Improvements to the data search and validation functionality in the Crystallography Open Database

A. Vaitkus[1], A. Merkys[1], A. Grybauskas[1], A. Konovalovas[1], M. Quirós Olozábal[2], S. Gražulis[1,3]

[1]Vilnius University, Life Sciences Center, Institute of Biotechnology, Saulėtekio 7, LT-10257 Vilnius, Lithuania, [2]Departamento de Química Inorgánica, Universidad de Granada, 18071, Granada, Spain, [3]Vilnius University, Faculty of Mathematics and Informatics, Naugarduko 24, LT-03225 Vilnius, Lithuania

antanas.vaitkus@bti.vu.lt

Crystallography Open Database (COD) [1] is the largest open-access FAIR [2] collection of small-molecule crystal structures that currently contains over 475 000 entries. In recent years, several notable improvements have been made to enhance the data curation process as well as expand the data search capabilities.

Data curation tasks of the COD heavily rely on the Crystallographic Information Framework (CIF) therefore recent CIF-related IUCr innovations stipulated the appropriate changes to the COD software. The F/LOSS cod-tools software package was updated to support the CIF2 data format [3] and the DDLm [4] dictionary language thus enabling the routine formal validation of all COD CIF files against the latest generation of CIF dictionaries [5]. The collected validation results were compiled in a publicly available CIF validation issue database that has already proven useful in data maintenance and ontology development tasks. A set of programs intended to aid in the dictionary migration from the now deprecated DDL1 language to the novel DDLm language was also created.

Effective search is another aspect of the COD database that has been greatly improved. Efficient chemical structure search in a crystallographic database requires that certain properties of the crystallised materials, such as molecular connectivity and other chemical features, be described in a machine-readable way. However, completely automated derivation of such information from CIF files is difficult and often provides suboptimal results. With this in mind, a set of high-quality manually curated SMILES that cover more than 40% of all COD entries have been made publicly available and can be used for chemical substructure search in the COD or for any other purpose on an open-access basis. The conventions that have been followed to represent various types of compounds as well as description of the semi-automatic SMILES derivation pipeline have also been extensively described [6] to improve the reusability and reproducibility of the data.

The COD data search capabilities were even further enhanced by implementing the OPTIMADE application programming interface (API) [7, 8] that aims to improve the interoperability between materials databases. It is extremely beneficial to be able to access information from multiple materials databases as they often differ in fidelity and focus across material classes and properties. However, retrieving data from multiple databases is difficult as each database has its own specific API. Moreover, as the APIs of individual databases inevitably evolve, existing clients must also evolve and are required to translate the responses from the new API to the internal representation of the client, which can require significant effort. The OPTIMADE consortium aims to alleviate most of these problems by providing a common RESTful API based on the JSON:API specification [9].

These recent changes to the COD are aimed at improving the data quality assurance process as well as ensuring that the data remain open, FAIR and readily available for a diverse range of applications in fields such as cheminformatics and materials science.

[1] Gražulis, Saulius et al. (2012). *Nucleic Acids Res*. **40**, D420. doi: 10.1093/nar/gkr900

[2] Wilkinson, Mark D. et al. (2016). *Sci. Data*. **3**. doi: 10.1038/sdata.2016.18

[3] Bernstein, Herbert J. et al. (2016). *J. Appl. Crystallogr*. **49**, 277. doi: 10.1107/S1600576715021871

[4] Spadaccini, N. & Hall, S. R. (2012). *J. Chem. Inf. Model*. **52**, 1907. doi: 10.1021/ci300075z

[5] Vaitkus, Antanas et. al. (2021). *J. Appl. Crystallogr*. **50**, 661. doi: 10.1107/S1600576720016532

[6] Quirós, Miguel et. al. (2018). *J. Cheminformatics*. **10**. doi: 10.1186/s13321-018-0279-6

[7] Andersen, Casper W. et al. (2020). *The OPTIMADE Specification*. doi: 10.5281/zenodo.4195050

[8] Andersen, Casper W. et al. (2021). *OPTIMADE: an API for exchanging materials data*. url: https://arxiv.org/abs/2103.02068

[9] *JSON:API v1.0*. url: https://jsonapi.org/format/1.0/

Keywords: COD; database; DDLm; SMILES; OPTIMADE

This research has received funding from the Research Council of Lithuania under grant agreement No. MIP-20-21.