Poster Presentation

MS107.P01

The use of interconnected open data for material identification

Antanas Vaitkus[1], Andrius Merkys[1], Saulius Gražulis[1]
[1]*Vilnius University Institute Of Biotechnology, Vilnius, Lithuania*
E-mail: antanas.vaitkus90@gmail.com

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 13 years it has become the largest curated and validated open-access collection of inorganic and non-polymeric organic crystal structures encompassing over 375 000 entries. More than 135 000 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 carried out choosing CIF as the homogeneous data carrier format for both XRD and spectral data; this, in turn, stipulated the development of the spectroscopy oriented CIF dictionary. These new developments of the CIF dictionaries 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.

[1] Gražulis, S. et al. (2012). Nucleic Acids Research, 40, D420–D427.
[2] Gražulis, S. et al. (2015). Journal of Applied Crystallography, 48, 85-91.

Keywords: COD, Spectral data, SOLSA