SciDataCon 2016

Conference paper

The Crystallography Open Database – new perspectives.

Summary

Today's connected world crucially depends on the open availability of data on the Internet. Great success of Wikipedia, PDB (Berman et al. 2012) or open sequence databases like UniProt ({The UniProt Consortium} 2015) demonstrate the power of data sharing when it is unhindered by paywalls and copy restrictions.

The Crystallography Open Database (COD) builds on the experience of the open databases and harnesses the power of the community to build an openly-available chemical crystallography database on the net. The COD ingests data in CIF format, validates it according to IUCr dictionaries and quality criteria, and offers consolidated data to COD users again in a standard CIF format. Currently, the COD contains over 360 thousand records, covering the year span from 1915 to present.

The COD story

Numerous people contributed to the COD (http://www.crystallography.net), and the COD now contains data from several databases such as the AMCSD (Rajan et al. 2006), CrystalEye (Day et al. 2012) or the PCOD (Le Bail 2005). Most structures published in electronic format are represented in the COD, provided they were available on the Internet freely or were donated by authors or their institutions. Some prominent structures published in paper form were also digitalized and included into the COD.

The COD database stores experimental structures of organic, metal organic and inorganic compounds and minerals. The COD collects the result of all types of crystallographic diffraction experiments (X-rays, electrons, or neutrons diffracted from single crystals and powders). In recent years, however, quantum mechanics computations using DFT and other methods became powerful enough to yield reliable structural descriptions, either *ab initio* or in conjunction with experimental crystallographic techniques. Such structures are also collected and are stored in a sister database, the TCOD (Theoretical Crystallography Open Database). Since the COD, TCOD and PCOD all use the same CIF framework for data representation, all databases can be searched and processed using the same software tools.

To make the COD simple but efficient several ingredients are crucial. Free/Libre Open Source software (F/LOSS) is at the core of COD development, making it easy to reuse COD data and algorithms. An open data standard – the Crystallographic Interchange Framework – and its ontologies maintained by the IUCr are essential for efficient data exchange (Hall et al. 1991; Bernstein et al. 2016).

The COD maintains stable REST interfaces since its inception (Gražulis et al. 2009; Gražulis et al. 2012). You know, cool URIs do not change! (Berners-Lee 1998). As a result, the COD is ideally suitable for the 21st century's open, connected world, providing stable links to crystal structures on the Web. The stable setup allows instantaneous reuse of COD data in various

sites, and provides a mechanism for data citation using stable, unique COD numbers for every structure.

The COD contains simple search and retrieve Web interface as demonstrated in **Figure 1**. In addition to web search, the COD offers download of all data using various protocols (http, svn, rsync) and querying the database directly using a MySQL client. This gives maximum flexibility in cases where the Web interface is not enough, and permits integration of the COD into other websites or standalone programs.

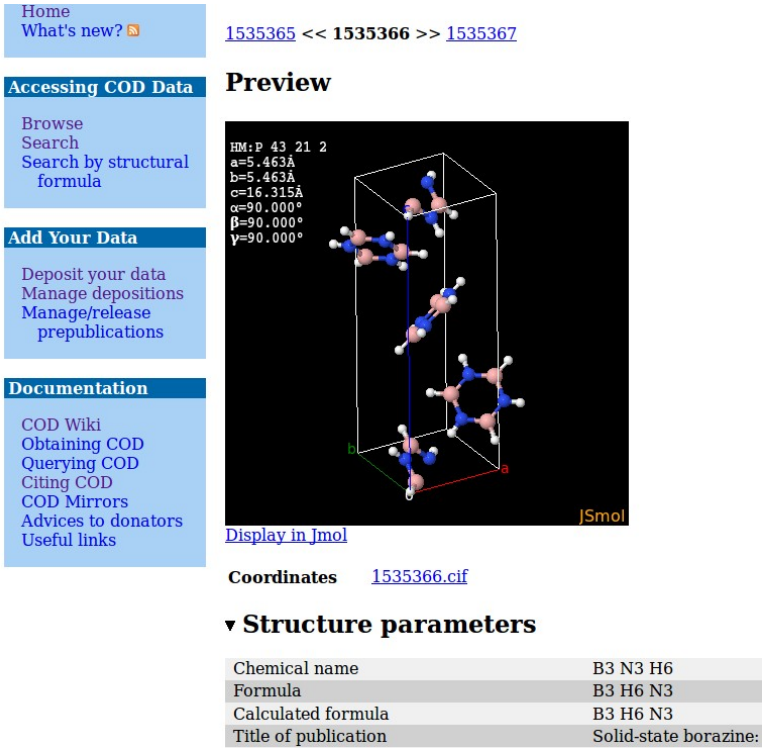

Figure 1: The COD search result interface and structure view page.

The primary information in the COD is atomic coordinates and crystal descriptions. Moreover, the COD systematically stores Fobs and powder trace data when these are available. COD data entries are manually curated as well as automatically checked. From the ingested data, the COD formulates a new CIF file which is guaranteed to be syntactically correct and contains all original data with attached metadata. Should a change of the COD entry be needed for data curation, the COD meticulously stores all changes in a version control database (currently Subversion), thus providing full traceability and data provenance. With the implementation of these procedures, the COD recently ranked 5 in the Thomson Reuters Databases citation index, and is now recommended by the Nature Publishing Group for all structure depositions.

COD applications

The COD enjoys growing field of applications. Teaching (Gražulis et al. 2015), Powder identification (Lutterotti et al. 2015), source of data for computational material science (First and Floudas 2013; Pizzi et al. 2016) are among the latest areas where the COD proved useful as a source of data. Moreover, the open nature of the COD allowed to match structure descriptions against material properties (Pepponi et al. 2012) yielding the first open collection of material properties and property-structure relations.

In all cases, we benefited from the exisiting crystallographic CIF dictionaries from the IUCr to describe crystallographic entities. We also found it very useful to make a domain-specific CIF dictionaries to describe material properties, computational settings and structural metadata. In this way we could reuse the existing CIF dictionaries without duplication, and could reuse our existing software and database structure for new fields of enquiry.

COD new directions

The COD continuity and large collection of crystal data allows us to use it as a data source and management platform for industrial processes: in the recently started SOLSA project, the COD will provide crystal data for mineral identification in mines in real time. Also, the COD will ingest public data that is obtained during such identification runs, further enriching the COD collection and increasing identification precision.

In the year to come, the COD will proceed to store diffraction image data. Storing raw diffraction images currently poses a challenge to meet reasonable management and storage costs, along with enough bandwidth for image set distribution. Here again the COD pursues community backed, distributed and open-source solution. COD starts to use a Tahoe-LAFS storage engine (Wilcox-O'Hearn and Warner 2008), which is currently capable to provide affordable redundant storage up to the order of petabytes.

Acknowledgements

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

Competing Interests

The authors declare that they have no competing interests.

References

Berman, H; Kleywegt, G; Nakamura, H and **Markley, J** 2012 The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure,* 20: 391-396. DOI: 10.1016/j.str.2012.01.010

Berners-Lee, T 1998 *Cool URIs don't change*. Available at https://www.w3.org/Provider/Style/URI.html [Last accessed 2016-05-29]

Bernstein, H J; Bollinger, J C; Brown, I D; Gražulis, S; Hester, J R; McMahon, B; Spadaccini, N; Westbrook, J D and **Westrip, S P** 2016 Specification of the Crystallographic Information File format, version 2.0. *Journal of Applied Crystallography,* 49: 277-284. DOI: 10.1107/s1600576715021871

Day, N; Downing, J; Adams, S; England, N W and **Murray-Rust, P.** 2012 CrystalEye: automated aggregation, semantification and dissemination of the world's open crystallographic data. *Journal of Applied Crystallography,* 45: 316-323. DOI: 10.1107/S0021889812006462

First, E L and **Floudas, C A** 2013 MOFomics: Computational pore characterization of metal–organic frameworks. *Microporous and Mesoporous Materials,* 165: 32-39. DOI: 10.1016/j.micromeso.2012.07.049

Gražulis, S; Chateigner, D; Downs, R T; Yokochi, A F T; Quirós, M; Lutterotti, L; Manakova, E; Butkus. J; Moeck, P and **Le Bail, A** 2009 Crystallography Open Database -- an open-access collection of crystal structures. *Journal of Applied Crystallography,* 42: 726-729. DOI: 10.1107/S0021889809016690

Gražulis, S; Daškevič, A; Merkys, A; Chateigner, D; Lutterotti, L; Quirós, M; Serebryanaya, N R; Moeck, P; Downs, R T and **Le Bail, A** 2012 Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research,* 40: D420-D427. DOI: 10.1093/nar/gkr900

Gražulis, S; Sarjeant, A A; Moeck, P; Stone-Sundberg, J; Snyder, T J; Kaminsky, W; Oliver, A G; Stern, C L; Dawe, L N; Rychkov, D A; Losev, E A; Boldyreva, E V; Tanski, J M; Bernstein, J; Rabeh, W M and **Kantardjieff, K A** 2015 Crystallographic education in the 21st century. *Journal of Applied Crystallography,* 48: 1964-1975. DOI: 10.1107/S1600576715016830

Hall, S R; Allen, F H and **Brown, I D** 1991 The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A,* 47: 655-685. DOI: 10.1107/S010876739101067X

Le Bail, A 2005 Inorganic structure prediction with it GRINSP. *Journal of Applied Crystallography,* 38: 389-395. DOI: 10.1107/S0021889805002384

Lutterotti, L; Chateigner, D; Pillière, H and **Fontugne, C** 2015 *Full-pattern search-match using the Crystallography Open Database: an Internet tool*. Available at http://www.ecole.ensicaen.fr/~chateign/danielc/abstracts/Lutterotti_abstract_RXMatiere2013_FPSM.pdf [Last accessed 2016-05-29]

Pepponi, G; Gražulis, S and **Chateigner, D** 2012 MPOD: A Material Property Open Database linked to structural information. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms,* 284: 10-14. DOI: 10.1016/j.nimb.2011.08.070

Pizzi, G; Cepellotti, A; Sabatini, R; Marzari, N and **Kozinsky, B** 2016 AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science,* 111: 218-230. DOI: 10.1016/j.commatsci.2015.09.013

Rajan, H; Uchida, H; Bryan, D; Swaminathan, R; Downs, R and **Hall-Wallace, M** 2006 *Building the American Mineralogist Crystal Structure Database: A recipe for construction of a small Internet database*. In: Sinha A. (Ed.), *Geoinformatics: Data to Knowledge*, Geological Society of America. DOI: 10.1130/2006.2397(06)

The UniProt Consortium 2015 UniProt: a hub for protein information. *Nucleic Acids Research,* 43: D204-D212. DOI: 10.1093/nar/gku989

Wilcox-O'Hearn, Z and **Warner, B** 2008 Tahoe: The Least-authority Filesystem. 21-26. Available at https://gnunet.org/sites/default/files/lafs.pdf.