Poster

Lost in translation – using CIF dictionaries for semantic data validation

A. Vaitkus[1], A. Merkys[1], S. Gražulis[1]

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

antanas.vaitkus@bti.vu.lt

Over the last 30 years, the Crystallographic Information Framework (CIF) has become one of the main means for recording and exchanging crystallographic data. As such, it has been widely adopted by crystallographers, scientific journals and crystallographic databases alike. However, a recent analysis of crystallographic data collected from peer-reviewed sources [1] suggests that the framework might still be underutilised, especially in regard to the adoption of the relatively new features like the CIF 2.0 [2] data format and DDLm [3] dictionaries. While it is possible to use the CIF framework as a black box by fully dedicating the creation and interpretation of the underlying files to third-party software, having a more in-depth understanding of CIF dictionaries and their relationship to CIF data files opens up new possibilities such as describing custom data fields in a formal standardised way that is both human- and machine-readable. Due to this, our team at the Crystallography Open Database (COD [4], https://www.crystallography.net/cod/) has developed a set of open resources aimed at easing main CIF-related tasks as well as identifying and elucidating the most common semantic mistakes encountered in CIF data files.

One of these resources is the open-source *cod-tools* software package [5] that contains a variety of tools designed for working within the CIF framework. Functionality provided by this software suite includes parsing CIF files and validating them against CIF dictionaries, translating dictionaries between the now-deprecated DDL1 and the modern DDLm dictionary definition language, and checking dictionaries against a set of dictionary development best practices. These tools have proven useful both for large-scale applications such as the data quality assurance in the COD and the automated post-commit checks in the official IUCr dictionary repositories, as well as for individual researchers aiming to identify and correct semantic issues in their data.

Another useful resource is the publicly available database of validation issues (https://sql.crystallography.net/db/cod_validation/) detected by validating the entire COD database against a set of standard IUCr and COD CIF dictionaries [6,7]. Since data in the COD are collected from various peer-reviewed sources, this derivative database serves as a reasonable reflection of the CIF validation practices adopted by the wider crystallographic community. Furthermore, since the data is checked against both the DDL1 and the DDLm dictionaries, the results allow to more easily detect incompatibilities that may arise when migrating from the one version to the other. In fact, findings of an earlier analysis [1] have already led to a series of fixes to the IUCr DDLm dictionaries aimed at improving backwards compatibility. Finally, software developers might also find these validation results interesting since certain issues seem to be routinely generated by the same pieces of software, most likely due to a slight coding errors or misinterpretation of data item definitions. Currently, the database contains over 10 million issues collected from more than 510 thousand CIF files.

The CIF framework is still very much alive with new features and dictionary updates being released on a regular basis. While the CIF file format seems to be ubiquitous these days with many programs even implementing their own built-in parsers, the semantic side of the framework is often neglected. It is completely understandable that development of a fully-fledged CIF dictionary might not be within the scope of every research group, however, even the less time-consuming changes such as the introduction of an additional semantic validation step to one's current workflow, may greatly improve the rate of early error detection and aid in ensuring that the data continues to be interpreted as originally intended by the authors.

[1] Vaitkus, A., Merkys, A. & Gražulis, S., (2021). *J. Appl. Crystallogr.,* **54**(2), 661-672.

[2] Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S. Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. & Westrip, Simon P., (2016). *J. Appl. Crystallogr.*, **49**(1), 277-284.

[3] Spadaccini, N. & Hall, S. R. (2012). *J. Chem. Inf. Model.*, **52**(8), 1907-1916.

[4] Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. & Le Bail, A., (2012). *Nucleic Acids Res.*, **40**, D420-D427.

[5] Vaitkus, A., Merkys, A. & Gražulis, S., (2024). cod-tools, version 3.9.0. URL: svn://www.crystallography.net/cod-tools/tags/v3.9.0

[6] IUCr. IUCr CIF dictionaries. https://www.iucr.org/resources/cif/dictionaries/. [Last accessed: 2024-05-07]

[7] Vaitkus, A., Merkys, A. & Gražulis, S. COD CIF dictionaries. https://www.crystallography.net/cod/cif/dictionaries/. [Last accessed: 2024-05-07]

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