Small-molecule extension dictionaries

James Hester

Sept 2nd, 2019

More than one dictionary?

No one person or group has the specialist expertise

COMCIFS checks for style and harmony with other dictionaries

- Duplicated concepts
- Duplicated names
- Syntax, well-written definition text

What you should get out of this talk:

- Information for your particular field now and in the future
- An idea of dictionary construction problems and their solutions

Extra small-molecule dictionaries

- Constraints and restraints
- Twinning
- Electron density
- Images
- Topology
- Powder diffraction
- Modulated and composite structures
- Magnetism

Constraints and restraints

Restraint: one or more parameters are restrained to lie within some range; like having an extra observation

Constraint: specific value required for one or more parameters; reduces number of free parameters.

Vital for assessing the value of published refinements.

Twinning

A twinned crystal in physical space is composed of two or more separate crystallites.

In reciprocal space the peaks from each crystallite are related to the others by some rotation.

If the underlying space group and “twin rule” relating the two structures can be found, h,k,l can be assigned and measured intensities used to solve and refine a structure.

The twinning CIF dictionary

Twinning classification information

Twin individual information: relative mass (refined) and orientation matrix relative to first individual

`_twin_refln`: list of reflections, each reflection assigned to a particular twin

Electron density dictionary (rhoCIF)

One way of describing the electron density in a crystal: as a sum of a series of multipole functions centred on each atom. Refine the parameters of these multipoles as an improvement on the simple “independent (spherical) atom model”.

This dictionary provides data names corresponding to the parameters and axis definitions

Note the string `_local_` in some of these data names: remember, the convention for your local data names is to include `_[local]_`.

Recording Images

Raw image data from a 2D detector. As well as the raw data, need to describe:

- goniometer and detector geometry
- method of intensity encoding and compression
- size and layout of the pixels in real space
- organisation of images into frames (for composite detectors) and scans

...and then you can process the data with confidence

CIF dictionary for raw images: imgCIF

Usually associated with mmCIF, but is universal.

Axis category: define arbitrary axes, which are then referred to when describing the detector geometry. They can depend on one another, e.g. a goniometer with stacked motors, and they can be used to describe curved or spiral pixel layouts.

The imgCIF model

An experiment consists of a series of scans. Each scan consists of a series of frames, with one or more axes changing position between frame measurement. A single “frame” is composed of data from one or more “arrays”. These data are included in encoded strings (imgCIF) or as binary (imgCBF). A single data block “should” include data from the entire experiment.

Limitations of CIF for images

An unordered format is not suitable for storing MB or GB of image data (slow access)

imgCBF format: one file = one frame, image in binary

Alternatives to CIF format for raw data

Large-scale neutron and X-ray facilities tending towards NeXus (=HDF5 + “application definitions” $\approx$ dictionaries). COMCIFS and NIAC working together to harmonise definitions.

See `nexusformat.org`

Remember also: a CIF dictionary is (should be) format-agnostic so can be created to describe the contents of a NeXus file.

Topology

Classify a continuous structure in terms of the network of connected nodes. Nodes can be atoms or groups (e.g. carbonate). Many different compounds can have identical topology. Distinct topologies are named and classified in various databases. Work ongoing in the area of classification of interpenetrating topologies.

Limitations of the core CIF dictionary

Assumes a single crystal is measured once at a single wavelength and temperature to produce a structure refined against integrated peak intensities and described by a single space group and list of atoms.

Focuses on final data at time of journal submission.

Powder diffraction

A powder diffraction experiment:

- Often measures from samples composed of more than one compound (“phase”)
- Results in a “diffractogram” of intensity vs 2θ (or d)
- Fits the entire diffractogram (peaks and background)
- Involves a variety of geometries
- Routinely uses both X-ray and neutron data in joint refinements

The powder diffraction dictionary “pdCIF”

All names begin with `_pd_`

Allows recording and transfer of reduced data, before any fitting results

The `<category>.<object>` naming scheme is not always followed, to allow the diffractogram at each stage of correction to be tabulated together with/without calculated intensities

Compare core CIF `refln` and `diffrn_refln` - separate categories

The pdCIF block pointer mechanism

A unique block pointer:

`2019-09-02T09:20|Naples|S357_phase1|J_R_Hester`

Used to link data blocks containing relevant information:

- Structure to experimental information (`_pd_block_diffractogram_id`)
- A data set to a constituent compound (`_pd_phase_block_id`)
- A data set to a block containing a calibration (`_pd_calib_std_external_block_id`)

A way to overcome the limitations of the original CIF: can now have multiple data sets, multiple constituent compounds, and even calibration information.

Modulated and composite structures

Some structures are best described as either:

- 2 or more interpenetrating structures
- a parent structure perturbed periodically in one or more dimensions

Both descriptions can be possible for a single compound.

So a result can contain multiple cells for a single experimental measurement.

Describing modulated structures

A higher-dimensional cell requires expanding category keys wherever h,k,l are listed:

`diffrn_refln`, `refln_index`

Describing composite structures

The structural description lists component cells and the modulation waves.

`_cell_subsystem_code`: a code for the particular component

`_cell_subsystem_matrix`: defines the relationships between components

`_space_group_ssg`: superspace group descriptions

Multiple blocks

One is the parent block, one (modulated structure) or more (composite structures) subsystems. A block pointer is specially encoded to explain the relationships, using the pdCIF approach and appending:

_REFRNCE(_code) Unmodulated structure information, common information. Relevant cell may be appended

_MOD(_code) Modulation-specific details

```
_audit_block_code      1997-07-24|LaSNbS2|G.M.|
loop_
_audit_link_block_code
_audit_link_block_description
    .   'common experimental and publication data'
    1997-07-24|LaSNbS2|G.M.|_REFRNCE
        'reference structure (global data)'
    1997-07-21|LaSNbS2|G.M.|_MOD
        'modulated structure (global data)'
    1997-07-24|LaSNbS2|G.M.|_REFRNCE_NbS2
        'reference structure (1st subsystem)'
    1997-07-21|LaSNbS2|G.M.|_MOD_NbS2
        'modulated structure (1st subsystem)'
    1997-07-24|LaSNbS2|G.M.|_REFRNCE_LaS
        'reference structure (2nd subsystem)'
    1997-07-21|LaSNbS2|G.M.|_MOD_LaS
        'modulated structure (2nd subsystem)'
```

Is this possible?

Reporting results of a powder diffraction experiment on a modulated structure

Reporting the constraints used in the refinement of the modulated structure