The use of interconnected open data for material identification

Antanas Vaitkus[a], Andrius Merkys[a], Yassine El Mendili[b] and Saulius Gražulis[a]

[a]Vilnius University Institute of Biotechnology, Saulėtekio av. 7, LT-10257 Vilnius, Lithuania
[b]Normandie Université, CRISMAT-ENSICAEN, UMR CNRS 6508, Université de Caen Normandie, 14050 Caen, France

Introduction

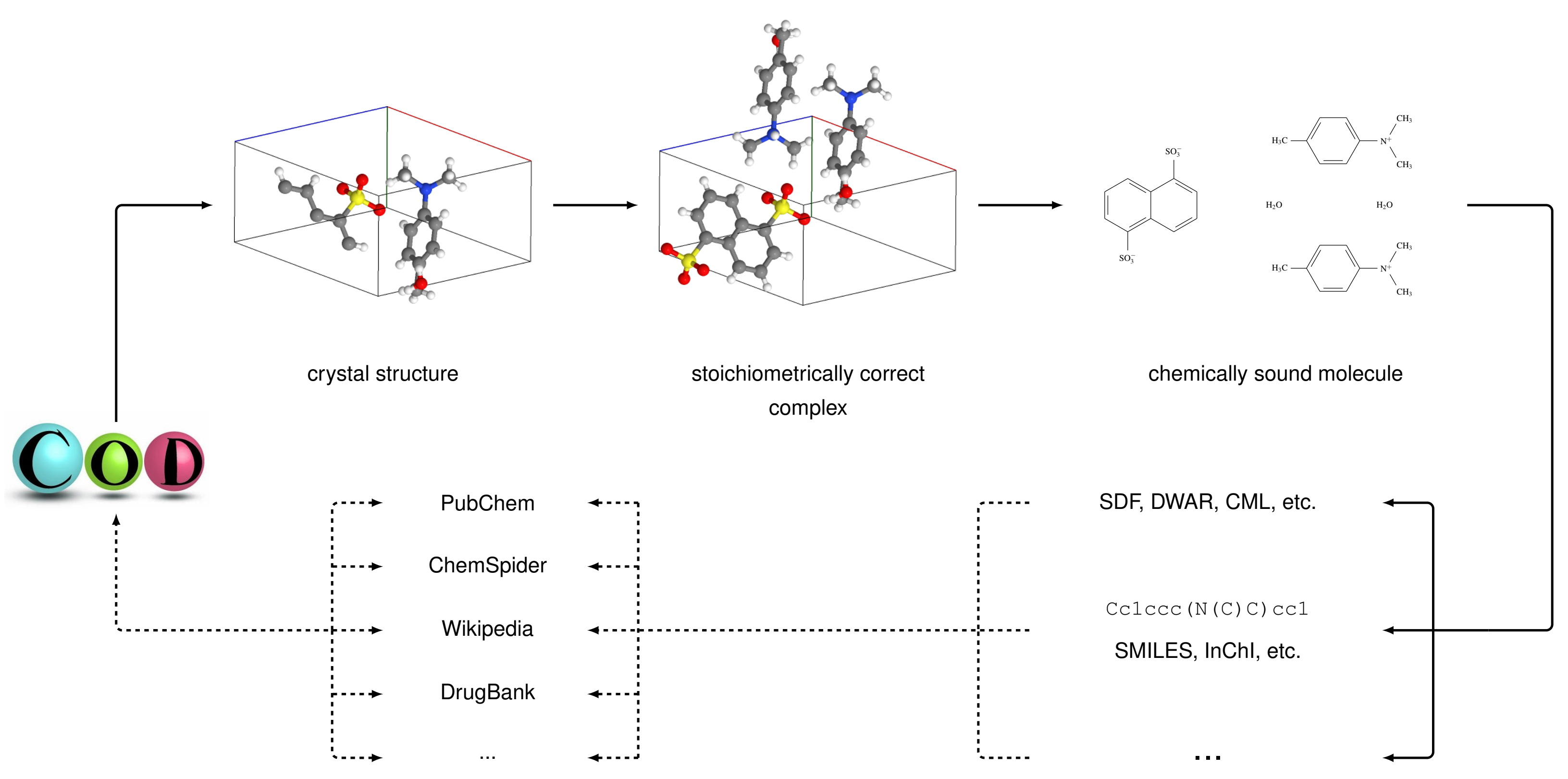
One of the main driving forces behind modern day scientific research is openness. As a result, open-access data repositories play an increasingly important role in the scientific community. The Crystallography Open Database (COD, `http://www.crystallography.net/cod`) [1] is one such resource – over the last 15 years it has become the largest curated and validated open-access collection of inorganic and non-polymeric organic crystal structures encompassing over 390 000 entries. More than 160 000 of these entries have been enhanced by manually adding the SMILES descriptors and as a result enabling the substructure search within the given subset. Recently, a number of computer programs capable of automatically determining stoichiometrically [2] and chemically sound molecules from the crystallographic data have also been developed; this, in turn, enabled the automated generation of structural formulae descriptors and eased the establishment of cross-links between the COD and other open-access resources such as PubMed, DrugBank and Wikipedia. New strides have also been made in relating spectral data to their corresponding crystal structures. The COD was chosen as the back-end database in the wide scale on-site sample analysis of the "Sonic Drilling coupled with Automated Mineralogy and chemistry On-Line-On-Mine-Real-Time" (SOLSA, `http://www.solsa-mining.eu`) project that focuses on developing highly efficient, cost-effective and sustainable exploration technologies. Since part of the sample analysis involves material identification via the means of Raman spectroscopy, reference spectra aggregation from various sources was initialised choosing CIF [3, 4] as the homogeneous data carrier format for both XRD and spectral data; this, in turn, stipulated the development of spectroscopy oriented DDLm dictionary [5] and the creation of the Raman Open Database (ROD, `http://solsa.crystallography.net/rod`). These new developments will allow the SOLSA project to present various aspects of mineral characterization such as Raman spectra, XRD structures and fluorescence data in the COD database in a uniform, computer-readable way.

Crystallography Open Database (COD)

`http://www.crystallography.net/cod`

- Open-access;
- Contains small-molecule organic, inorganic, and metal-organic crystal structures;
- Uses CIF [3] as the carrier format;
- Over 390 000 entries.

Chemical information extraction


crystal structure

stoichiometrically correct complex

chemically sound molecule

PubChem

ChemSpider

Wikipedia

DrugBank

...

SDF, DWAR, CML, etc.

`Cc1ccc(N(C)C)cc1`

SMILES, InChI, etc.

...

Chemical information in the COD

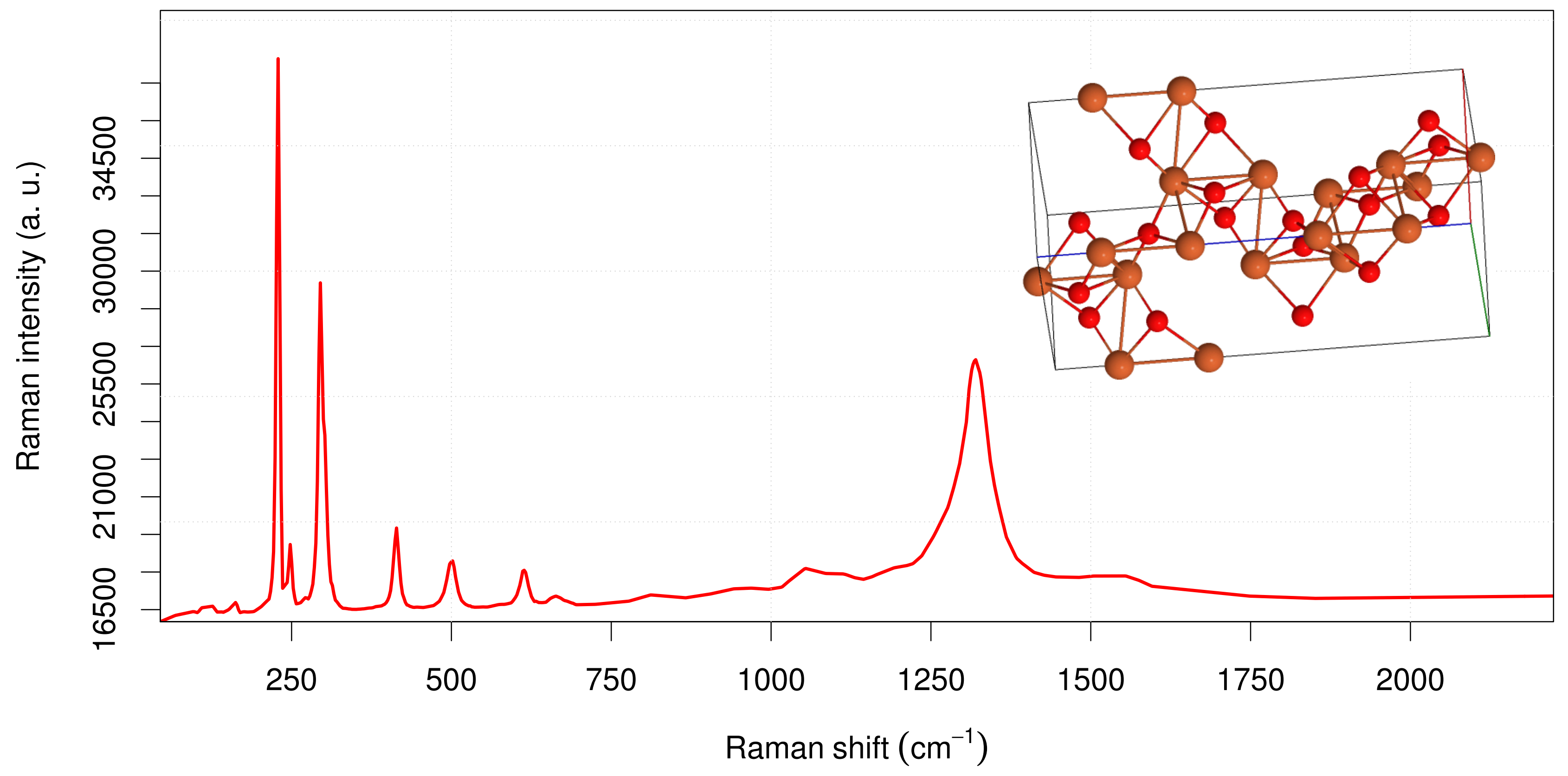
- Periodically generated from crystallographic data in an automated way using open-source software [6, 2, 7];
- Enables a more efficient substructure search;
- Used to establish cross-links to other resources;
- Available in its entirety as a DataWarrior [8] file at `http://www.crystallography.net/dwar`

Raman Open Database (ROD)

`http://solsa.crystallography.net/rod`

- Open-access;
- Contains Raman spectroscopy data;
- Uses CIF2 [4] as the carrier format;
- Validates input files upon deposition;
- Spectral data is cross-linked with the XRD data in the COD;
- Used in the SOLSA project for material identification.

COD on a ROD



The Raman spectra and the related crystal structure of hematite. ROD ID 1000001, COD ID 1546383.

Raman Spectroscopy Ontology

- Developed and maintained by an international team of Raman spectroscopy experts;
- Expressed as a DDLm [5] conforming CIF dictionary;
- Latest version available at `http://solsa.crystallography.net/rod/cif/dictionaries/cif_raman.dic`

Conclusions

- The COD is now enhanced with chemical data and interconnected with other open access resources;
- There is a need for a curated set of Raman spectroscopy data;
- CIF2 is suitable format for storing scientific data of all sorts;
- Open access data repositories are a viable alternative to proprietary databases.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 689868.

References

[1] S. Gražulis et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40(D1):D420–D427, 2012.

[2] S. Gražulis et al. Computing stoichiometric molecular composition from crystal structures. *Journal of Applied Crystallography*, 48:85–91, 2015.

[3] S. R. Hall et al. The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47(6):655–685, 1991.

[4] H. J. Bernstein et al. Specification of the Crystallographic Information File format, version 2.0. *Journal of Applied Crystallography*, 49(1):277–284, 2016.

[5] N. Spadaccini et al. DDLm: A new dictionary definition language. *Journal of Chemical Information and Modeling*, 52(8):1907–1916, 2012.

[6] A. Merkys et al. COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *Journal of Applied Crystallography*, 49(1):292–301, 2016.

[7] OpenChemLib. Open source Java-based chemistry library. `https://github.com/Actelion/openchemlib`.

[8] T. Sander et al. DataWarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling*, 55(2):460–473, 2015.

Antanas Vaitkus has no conflict of interest.
Andrius Merkys has no conflict of interest.
Yassine El Mendili has no conflict of interest.
Saulius Gražulis has no conflict of interest.

On-line version of the poster:
`http://j.mp/2vXMfXP`