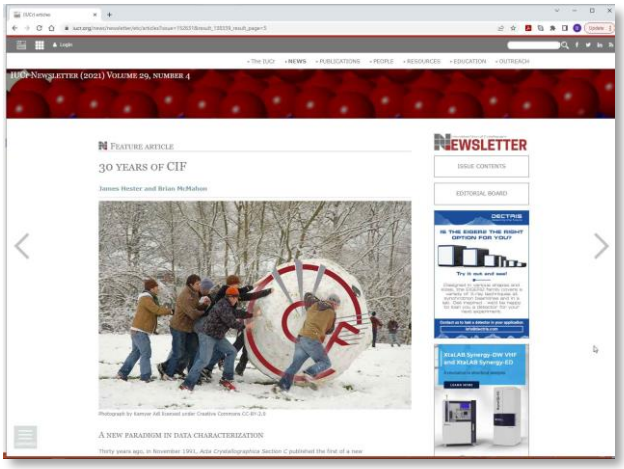

The similarity of the category/included data names structure that was established in DDL1 to the structure of a relational database became apparent to the working group developing the mmCIF dictionary for describing protein and nucleic acid structures. They developed a variant DDL (designated DDL2) that was almost completely isomorphous to a relational database schema. This is the basis for the current wwPDB implementation, is fully mature, and continues to be extended as needed to accommodate the specific requirements of the PDB and structural biology community. It won't be discussed here (but details can be found in the handout for the last COMCIFS workshop at https://www.iucr.org/resources/cif/comcifs/workshop-2017)

However, the relational approach was a significant enhancement to the original DDL, and in 2013 a new language, DDLm, was introduced that was capable of expressing the relational attributes that were in DDL2, but that also allowed the inclusion of algorithmic methods to express or calculate relationships between different data items. This new DDLm would also be able to describe the attribute behaviours of DDL1, and so was potentially universal in scope. A new structuring of data names was encouraged through a **`_CATEGORY.item`** construction (*i.e.* using a dot separator); but legacy files with different data names (all underscores) could be supported through an aliasing mechanism.

30(+) Years of CIF

- *IUCr Newsletter* article (2021)
  - Historical background
  - Contributors
  - Diversification of uses
  - Importance to journals and databases
  - Future prospects

Reprinted in workshop booklet



Some more historical background and a description of how CIF became established among different structural science communities is found in a relatively recent *IUCr Newsletter* article, which is also reprinted in the workshop booklet.

The CORE of the CIF ontology

- Scope of the original CIF specification: single-crystal structures
  - Description of the diffraction experiment (X-ray, neutron, electron)
  - Crystal characteristics (size, shape, composition, space group)
  - Diffracted beam intensities (point detectors, area detectors)
  - Structure solution and refinement (methodology, software)
  - Structure factors (allowing reprocessing, replicability)
  - 3D structure (molecular geometry, displacement factors)
  - Publication (authors, citations, database entries, commentary)

So we turn to an overview of the core CIF dictionary that was developed for the 1991 CIF article, and the extensions that have subsequently been built under the aegis of the IUCr and the wwPDB. These canonical dictionaries we are referring to collectively as the *CIF Ontology*.

Definitions organised by CATEGORY

Category EXPTL

```
_exptl.crystals_number
_exptl.method
_exptl.method_details
_exptl.special_details
_exptl.transmission_factor_max
_exptl.transmission_factor_min
```
etc.

Category EXPTL_CRYSTAL_FACE

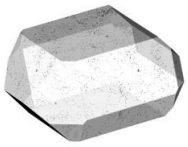
```
_exptl_crystal_face.index_h
_exptl_crystal_face.index_k
_exptl_crystal_face.index_l
_exptl_crystal_face.diffr_chi
_exptl_crystal_face.diffr_kappa
_exptl_crystal_face.diffr_phi
_exptl_crystal_face.diffr_psi
_exptl_crystal_face.perp_dist
```
etc.

Category EXPTL_CRYSTAL_APPEARANCE

```
_exptl_crystal_appearance.general
_exptl_crystal_appearance.hue
_exptl_crystal_appearance.intensity
```

Example

```
_exptl_crystal_appearance.general
                                'opaque'

loop_
  _exptl_crystal_face.index_h
  _exptl_crystal_face.index_k
  _exptl_crystal_face.index_l
  _exptl_crystal_face.perp_dist

   -1 -1  0   0.725
   -1 -1  1   0.764
   -1  0  0   0.548
   -1  0  1   0.597
    0 -1 -1   0.616
    0  0 -1   0.391
    0  1 -1   1.355
```



Example from Kaminsky, W. (2007). From CIF to virtual morphology using the *WinXMorph* program. *J. Appl. Cryst.* **40**, 382-385

The essential characteristic of the CIF dictionaries is that related items are gathered together into categories. If you are storing CIF information in a relational database, then each category can simply be considered as a table. However, some items will only occur once (at least within the scope of the subject area covered by that dictionary). It is possible to display everything in tabular format, but in the case of these 'degenerate' or 'scalar' categories, it can seem more natural to list the items where the data value immediately follows the declared data name. This could, for example, allow items within such a category to appear at several locations within a file (potentially useful if data is accumulated at different stages along a workflow). Categories that are expected to have single-instance data names are indicated with the attribute **`_definition.class 'Set'`**, while the more typical multi-value (tabular) categories are identified with **`_definition.class 'Loop'`**.

Category hierarchy of core CIF dictionary

- HEAD category : CIF_CORE (establishes 'root' of definition family)
- 6 'themes'
  - DIFFRACTION
  - EXPTL
  - FUNCTION
  - MODEL
  - PUBLICATION
  - STRUCTURE

There is an organisational hierarchy amongst categories in the core and other CIF dictionaries, though relationships between categories are expressed through foreign keys and are not dependent on this notional hierarchy. There are also parent-child category relationships, which express 'projections' or instances when sub-categories are separated out into different tables to escape sparse presentations. Ultimately, relationships are all expressed through inter-category pointers independently of the notional organisational hierarchy. Nevertheless, for understanding large dictionaries it is helpful to use this hierarchy as a way of collecting together categories that are thematically related, and we use the notion of 'theme' to structure our discussion of the core dictionary.

Core categories by 'theme': PUBLICATION

In practice, much broader than the term 'publication' suggests: essentially metadata relating to the overall research project, and so common to any scientific programme

AUDIT
- AUDIT_AUTHOR
- AUDIT_AUTHOR_ROLE
- AUDIT_CONFORM
- AUDIT_CONTACT_AUTHOR
- AUDIT_LINK
- AUDIT_SUPPORT

CITATION
CITATION_AUTHOR
CITATION_EDITOR
COMPUTING
DATABASE
- DATABASE_CODE

DATABASE_RELATED
DISPLAY
- DISPLAY_COLOUR

JOURNAL
- JOURNAL_COEDITOR
- JOURNAL_DATE
- JOURNAL_INDEX
- JOURNAL_TECHEDITOR

PUBL
- PUBL_AUTHOR
- PUBL_BODY
- PUBL_CONTACT_AUTHOR
- PUBL_MANUSCRIPT
  - PUBL_MANUSCRIPT_INCL_EXTRA

PUBL_REQUESTED
PUBL_SECTION

The 'publication' group is, as indicated on the slide, really a broad collection of generic metadata that could apply to any information collection. It is possible that these categories might even be separated out into a distinct ('common' ?) dictionary.

Core categories by 'theme': EXPTL

Experimental work carried out prior to diffraction measurements

CHEMICAL
- CHEMICAL_CONN_ATOM
- CHEMICAL_CONN_BOND
- CHEMICAL_FORMULA

EXPTL_ABSORPT

EXPTL_CRYSTAL
- EXPTL_CRYSTAL_APPEARANCE
- EXPTL_CRYSTAL_FACE

SPACE_GROUP
- SPACE_GROUP_GENERATOR
- SPACE_GROUP_SYMOP
- SPACE_GROUP_WYCKOFF

Categories in this 'exptl' group are associated with experimental work carried out prior to diffraction measurements. Many of these are specific to crystallographic experiments (chemical characterisation, crystal preparation, space-group determination), and it is clear that any other experimental science would have a parallel requirement to record preparation and set-up information about an experiment.

Core categories by ‘theme’: DIFFRACTION

Description of a diffraction experiment and its data collection and recording

DIFFRN
- CELL
  - CELL_MEASUREMENT
    - CELL_MEASUREMENT_REFLN

DIFFRN_ATTENUATOR
DIFFRN_DETECTOR
DIFFRN_MEASUREMENT
DIFFRN_ORIENT
- DIFFRN_ORIENT_MATRIX
- DIFFRN_ORIENT_REFLN

DIFFRN_RADIATION
- DIFFRN_RADIATION_WAVELENGTH

DIFFRN_REFLN
DIFFRN_REFLNS
- DIFFRN_REFLNS_CLASS
- DIFFRN_REFLNS_TRANSF_MATRIX

DIFFRN_SCALE_GROUP
DIFFRN_SOURCE
DIFFRN_STANDARDS
- DIFFRN_STANDARD_REFLN

REFLN
- REFLNS
- REFLNS_CLASS
- REFLNS_SCALE
- REFLNS_SHELL

These categories cover the data collected during a diffraction experiment. They also include much experimental metadata, and numerical data that could be, in the nomenclature favoured by the IUCr Committee on Data, ‘raw’ (collected straight from the instrument) or ‘processed’ (in the crystallographic context, typically structure factors).

Core categories by 'theme': STRUCTURE

Other structure-related information, including refinement strategy

- ATOM
  - ATOM_ANALYTICAL
    - ATOM_ANALYTICAL_MASS_LOSS
    - ATOM_ANALYTICAL_SOURCE
  - ATOM_SITE
    - ATOM_SITE_ANISO
  - ATOM_SITES
    - ATOM_SITES_CARTN_TRANSFORM
    - ATOM_SITES_FRACT_TRANSFORM
  - ATOM_TYPE
  - ATOM_TYPE_SCAT
- REFINE
  - REFINE_DIFF
  - REFINE_LS
    - REFINE_LS_CLASS

These categories generally provide information about the structural model constructed after solution and refinement (in CommDat terminology, 'derived' data).

Core categories by ‘theme’: MODEL

Details of molecular or ionic structure derived from the inferred positions of atoms in a determined structure

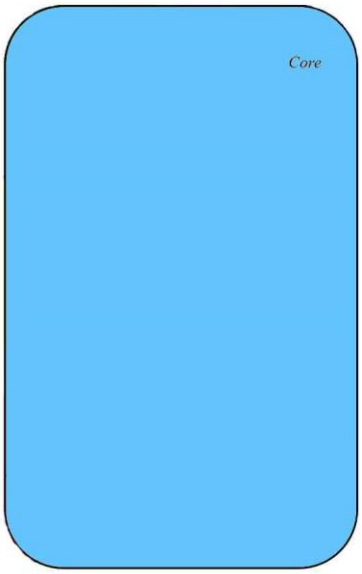
- GEOM
  - GEOM_ANGLE
  - GEOM_BOND
  - GEOM_CONTACT
  - GEOM_HBOND
  - GEOM_TORSION
- MODEL_SITE
- VALENCE
  - VALENCE_PARAM
  - VALENCE_REF

Broadly speaking, these cover the description of the chemical characteristics of the species investigated that are largely independent of its crystalline state.

Core categories by 'theme': FUNCTION

Internal functional relationships used in DDLm methods	FUNCTION

This is a purely internal category that allows computational methods to be included for validating or evaluating relationships between data items.



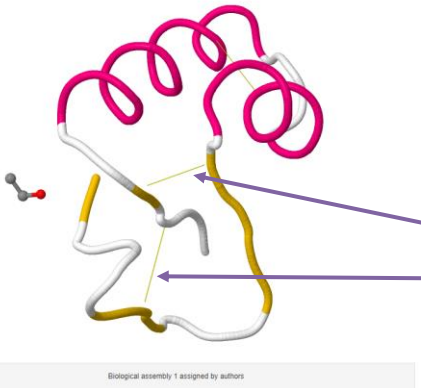
We will present a series of diagrams showing how categories are developed amongst the various dictionaries that have extended the core ontology since 1991.

mmCIF – describing biological molecules

1CNR

CORRELATED DISORDER OF THE PURE PRO22(SLASH)LEU25 FORM OF CRAMBIN AT 150K REFINED TO 1.05 ANGSTROMS RESOLUTION

Yamano, A. & Teeter, M.M. (1994). Correlated disorder of the pure Pro22/Leu25 form of crambin at 150 K refined to 1.05-Å resolution. *J. Biol. Chem.* **269**, 13956-13965.



Biological assembly 1 assigned by authors

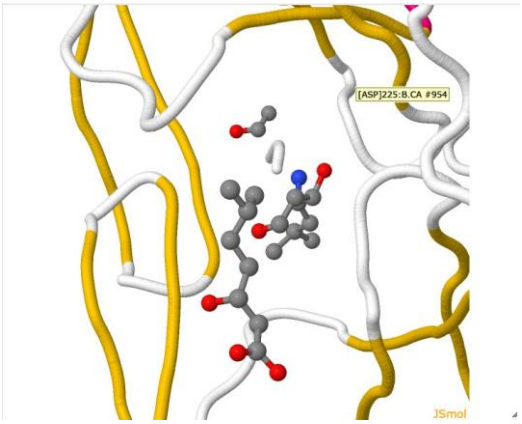
Interactions between portions of the crambin structure

```
loop_
  _struct_conn.id
  _struct_conn.conn_type_id
  _struct_conn.ptnr1_label_comp_id
  _struct_conn.ptnr1_label_asym_id
  _struct_conn.ptnr1_label_seq_id
  _struct_conn.ptnr1_label_atom_id
  _struct_conn.ptnr1_role
  _struct_conn.ptnr1_symmetry
  _struct_conn.ptnr2_label_comp_id
  _struct_conn.ptnr2_label_asym_id
  _struct_conn.ptnr2_label_seq_id
  _struct_conn.ptnr2_label_atom_id
  _struct_conn.ptnr2_role
  _struct_conn.ptnr2_symmetry
  _struct_conn.details
   SS2 disulf CYS chain_a  4 S  .        1_555
              CYS chain_a 32 S  .        1_555 .
   SS1 disulf CYS chain_a  3 S  .        1_555
              CYS chain_a 40 S  .        1_555 .
   HB1 hydrog SER chain_a  6 OG positive 1_555
              LEU chain_a  8 O  negative 1_556 .
   HB2 hydrog ARG chain_a 17 N  positive 1_555
              ASP chain_a 43 O  negative 1_554 .
```


We begin with the macromolecular CIF dictionary, which rapidly grew into a large-scale undertaking and has become the responsibility of the Worldwide Protein Data Bank for future maintenance and extension. We will not therefore discuss it in any detail here, but present a couple of slides to demonstrate the richness and complexity of relationships that it is capable of expressing. The STRUCT_CONN category records details about the connections between portions of a macromolecular structure. These can be hydrogen bonds, salt bridges, disulfide bridges and so on. The STRUCT_CONN_TYPE records define the criteria used to identify these connections. This example describes two disulfide bonds between portions of the folded protein backbone, as well as two hydrogen bonds (not illustrated).

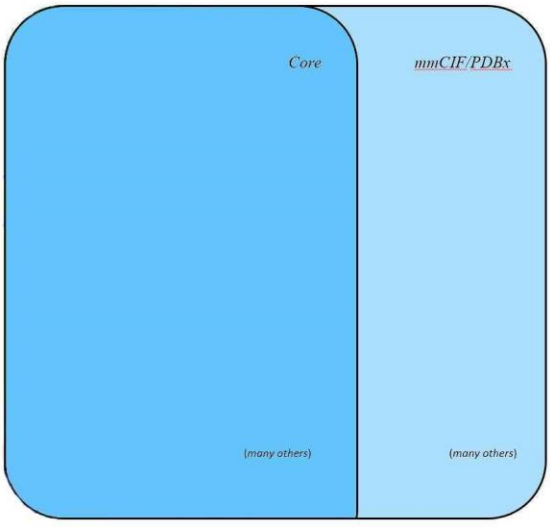
mmCIF structure – alternative conformations

```
loop_
_atom_sites_alt_ens.id
_atom_sites_alt_ens.details
  'Ensemble 1'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  This conformational ensemble includes the more
  populated conformation of the inhibitor (id=1) and
  the amino acid side chains that correlate with this
  inhibitor conformation.
;
  'Ensemble 2'
; The inhibitor binds to the enzyme in two, roughly
  twofold symmetric, alternative conformations.

  This conformational ensemble includes the less
  populated conformation of the inhibitor (id=2) and
  the amino acid side chains that correlate with this
  inhibitor conformation.
;

loop_
_atom_sites_alt_gen.ens_id
_atom_sites_alt_gen.alt_id
   'Ensemble 1'  .
   'Ensemble 1'  1
   'Ensemble 2'  .
   'Ensemble 2'  2
```

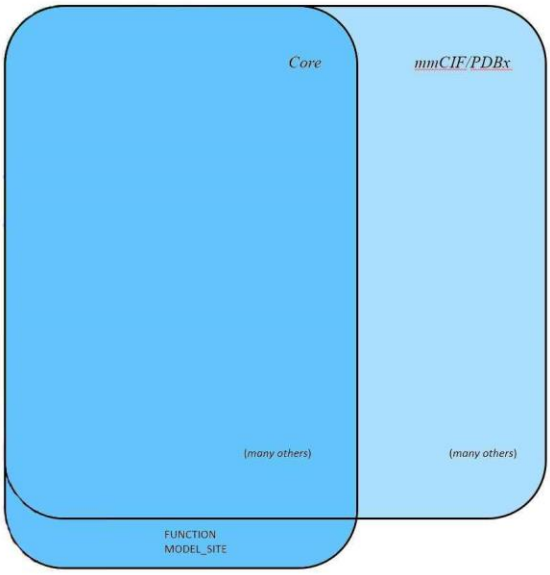


Alternative conformations in an HIV-1 protease structure (Fitzgerald *et al.*, 1990) described with data items in the ATOM_SITES_ALT, ATOM_SITES_ENS and ATOM_SITES_GEN categories.

This example demonstrates how alternative conformations in the same structure can be handled. There are two conformations in which the inhibitor binds to the enzyme in a HIV-a protease structure. Reference: Fitzgerald, P.M., McKeever, B.M., VanMiddlesworth, J.F., Springer, J.P., Heimbach, J.C., Leu, C.T., Herber, W.K., Dixon, R.A., Darke, P.L. (1990). Crystallographic analysis of a complex between human immunodeficiency virus type 1 protease and acetyl-pepstatin at 2.0-Å resolution. *J. Biol. Chem.* **265**, 14209-14219.



The original mmCIF dictionary included all the existing core categories, and introduced many new ones. (Some, such as those describing bibliographic citations, were more detailed than existing treatments in the core, but were developed as completely distinct categories.) The mmCIF dictionary was thus a proper superset of the core.



With the introduction of DDLm methods, some new categories have been added to the core dictionary that are not present in mmCIF.

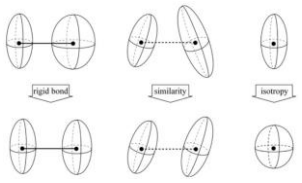
‘Simple’ extension dictionaries - restraints

Description of restraints and constraints that can be applied in a structure refinement program

RESTR
RESTR_ANGLE
RESTR_DISTANCE
RESTR_DISTANCE_MIN
RESTR_EQUAL_ANGLE
RESTR_EQUAL_ANGLE_CLASS
RESTR_EQUAL_DISTANCE
RESTR_EQUAL_DISTANCE_CLASS
RESTR_EQUAL_TORSION
RESTR_EQUAL_TORSION_CLASS
RESTR_PARAMETER
RESTR_PARAMETER_CLASS
RESTR_PLANE
RESTR_PLANE_CLASS
RESTR_RIGID_BODY
RESTR_RIGID_BODY_CLASS
RESTR_TORSION
RESTR_U_ISO
RESTR_U_RIGID
RESTR_U_SIMILAR

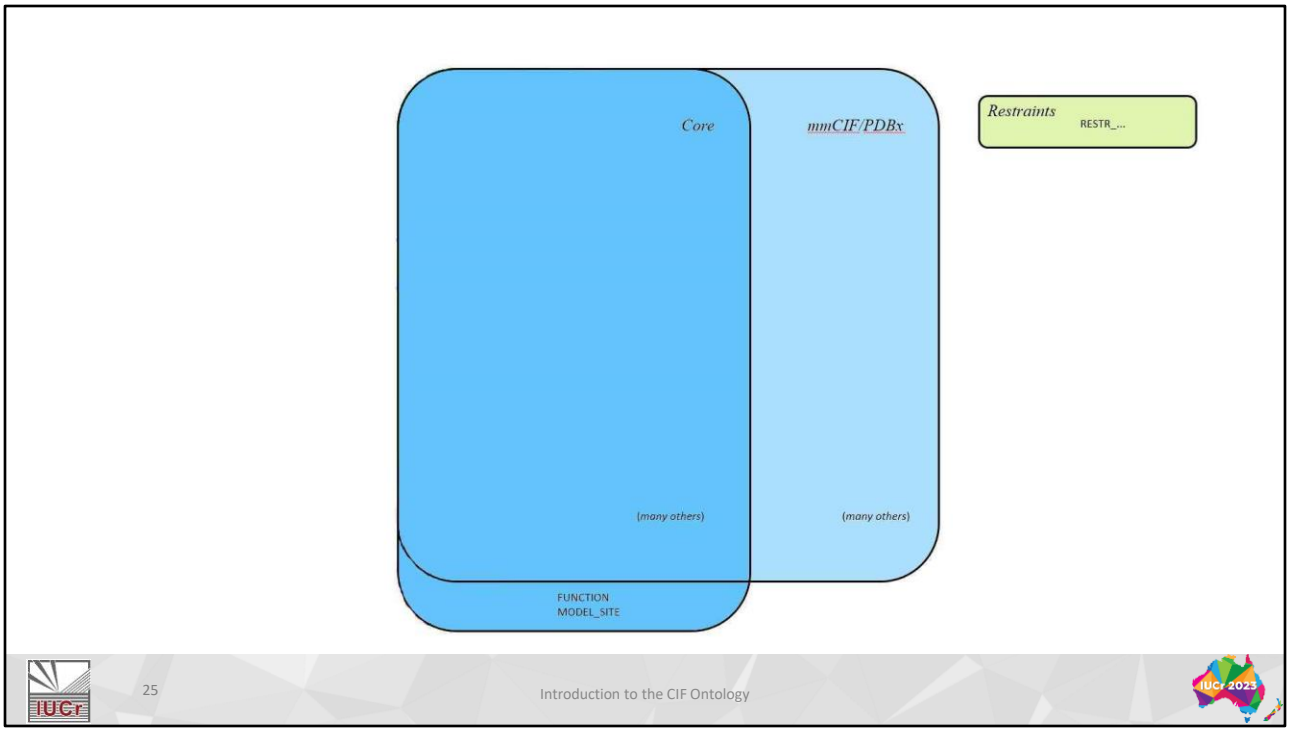
```
loop_
  _restr_distance.atom_site_label_1
  _restr_distance.atom_site_label_2
  _restr_distance.site_symmetry_2
  _restr_distance.target
  _restr_distance.target_weight_param
  _restr_distance.diff
  _restr_distance.details
    C1 C2 1_555 1.524 0.04 -0.032
                  'generated by SHELX DFIX'
    C2 C3 1_555 1.340 0.04 0.051
                  'generated by SHELX DFIX'
    Na1 Ca1 1_555 0.0 0.0 0.0
     'Na1 is constrained to occupy Ca site'
    Fe1 Ca1 1_555 0.0 0.0 0.0 .
    Al1 Ca1 1_555 0.0 0.0 0.0 .
```

Example 1: listing of constrained bond distances.



Example 2: types of constraints on atomic displacement parameters described by data items in the RESTR_U_ISO category.

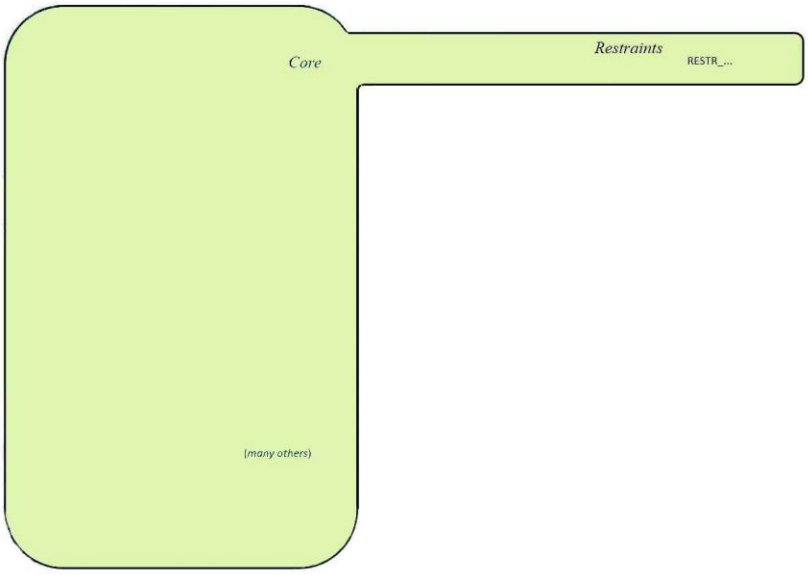
We now consider the individual extension dictionaries that have been developed with COMCIFS supervision. We will not go through them in great detail, but in each case we will try to illustrate the novel descriptive features that the dictionary is interested in developing. The restraints dictionary describes the types of geometric or ADP restraints that have been applied during a refinement strategy. Because of the dynamic way in which refinement software can arbitrarily restrain parameters, this cannot be done with complete generality. Nevertheless, a series of descriptions has been developed that provide insight into the refinement strategy. The rigid-bond restraint assumes that for atoms bound to one another the amplitude of motion along the direction of the bond is similar. The similar ADP restraint assumes that atoms close to one another would move similarly with respect to direction and amplitude. The isotropy restraint assumes that the atomic motion is approximately spherical. From Müller (2009).



The categories in the restraints dictionary all begin **RESTR_** and so are distinct from any in the core.

An aside

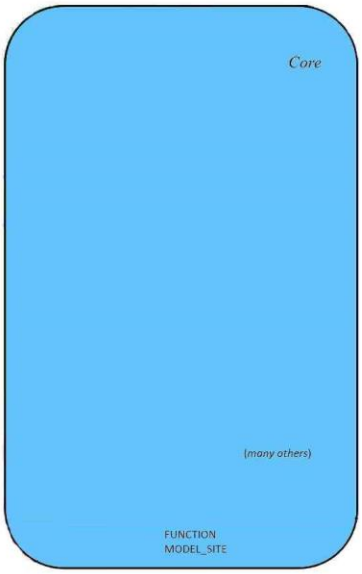
Strictly speaking, from the ontological (*i.e.* conceptual) point of view, the extension dictionaries such as cif_restraints.dic are proper **supersets** of the core, as they import all its definitions

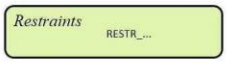


A brief visual diversion to reinforce that these diagrams are sets of discrete category names, rather than an attempt to indicate the ontological completeness of an assembly of dictionaries. Ontologically, an extension dictionary such as cif_restraints.dic includes all the definitions in the core (and any other dictionaries it subsumes through **`_import.get`** statements or implicitly, by inheritance).

An aside

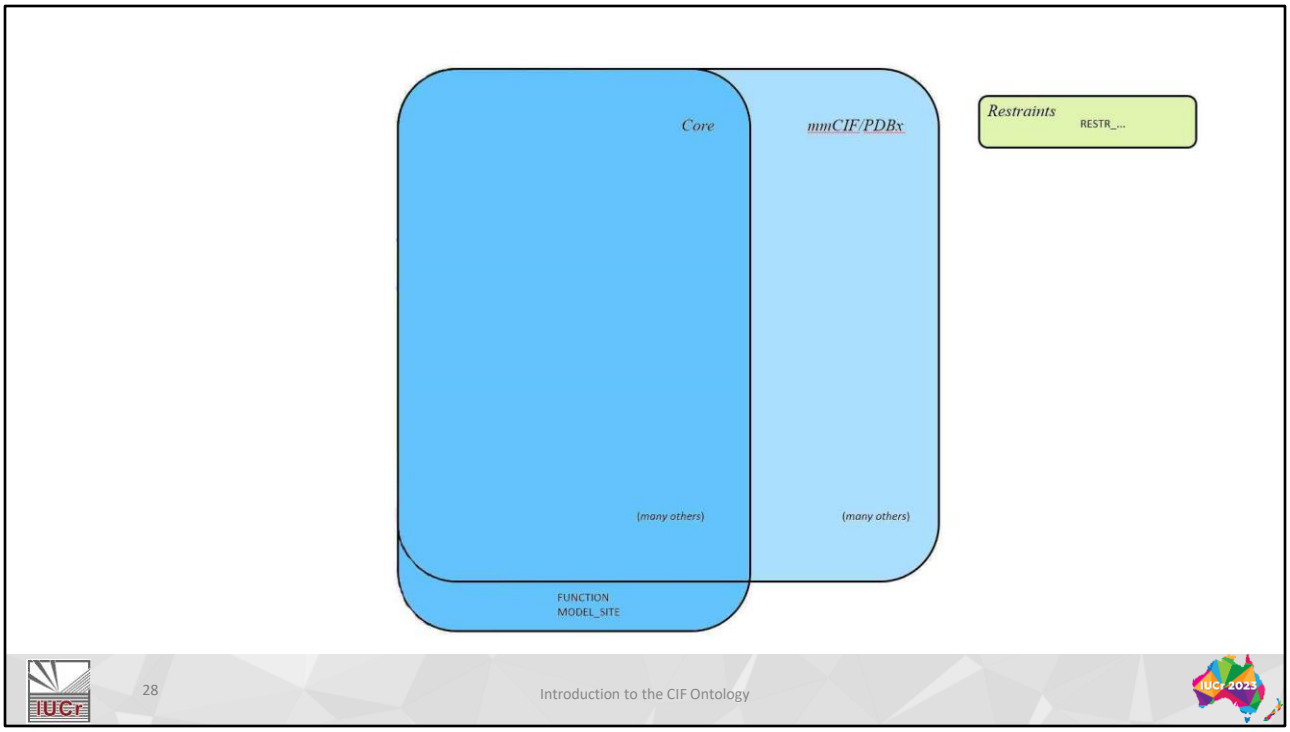
Strictly speaking, from the ontological (*i.e.* conceptual) point of view, the extension dictionaries such as cif_restraints.dic are proper **supersets** of the core, as they import all its definitions





However, for diagrammatic simplicity these figures indicate the categories defined explicitly in each of the dictionaries (*i.e.* outside of any _import.get statements including other dictionaries)

So, to reiterate, we draw the separate dictionaries as separate sets (unless they do share categories, as will be seen later).

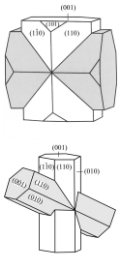


And we'll include the mmCIF for good measure, as we build up our collection.

‘Simple’ extension dictionaries - twinning

Description of twinning. The TWIN_REFLN category partitions reflections according to the twin component. Other categories capture the methodology of twin assignment.

TWIN
- TWIN_INDIVIDUAL
- TWIN_REFLN



```
_twin.formation_mechanism          gt
_twin.dimensionality               triperiodic
_twin.morphology                   polysynthetic
_twin.special_details
;       This is an example of a twin in space group
        P 32 by the 2 x,x,0 twin operation.
;

loop_
    _twin_individual.id
    _twin_individual.mass_fraction_refined
    _twin_individual.twin_lattice_type
    _twin_individual.twin_matrix_11
    _twin_individual.twin_matrix_12
    _twin_individual.twin_matrix_13
    _twin_individual.twin_matrix_21
    _twin_individual.twin_matrix_22
    _twin_individual.twin_matrix_23
    _twin_individual.twin_matrix_31
    _twin_individual.twin_matrix_32
    _twin_individual.twin_matrix_33
       1 0.545(2) ref   1  0  0  0  1  0  0  0  1
       2 0.455(2) mt_II 0  1  0  1  0  0  0  0 -1
```

Example showing twinning by merohedry for a trigonal crystal system. Reflection list (right) shows assignment to twin components 1 and 2.

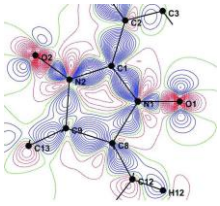
```
loop_
   _twin_refln.datum_id
   _twin_refln.individual_id
   _twin_refln.index_h
   _twin_refln.index_k
   _twin_refln.index_l
   _twin_refln.F_squared_calc
   _twin_refln.F_squared_meas
   _twin_refln.F_squared_sigma
   _twin_refln.include_status
     1    1  -1   1 -32     40.17     55.86      7.39 o
     1    2   1  -1  32     40.17     55.86      7.39 o
     2    1   0   1 -32    125.83    145.63      9.36 o
     2    2   1   0  32    125.83    145.63      9.36 o
     3    1   1   1 -32     10.33     30.89      7.39 o
     3    2   1   1  32     10.33     30.89      7.39 o
     4    1  -3   2 -32     28.26     26.84      9.47 o
     4    2   2  -3  32     28.26     26.84      9.47 o
     5    1  -2   2 -32      4.31     13.21      9.78 o
     5    2   2  -2  32      4.31     13.21      9.78 o
     6    1  -1   2 -32     10.71     17.48      9.88 o
     6    2   2  -1  32     10.71     17.48      9.88 o
     7    1   0   2 -32     13.94     37.66     10.51 o
     7    2   2   0  32     13.94     37.66     10.51 o
     8    1   1   2 -32     58.60     75.41      9.88 o
     8    2   2   1  32     58.60     75.41      9.88 o
     9    1  -2   3 -32     27.75     41.40      8.01 o
     9    2   3  -2  32     27.75     41.40      8.01 o
    10    1  -1   3 -32     59.22     65.53      8.11 o
    10    2   3  -1  32     59.22     65.53      8.11 o
```


The twinning dictionary actually contains only these three categories (apart from the TWIN_GROUP parent) which are distinct from any categories in the core. Note that the TWIN_REFLN list may contain numeric values that also occur in the core REFLN list (*e.g.* measured and calculated squared structure-factors associated with diffraction peaks) but are partitioned differently, and must therefore be indexed with different category keys. The illustration shows two macroscopic twin forms of the mineral staurolite. The example listings are for a different structure.

‘Simple’ extension dictionaries – electron density

Multipole expansion description of electron density associated with atom sites in the crystal structure

ATOM_LOCAL_AXES
ATOM_RHO_MULTIPOLE
  ATOM_RHO_MULTIPOLE_COEFF
  ATOM_RHO_MULTIPOLE_KAPPA
  ATOM_RHO_MULTIPOLE_RADIAL_SLATER




```
loop_
_atom_rho_multipole_atom_label    _atom_rho_multipole_coeff_Pv
_atom_rho_multipole_coeff_P00     _atom_rho_multipole_coeff_P11
_atom_rho_multipole_coeff_P1-1    _atom_rho_multipole_coeff_P10
_atom_rho_multipole_coeff_P20     _atom_rho_multipole_coeff_P21
_atom_rho_multipole_coeff_P2-1    _atom_rho_multipole_coeff_P22
_atom_rho_multipole_coeff_P2-2    _atom_rho_multipole_coeff_P30
_atom_rho_multipole_coeff_P31     _atom_rho_multipole_coeff_P3-1
_atom_rho_multipole_coeff_P32     _atom_rho_multipole_coeff_P3-2
_atom_rho_multipole_coeff_P33     _atom_rho_multipole_coeff_P3-3
_atom_rho_multipole_coeff_P40     _atom_rho_multipole_coeff_P41
_atom_rho_multipole_coeff_P4-1    _atom_rho_multipole_coeff_P42
_atom_rho_multipole_coeff_P4-2    _atom_rho_multipole_coeff_P43
_atom_rho_multipole_coeff_P4-3    _atom_rho_multipole_coeff_P44
_atom_rho_multipole_coeff_P4-4    _atom_rho_multipole_kappa
_atom_rho_multipole_kappa_prime0
_atom_rho_multipole_kappa_prime1
_atom_rho_multipole_kappa_prime2
_atom_rho_multipole_kappa_prime3
_atom_rho_multipole_kappa_prime4
Ni2+(1) 2.38(4) 0.32(4) 0.00 0.00 -0.02(1) 0.00(2) 0.00 0.00
0.00 0.00 -0.08(1) 0.00 0.00 0.00 0.00 0.06(1) -0.04(1) 0.05(1)
0.00 0.00 0.00 0.00 -0.20(1) 0.08(1) 0.00 0.00 1.04(1) 0.44(1)
0.44 1.15(4) 0.44 1.15
```

Example 1 - Multipole coefficients for the nickel ion in [Ni(H3L)][NO3][PF6].

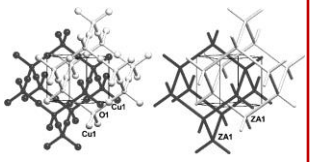
The detailed electron density around individual atoms can be described by a multipole expansion method [Hansen, N. K. & Coppens, P. (1978). *Acta Cryst*. A**34**, 909–921]. In this formalism, local axes must be defined around each atom of interest, and the spherical harmonic coefficients according to the specific method of Hansen & Coppens may be enumerated. This is therefore a very specialised application. There is growing interest in returning to the topic of electron density to create a dictionary that can accommodate a wider range of modern quantum crystallographic formalisms.

‘Simple’ extension dictionaries – topology

Describes topological and connectivity properties of periodic nets

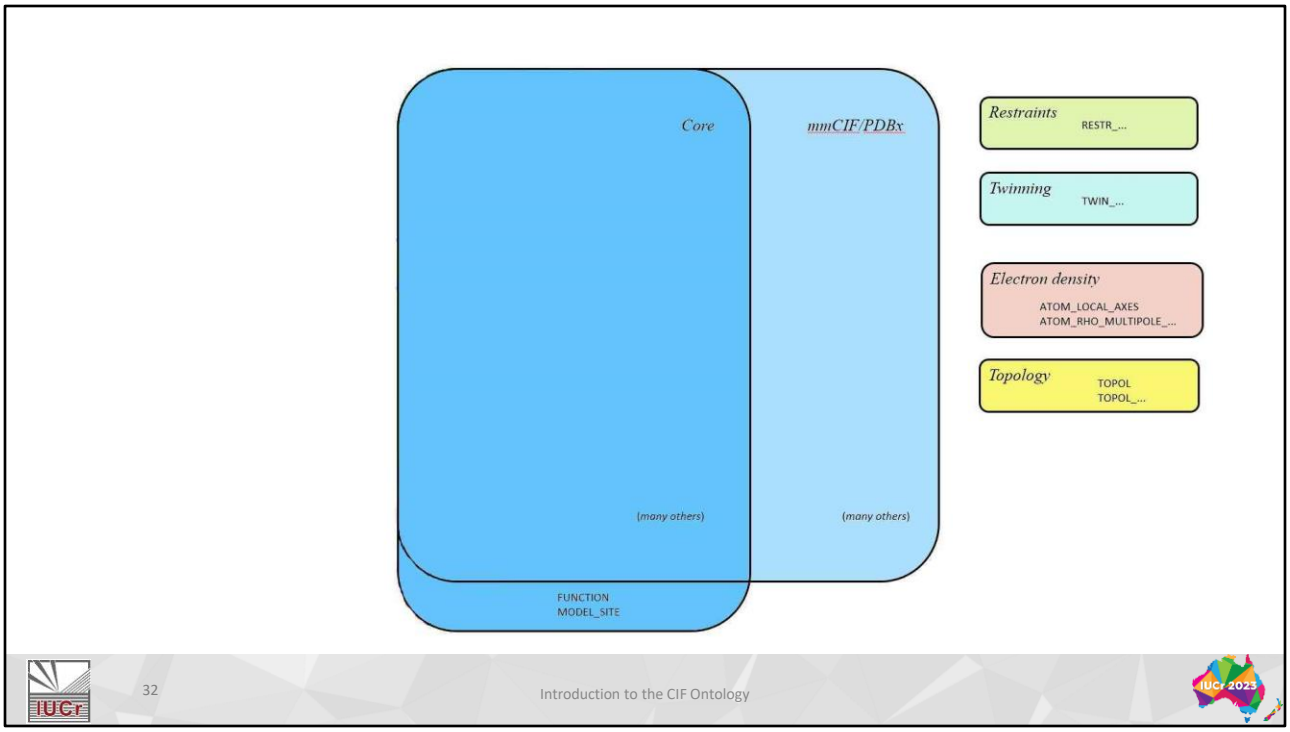
TOPOL
TOPOL_ATOM
TOPOL_ENTANGL
TOPOL_LINK
TOPOL_NET
TOPOL_NODE
TOPOL_TILING

```
_topol_net.id                    1
_topol_net.z_number              2
_topol_net.overall_topology_RCSR 'dia'

_topol_node.id         1
_topol_node.net_id     1

_topol_link.id         1
_topol_link.node_id_1  1
_topol_link.node_id_2  1
_topol_link.symop_id_2 13
_topol_link.type       gl

loop_
    _topol_atom.id
    _topol_atom.node_id
    _topol_atom.link_id
    _topol_atom.atom_label
    _topol_atom.element_symbol
        1  1  .  O1   O
        2  .  1  Cu1  Cu
```



Example: Representation of two interpenetrating diamondoid nets in the Cu_2O crystal structure

The topology dictionary is largely concerned with lattice descriptions rather than the physical structure of a crystal. However, it is clearly of interest to relate topological nets with an underlying distribution of atoms, and the TOPOL_ATOM category provides a link to a structure described in the core dictionary by requiring that values of **`_topol_atom.atom_label`** match corresponding values of **`_atom_site.label`** in a core category.

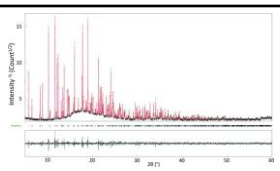


So all of these dictionaries contain their own discrete categories and can sit apart from the core in our diagrammatic representation.

Extension dictionaries – powder

Describes powder data collection, intensity profiles, time-of-flight, structure refinement, phase mixtures, etc.

PD_AMORPHOUS
PD_BACKGROUND
PD_BLOCK
PD_BLOCK_DIFFRACTOGRAM
PD_CALC_COMPONENT
PD_CALC_OVERALL
PD_CALIB
PD_CALIBRATION
PD_CALIB_DETECTED_INTENSITY
PD_CALIB_D_TO_TOF
PD_CALIB_INCIDENT_INTENSITY
PD_CALIB_OFFSET
PD_CALIB_STD
PD_CALIB_WAVELENGTH
PD_CHAR
PD_DATA
  PD_CALC
  PD_MEAS
  PD_PROC
PD_DIFFRACTOGRAM
PD_INSTR
PD_INSTR_DETECTOR
PD_MEAS_INFO_AUTHOR
PD_MEAS_OVERALL
PD_PEAK
PD_PEAK_OVERALL
PD_PHASE
PD_PHASE_BLOCK
PD_PHASE_MASS
PD_PREF_ORIENT
PD_PREF_ORIENT_MARCH_DOLLASE
PD_PREF_ORIENT_SPHERICAL_HARMONICS
PD_PREP
PD_PROC_INFO_AUTHOR
PD_PROC_LS
PD_PROC_OVERALL
PD_QPA_CALIB_FACTOR
PD_QPA_EXTERNAL_STD
PD_QPA_INTENSITY_FACTOR
PD_QPA_INTERNAL_STD
PD_QPA_OVERALL
PD_SPEC
¶REFLN



```
_pd_meas_number_of_points                5500
_pd_meas_2theta_range_min                5.000
_pd_meas_2theta_range_max                59.990
_pd_meas_2theta_range_inc                0.01

_diffrn_radiation_probe                  X-ray

_pd_proc_ls_profile_function             'Pseudo-Voigt'
_pd_proc_ls_background_function
 'Manual background combined with 20 Legendre polynoms'
_pd_proc_ls_pref_orient_corr             none
_pd_proc_ls_prof_R_factor                0.0586
_pd_proc_ls_prof_wR_factor               0.0785
_pd_proc_ls_prof_wR_expected             0.0580

_pd_proc_2theta_range_min                5
_pd_proc_2theta_range_max                59.99
_pd_proc_2theta_range_inc                0.01

loop_
 _pd_proc_point_id
 _pd_proc_2theta_corrected
 _pd_proc_intensity_net
 _pd_calc_intensity_net
 _pd_proc_ls_weight
    1      5.0500     211.01     200.96      14.53
    2      5.0600     209.87     200.66      14.49
    3      5.0700     212.65     200.37      14.58
    4      5.0800     203.98     200.09      14.28
    5      5.0900     185.85     199.80      13.63
    6      5.1000     192.20     199.53      13.86
    7      5.1100     190.23     199.25      13.79
```

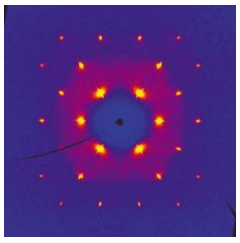
Example: Powder CIF and final Rietveld plot of phase II of griseofulvin at room temperature.

The powder dictionary is another large extension dictionary, and covers the rather diverse range of experimental types and structural characterisation met with in powder diffraction. Since a sample analysed by powder diffraction often contains distinct phase mixtures, powder CIFs are often organized with multiple data blocks, each data block describing one candidate phase. This dictionary introduced mechanisms for defining the relationship of each such data block to others present in the same file. There is an overlap with one core category (REFLN), as indicated by the pilcrow symbol, but I found it difficult to draw an overlap with the core diagram in a way that didn't interfere with the other overlaps I wanted to show.

Extension dictionaries – image

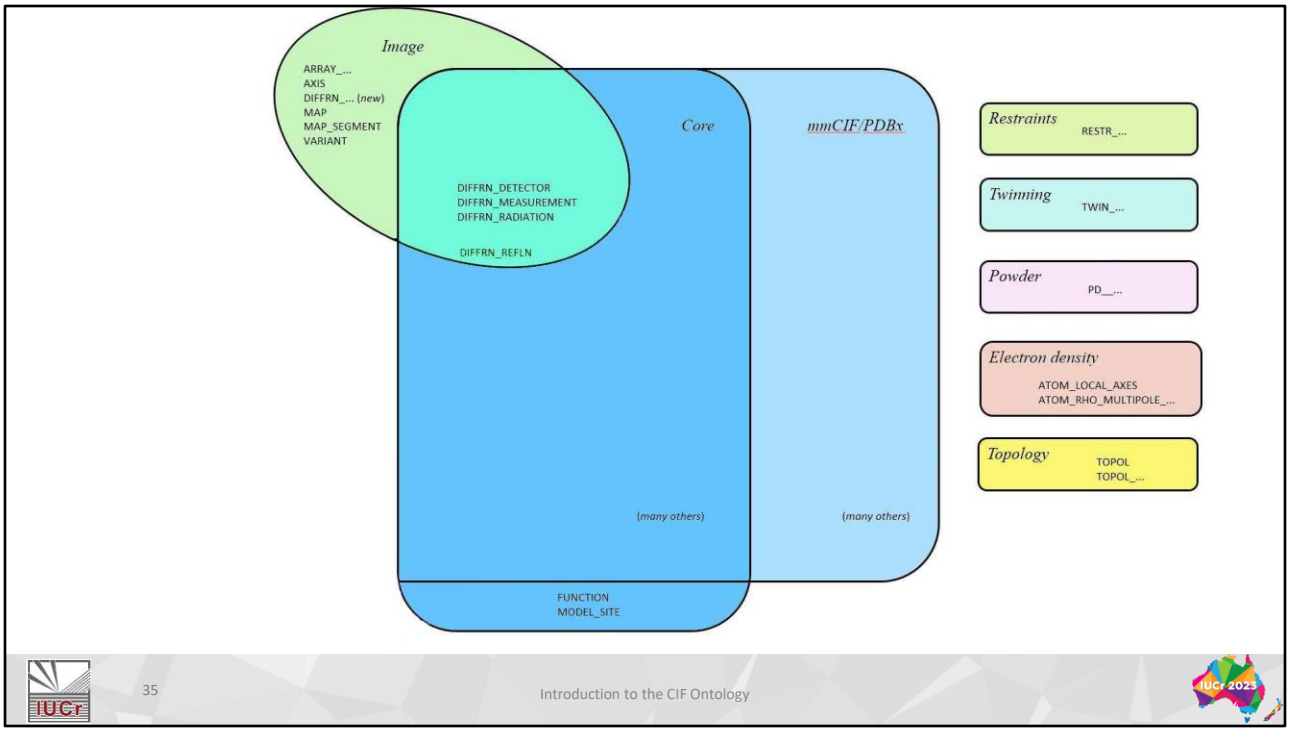
Description of raw image data collected in a diffraction experiment

ARRAY_DATA
ARRAY_DATA_EXTERNAL_DATA
ARRAY_ELEMENT_SIZE
ARRAY_INTENSITIES
ARRAY_STRUCTURE
ARRAY_STRUCTURE_LIST
ARRAY_STRUCTURE_LIST_AXIS
ARRAY_STRUCTURE_LIST_SECTION
AXIS
DIFFRN_DATA_FRAME
¶ DIFFRN_DETECTOR
DIFFRN_DETECTOR_AXIS
DIFFRN_DETECTOR_ELEMENT
DIFFRN_FRAME_DATA
¶ DIFFRN_MEASUREMENT
DIFFRN_MEASUREMENT_AXIS
¶ DIFFRN_RADIATION
¶ DIFFRN_REFLN
DIFFRN_SCAN
DIFFRN_SCAN_AXIS
DIFFRN_SCAN_COLLECTION
DIFFRN_SCAN_FRAME
DIFFRN_SCAN_FRAME_AXIS
DIFFRN_SCAN_FRAME_MONITOR
MAP
MAP_SEGMENT
VARIANT



Example: Reciprocal space of the plastic phase of cyclohexane, at 255 K, viewed in a projection normal to [011].

The core dictionary did allow for ‘raw’ data from point detectors, which were common at the time CIF was developed. However, area detectors were already beginning to become available, and a separate image dictionary was developed to handle diffraction images from a range of detector types. It was hoped that the image data could be standardised (using the ARRAY_DATA and related categories), but detector manufacturers proved resistant to conforming to such a standard, and imgCIF (and the associated binary CBF files) had relatively little uptake. However, latterly with growing community interest in archiving raw diffraction images, imgCIF has re-emerged as the natural vehicle at least for describing the experimental metadata, and it is used in IUCr journals to ensure reusability of data sets discussed in the *Raw Data Letters* section of the journal *IUCrData*. Increasingly high data capture rates make creation of CBF files on the fly increasingly difficult, but the ARRAY_DATA categories are still available for providing a standard image format if needed or desired for long-term archiving of selected data sets.

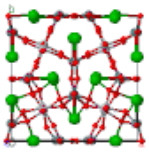


Because the image dictionary extended the description of raw data beyond the provisions already present in the core dictionary, it extended several core categories as well as introducing many new ones.

Extension dictionaries – modulated structures

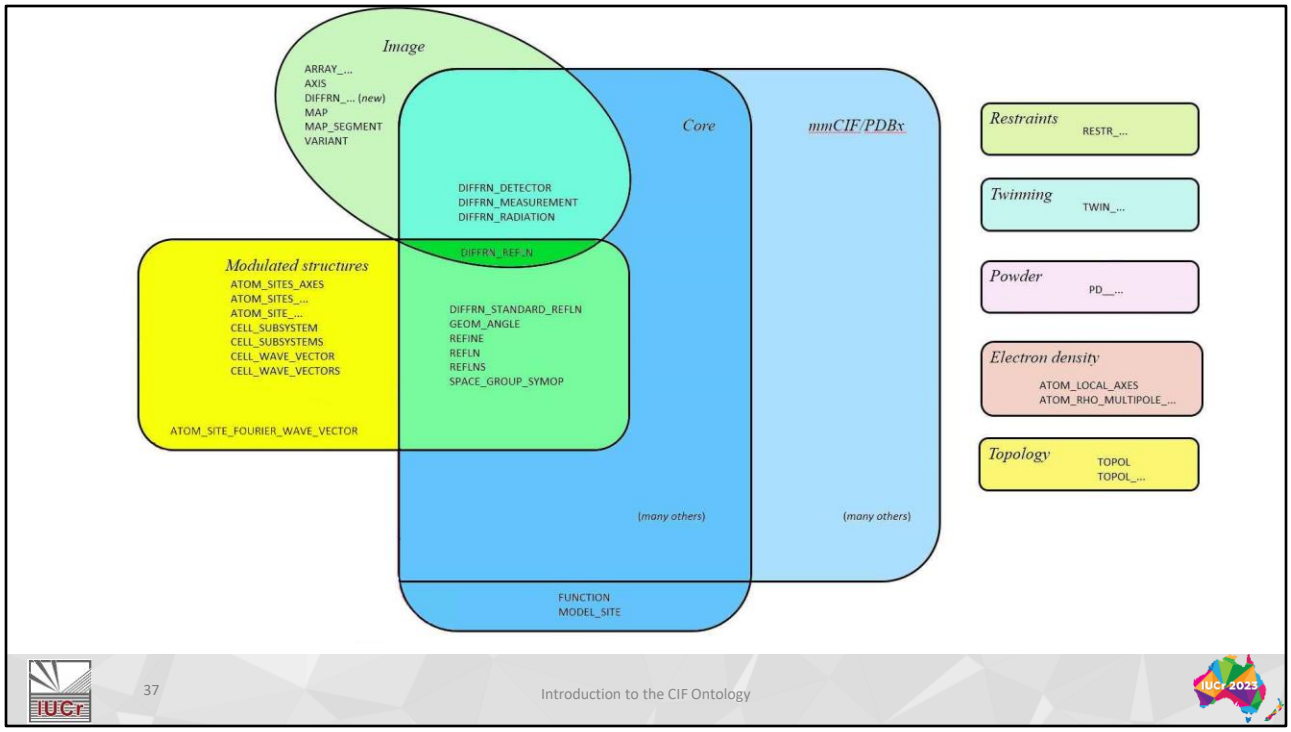
Description of modulated and incommensurate structures as superposition of distinct structural models or projection into 3-space of higher-dimensional symmetries

ATOM_SITES_AXES
ATOM_SITES_DISPLACE_FOURIER
ATOM_SITES_MODULATION
ATOM_SITES_ORTHO
ATOM_SITES_ROT_FOURIER
ATOM_SITE_DISPLACE_FOURIER
ATOM_SITE_DISPLACE_FOURIER_PARAM
ATOM_SITE_DISPLACE_LEGENDRE
ATOM_SITE_DISPLACE_ORTHO
ATOM_SITE_DISPLACE_SPECIAL_FUNC
ATOM_SITE_DISPLACE_XHARM
ATOM_SITE_FOURIER_WAVE_VECTOR
ATOM_SITE_OCC_FOURIER
ATOM_SITE_OCC_FOURIER_PARAM
ATOM_SITE_OCC_LEGENDRE
ATOM_SITE_OCC_ORTHO
ATOM_SITE_OCC_SPECIAL_FUNC
ATOM_SITE_OCC_XHARM
ATOM_SITE_PHASON
ATOM_SITE_ROT_FOURIER
ATOM_SITE_ROT_FOURIER_PARAM
ATOM_SITE_ROT_LEGENDRE
ATOM_SITE_ROT_ORTHO
ATOM_SITE_ROT_SPECIAL_FUNC
ATOM_SITE_ROT_XHARM
ATOM_SITE_U_FOURIER
ATOM_SITE_U_FOURIER_PARAM
ATOM_SITE_U_LEGENDRE
ATOM_SITE_U_ORTHO
ATOM_SITE_U_XHARM
CELL_SUBSYSTEM
CELL_SUBSYSTEMS
CELL_WAVE_VECTOR
CELL_WAVE_VECTORS
¶ DIFFRN_REFLN
¶ DIFFRN_REFLNS
¶ DIFFRN_STANDARD_REFLN
¶ GEOM_ANGLE *etc.*
¶ REFINE
¶ REFLN
¶ REFLNS
¶ SPACE_GROUP_SYMOP



Example: Incommensurate modulation in Fresnoite framework-structure type $Ba_2TiGe_2O_8$

Another challenge was addressed by the modulated and composite structures dictionary (first available in 2002). The need was to extend the core description of crystal structures that had conventional space-group symmetries into the growing area of aperiodic structures. There are two main approaches: to describe distinct structural components that overlap in a non-periodic manner; and to use superspace groups to represent the aperiodicities as projections into 3-space of higher-dimensional symmetries. Spatial modulations were described by a variety of harmonic modulation functions, and superspace group nomenclatures and symmetry operations were introduced.

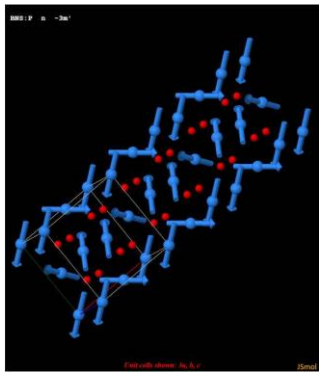


The resultant dictionary overlaps with and extends several categories already existing in the core. And, because the characteristics of the diffraction spots are changed by the aperiodicity, there is also overlap with the DIFFRN_REFLN category which is now also common to imgCIF.

Extension dictionaries – magnetic structures

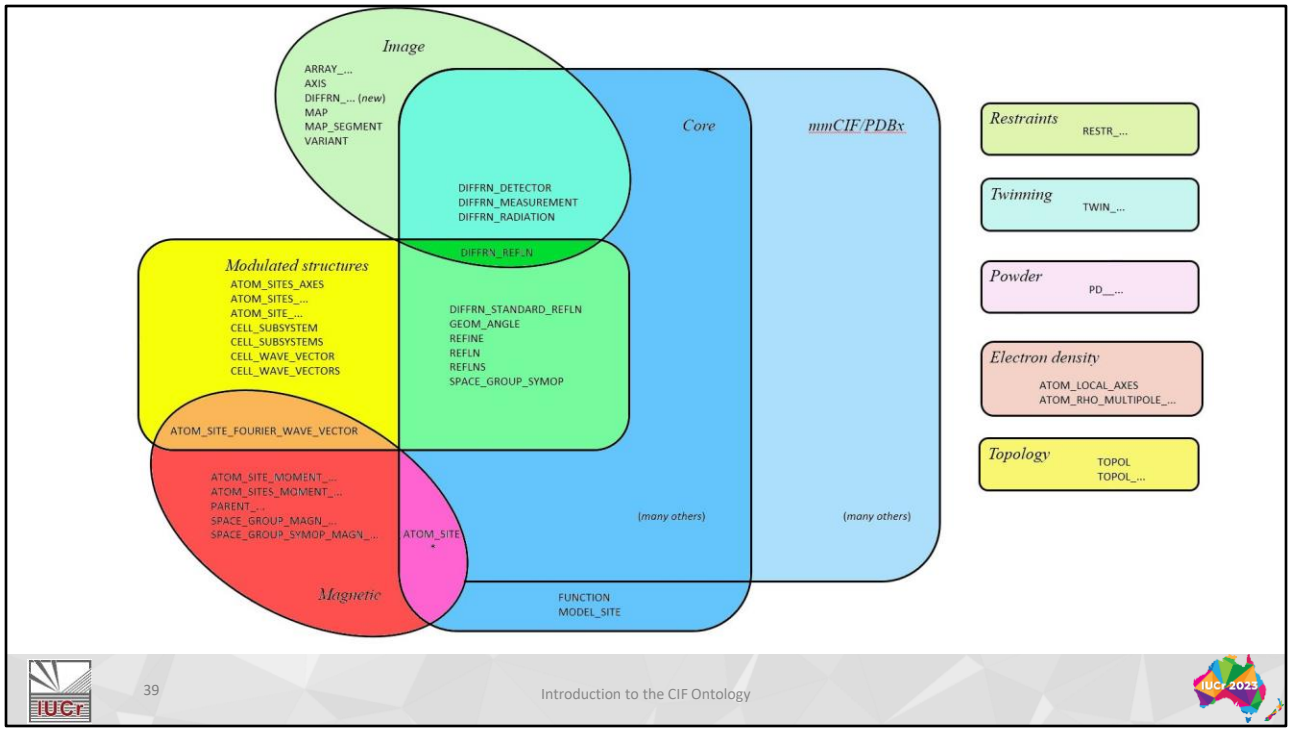
Description of magnetic properties of crystal structures

§ ATOM_SITE_FOURIER_WAVE_VECTOR
¶ ATOM_SITE_MOMENT
ATOM_SITE_MOMENT_FOURIER
ATOM_SITE_MOMENT_FOURIER_PARAM
ATOM_SITE_MOMENT_SPECIAL_FUNC
ATOM_SITES_MOMENT_FOURIER
PARENT_PROPAGATION_VECTOR
PARENT_SPACE_GROUP
SPACE_GROUP_MAGN
SPACE_GROUP_MAGN_SSG_TRANSFORMS
SPACE_GROUP_MAGN_TRANSFORMS
SPACE_GROUP_SYMOP_MAGN_OG_CENTERING
SPACE_GROUP_SYMOP_MAGN_CENTERING
SPACE_GROUP_SYMOP_MAGN_OPERATION
SPACE_GROUP_SYMOP_MAGN_SSG_CENTERING
SPACE_GROUP_SYMOP_MAGN_SSG_OPERATION



Example: Magnetic propagation vectors of UO_2.

On top of the description of the atomic positions and motions in a crystal structure, there is now also growing interest in magnetic properties, and the magnetic CIF dictionary under development adds descriptions of magnetic moments and symmetries to the core description.



As magnetic properties are found in both periodic and aperiodic structures, so there is overlap with categories in both the core and the modulated structures dictionaries.

Summary: data definitions in CIF dictionaries

Managed by COMCIFS

- Crystallographic Core (coreCIF) – 1991 and ongoing
- Crystallographic Restraints – 2011
- Crystallographic Powder Diffraction (pdCIF) – 1997
- Modulated and Composite Structures (msCIF) – 2002
- Multipole Electron Density (rhoCIF) – 2003
- Crystallographic Twinning – 2014
- Magnetic Structures (magCIF) – 2016
- Lattice topology (topoCIF) – 2021
- Crystallographic Symmetry (symCIF) – 2001
- Diffraction Images (imgCIF) – 2000
- High pressure – under development
- Crystallographic Macromolecular Structure (mmCIF) – 1997

Managed by wwPDB

- Crystallographic Macromolecular Structure (mmCIF) – 1997
- PDB Exchange Dictionary (PDBx/mmCIF) – 1997 and ongoing
- Integrative/Hybrid (I/H) methods – 2017
- 3DEM Extension Dictionary – 2004
- NMRSTAR Dictionary – 2013
- Biological Small Angle Scattering – 1998
- Model Archive Extension Dictionary – 2018
- BIOSYNC Extension Dictionary – 2000
- NMR Exchange Format Dictionary – 2016

This is the current list of distinct dictionaries openly available to the crystallographic community. There are also some local dictionaries (for specific software applications or database implementations), and a growing number of comparable dictionaries in other similar fields (materials science, biological NMR structures).