# Lost in translation – using CIF dictionaries for semantic data validation

Vilnius University

Antanas Vaitkus[1], Algirdas Grybauskas[1], Andrius Merkys[1], Saulius Gražulis[1]

[1]Vilnius University, Life Sciences Center, Institute of Biotechnology, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania

Email: antanas.vaitkus@bti.vu.lt

## Crystallographic Information Framework (CIF)

The Crystallographic Information Framework encompasses:

- Multiple data formats including `CIF 1.1` [1] and `CIF 2.0` [2].
- Three dictionary definition languages: `DDL1`, `DDL2` and `DDLm` [3].
- Numerous COMCIFS-approved and private dictionaries.

## What data discrepancies can a generic CIF validator detect?

```
#\#CIF_2.0
data_CIF
# ...

loop_
_audit_conform.dict_name
_audit_conform.dict_version
_audit_conform.dict_location
CIF_CORE       3.2.0       https://doi.org/10.1107/cifdic_000002

_exptl_crystal.colour                  [ opaque colourless whitish ]

_refine_ls.R_factor_all                0.030(17)

_cell.volume                           194.14(7)
_cell.volume_su                        0.3

_symmetry.cell_setting                 monoclinic

_reflns.threshold_expression           'I > 2\s(I)'
_reflns.observed_criterion             'I > 3\s(I)'

loop_
_space_group.IT_number                 4  11  104

_space_group.name_Haal                 '-P 2yb'

loop_
_atom_site.label
_atom_site.type_symbol
_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
_atom_site.calc_flag
C1   C    0.0347(8)  0.25  0.7405(7)   0.0204(8)  Uani  1  d
C2   C   -0.1098(9)   1/4  0.8805(7)   0.0217(9)  Uani  1  d
C3   C   -0.2811(8)  0.25  1.0390(7)   0.0229(9)  Uanj  1  d
N4   N   -0.4234(9)  0.25  1.1721(7)  -0.0314(9)  Uani  1  d
Br1  Br   0.2471(1)  0.25  0.5278(1)   0.0199(2)  Uani  1  d

loop_
_geom_contact.atom_site_label_1
_geom_contact.atom_site_label_2
_geom_contact.site_symmetry_1
_geom_contact.site_symmetry_2
_geom_contact.distance
_geom_bond.multiplicity
C1   C1    .   2_455  4.283(5)  1
C4   Br1   .       .  1.788(4)  1
C2   C3    .   1_556  1.378(6)  1
C2   N4    .   1_556  2.530(4)  2
C3   N4    .   1_556  2.723(5)  1
C2   N4    .   1_556  3.890(4)  0
# ...
```

- Dictionary conformance declaration
- Incorrect data container
- Inappropriate s.u. value
- Out-of-range value
- Missing parent data item value
- Inappropriate loop item
- Aliases with differing values
- Incorrect data type
- Invalid enumeration value
- Mismatching s.u. values
- Non-unique category key
- Incorrect category loop
- Deprecated data item
- Unrecognised data name

## Crystallography Open Database (COD)

https://www.crystallography.net/cod

- Open-access FAIR [4] repository of small molecule crystal structures.
- Data can be reused without any additional restrictions (CC0 license).
- Covers organic, inorganic, organometallic compounds and minerals.
- More than 515 000 entries and growing.
- Heavily utilises the CIF framework.

## The `cod-tools` software package

The open-source `cod-tools` [5] software package includes a variety of tools for handling CIF files and dictionaries such as:

- **COD::CIF::Parser** [6]. An error-correcting CIF parser that supports both `CIF 1.1` and `CIF 2.0` data files.
- **DDLm validator** [7]. A general purpose validator that supports dictionaries written either in DDL1 or DDLm.
- **DDL development tools.** A set of tools for comparing, migrating and checking DDL1 and DDLm dictionaries. Some of these tools are used in the official IUCr dictionary development repositories.

Usage examples:

- Validate a CIF file against a DDLm dictionary:

```
cif_validate --ddlm-add-dictionary cif_core.dic 1000000.cif
```

- Check a DDLm dictionary against a set of best practices:

```
cif_ddlm_dic_check cif_core.dic
```

## Validation issue database

CIF validation messages collected from the COD are stored in a publicly available relational database that can be accessed:

- By using a `RestfulDB` GUI at https://sql.crystallography.net/db/cod_validation.
- By connecting directly using a `MySQL` client:

```
mysql -u cod_reader -h sql.crystallography.net cod_validation
```

## Conclusions

- CIF dictionaries are crucial in ensuring crystallographic data quality.
- The CIF framework is still underused by the crystallographic community.
- The `cod-tools` package provides tools for CIF validation.

## References

[1] Hall et al. *Acta Crystallographica, Section A*, 47(6):655–685, 1991. https://doi.org/10.1107/S010876739101067X.

[2] Bernstein et al. *Journal of Applied Crystallography*, 49(1):277–284, 2016. https://doi.org/10.1107/S1600576715021871.

[3] Spadaccini et al. *Journal of Chemical Information and Modeling*, 52(8):1907–1916, 2012. https://doi.org/10.1021/ci300075z.

[4] Wilkinson et al. *Scientific Data*, 3(1), 2016. https://doi.org/10.1038/sdata.2016.18.

[5] Vaitkus et al. cod-tools, version 3.10.0, 2024. svn://www.crystallography.net/cod-tools/tags/v3.10.0.

[6] Merkys et al. *Journal of Applied Crystallography*, 49(1):292–301, 2016. https://doi.org/10.1107/S1600576715022396.

[7] Vaitkus et al. *Journal of Applied Crystallography*, 54(2):661–672, 2021. https://doi.org/10.1107/S1600576720016532.

## Acknowledgements

This research has received funding from the Research Council of Lithuania under grant agreement No. MIP-23-87.

Online version of the poster: https://bit.ly/4cvDbwf