Poster Presentation

CC.P25

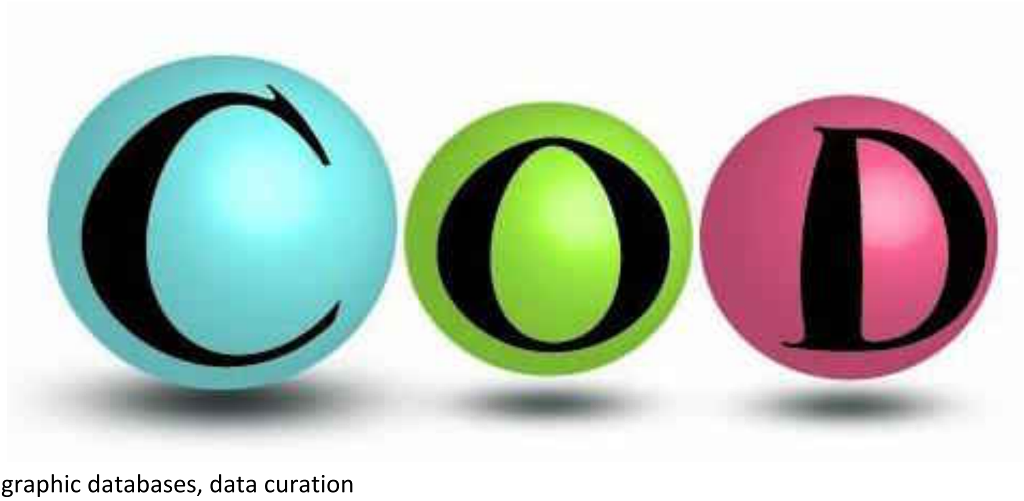
Chemical information presentation in the Crystallography Open Database

A. Merkys[1], A. Matusevičiūtė[2], A. Vaitkus[2,3], A. Le Bail[4], D. Chateigner[5], L. Lutterotti[6], M. Quirós-Olozábal[7], M. Okulič-Kazarinas[1,8], P. Moeck[9], P. Murray-Rust[10], R. Downs[11], S. Girdzijauskaitė[3], S. Gražulis[1,2]

[1]Vilnius University, Institute of Biotechnology, Department of Protein - DNA Interactions, Vilnius, Lithuania, [2]Vilnius University, Faculty of Mathematics and Informatics, Department of Mathematical Computer Science, Vilnius, Lithuania, [3]Vilnius University, Faculty of Mathematics and Informatics, Department of Software Engineering, Vilnius, Lithuania, [4]Universite du Maine, Laboratoire des Oxydes et Fluorures - CNRS UMR 6010, Faculté des Sciences, Le Mans, France, [5]Universite de Caen-Basse Normandie, CRISMAT-ENSICAEN, Caen, France, [6]University of Trento, Department of Materials Engineering, Trento, Italy, [7]Universidad de Granada, Facultad de Ciencias, Departamento de Quımica Inorganica, Granada, Spain, [8]Mykolas Romeris University, Faculty of Social Technologies, Department of Informatics and Software Systems, Vilnius, Lithuania, [9]Portland State University, Department of Physics, Portland, USA, [10]University of Cambridge, Department of Chemistry, Cambridge, United Kingdom, [11]University of Arizona, Department of Geosciences, Tucson, USA

Crystallography Open Database (COD, http://www.crystallography.net/) is the largest to date curated open-access collection of small to medium sized unit cell crystal structures [1,2]. Over 11 years of development, COD has accumulated over 1/4 million structures from the peer reviewed press and personal communications. COD has an automated data submission Web site, performs routine automatic quality checks on all incoming structures and is now recommended as a database for crystallographic deposition by several scientific journals. To facilitate automatic use and discoverability of COD data, and to increase usefulness of our database for chemists, two steps were undertaken. COD was now supplemented with software and data from the CrystalEye data aggregator. The new software permits extracting chemical data and presenting them as structural formula, unique moieties, and chemically significant fragments. We have also implemented search of crystal structures by the structural chemical formulae of the target compounds. The search is first of all performed among 70 000 hand-curated chemical structure descriptors, and can be extended to automatically generated descriptors. To facilitate data curation, a new software platform for data review is being developed. All COD structures will be evaluated using statistical distributions of observed geometrical and chemical properties (bond lengths, angles, dihedrals, planarities). The most statistically unusual structures will be forwarded to a COD reviewer Internet forum, where qualified reviewers will be asked whether they find provided evidence for a particular structure convincing or not. In this way, a set of human review indicators (convincing/unconvincing) will be available along with the match against the bulk of data (usual structure/unusual). Such indicators would be especially useful for deciding which COD records require special attention and which subsets of COD should be selected for reliable scientific inferences.

***[1]** Gražulis, S., Daškevič, A., Merkys, A., et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Research, 2012, 40, D420-D427, **[2]** Gražulis, S., Chateigner, D., Downs, R. T., et al., A. Crystallography Open Database - an open-access collection of crystal structures. Journal of Applied Crystallography, 2009, 42, 726-729*



Keywords: crystallographic databases, data curation