A FAIR share of crystallographic data: 17+ years of the Crystallography Open Database

Saulius Gražulis Andrius Merkys Antanas Vaitkus The COD Advisory Board

RDA Node Lithuania; VU Institute of Biotechnology; The COD Users' Group <mailto:C-O-D@yahoogroups.com>

17 years ago...

... a small team of engaged scientists decided to make **all** published crystal structures available openly and free of charge to everyone.[a]

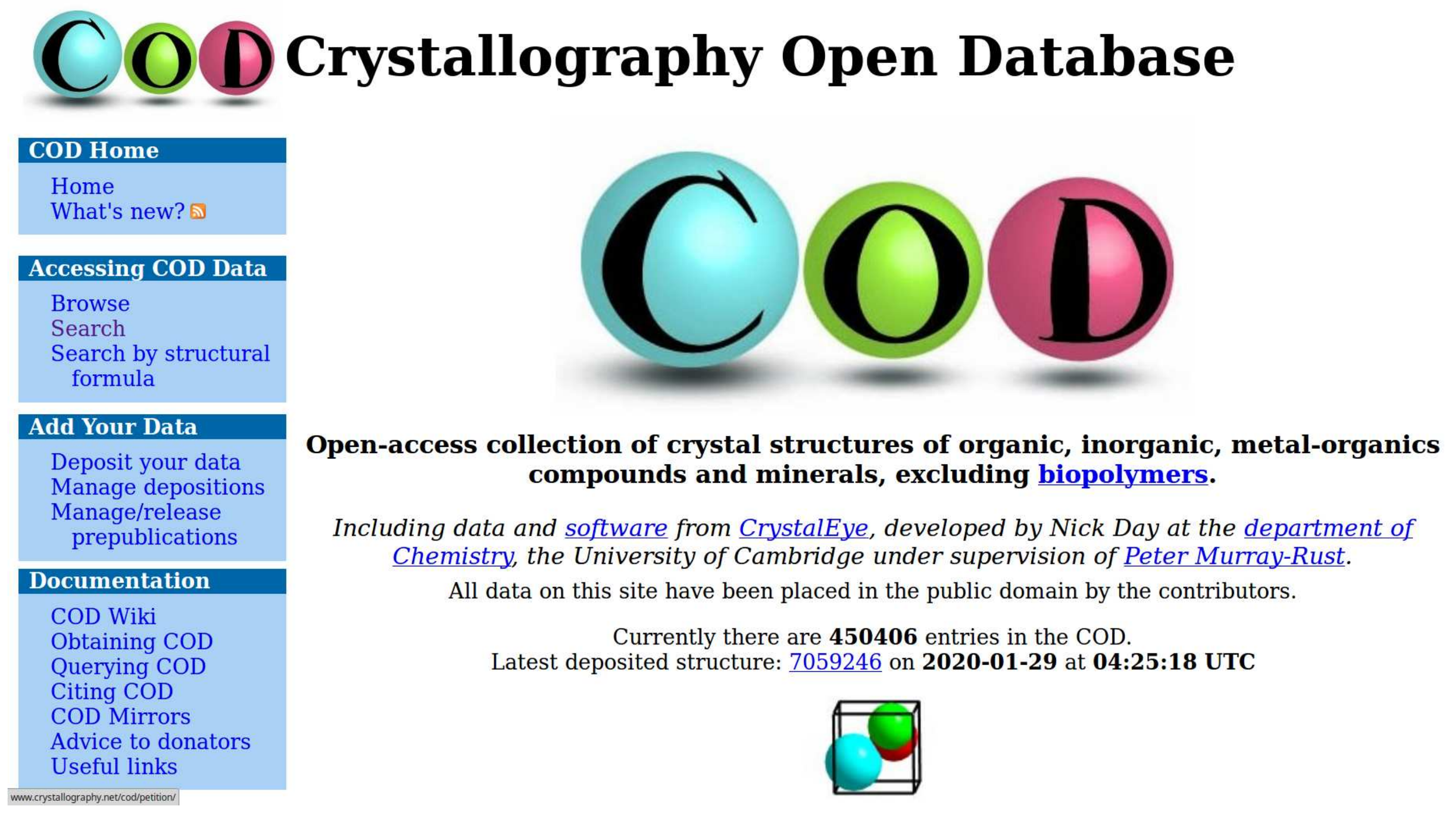
The Crystallography Open Database (COD) [1] project was started by the late Michael Berndt and the current Advisory Board members Daniel Chateigner, Robert T. Downs, Armel Le Bail, Luca Lutterotti, Peter Moeck, Miguel Quirós Olozábal, F.T. Yokochi and others.

During the past decade, more than **450 000 records** were collected in a **curated**, **versioned** scientific database.

[a] gemstonede (Dr. Michael BERNDT) Fri Feb 14, 2003 1:26 pm

The COD

COD is the largest FAIR crystal structure collection.
https://www.crystallography.net/cod



Crystallography Open Database

COD Home
Home
What's new?

Accessing COD Data
Browse
Search
Search by structural formula

Add Your Data
Deposit your data
Manage depositions
Manage/release prepublications

Documentation
COD Wiki
Obtaining COD
Querying COD
Citing COD
COD Mirrors
Advice to donators
Useful links

www.crystallography.net/cod/petition/

Open-access collection of crystal structures of organic, inorganic, metal-organics compounds and minerals, excluding biopolymers.

Including data and software from CrystalEye, developed by Nick Day at the department of Chemistry, the University of Cambridge under supervision of Peter Murray-Rust.

All data on this site have been placed in the public domain by the contributors.

Currently there are **450406** entries in the COD.
Latest deposited structure: 7059246 on **2020-01-29** at **04:25:18 UTC**

The FAIRness of the COD

The COD database is:

- **F**indable:
  - Has a unique, stable identifier for every record (a record ID, e.g. **COD 2006278** [8]);
  - Resolves record IDs to actual data (not just a landing page!);
  - No. 1 for DuckDuckGo and Google searches for the phrase "crystallography database" (checked 2020-01-29).

  Accessible:
  - Each COD ID can be resolved to a **human-readable** Web page: https://www.crystallography.net/cod/2006278.html
  - Each COD ID can be resolved to a **machine-readable** data stream: https://www.crystallography.net/cod/2006278.cif
  - Any previous **revision** of any COD record can be retrieved, at any time: https://www.crystallography.net/cod/2006278.cif@12345
- **I**nteroperable:
  The data stream is in the standard CIF [2] format and is suitable for immediate automated processing:

```
curl -sSL http://crystallography.net/cod/2006278.cif | \
cif_molecule -i --p1 | buffer jmol 2> /dev/null &
```

- **R**eusable:
  Used in a number of scientific and industrial applications:
  - as a material identification database for the EU SOLSA project [9];
  - as an inter-linked resource for other databases, e.g. Raman Open Database [5];
  - as a source of data for macromolecular refinement [4];
  - as a source for search-match applications of macromolecular vendors;
  - as data source for scientific research: see e.g. [7] and 600+ other citations (Google Scholar, checked 2020-01-29).

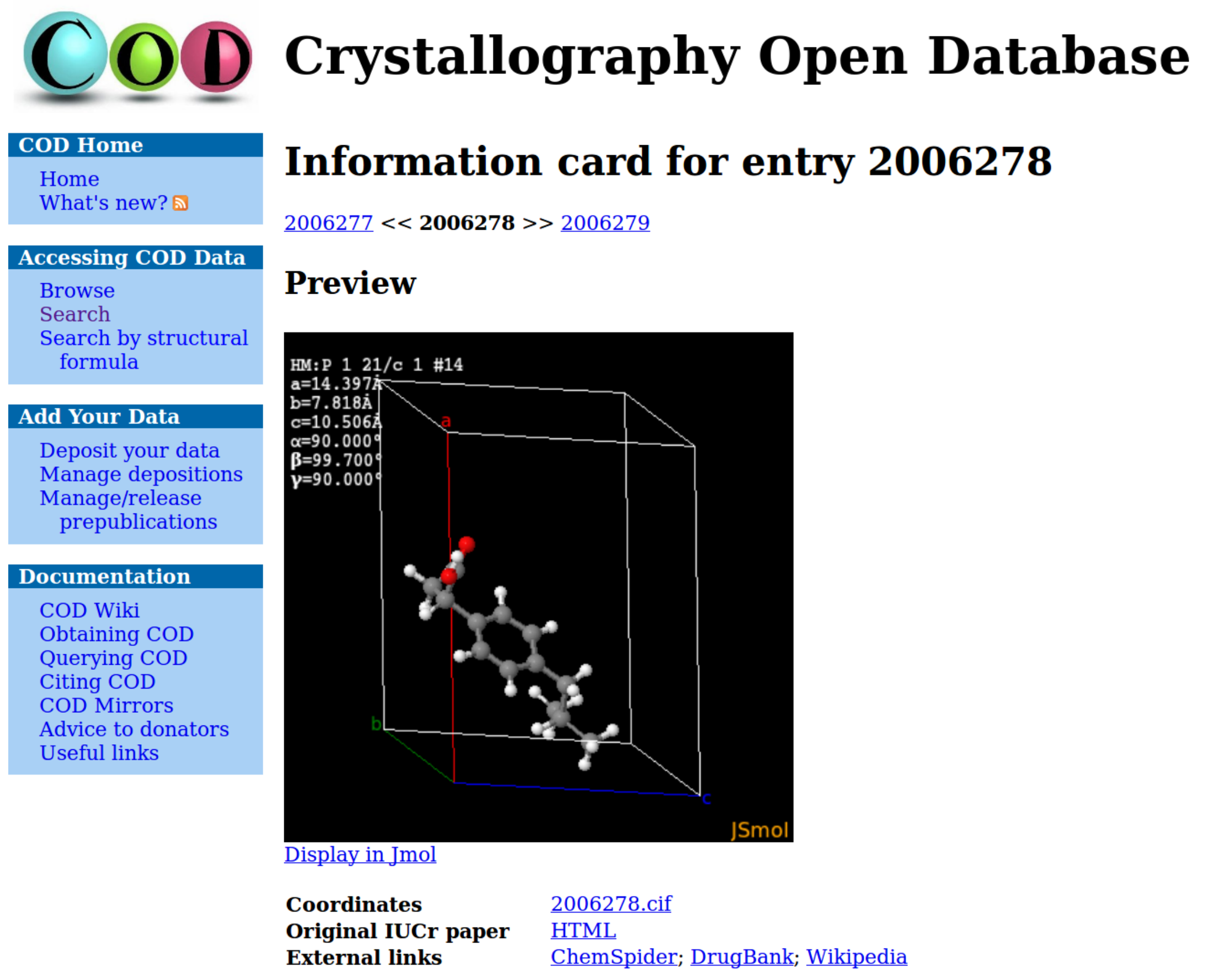
Tools and Methods

The main tools for the COD are **F/LOSS** software:

- Apache2 Web server hosts the Web site and the REST layer; **rewrite rules** ensure that **URIs** remain stable throughout the development of the COD;
- **Jmol** [3] molecule viewer is used to assess the crystal structures by the curators and users;
- **Perl** is an excellent tool for data processing in textual form; a strict Perl CIF parser [6] helps maintaining **correct syntax**.

Ways to access data

Individual COD records can be inspected and downloaded using the Web interface:



Crystallography Open Database

COD Home
Home
What's new?

Accessing COD Data
Browse
Search
Search by structural formula

Add Your Data
Deposit your data
Manage depositions
Manage/release prepublications

Documentation
COD Wiki
Obtaining COD
Querying COD
Citing COD
COD Mirrors
Advice to donators
Useful links

Information card for entry 2006278

2006277 << **2006278** >> 2006279

Preview

```
HM:P 1 21/c 1 #14
a=14.397Å
b=7.818Å
c=10.506Å
α=90.000°
β=99.700°
γ=90.000°
```

Display in Jmol

Coordinates	2006278.cif
Original IUCr paper	HTML
External links	ChemSpider; DrugBank; Wikipedia

Alternatively, multiple COD records can be downloaded for mass-processing using one of the available protocols:
`svn://crystallography.net/cod` (Subversion SVN)
`rsync://crystallography.net/cod-cif` (rsync)

Problems

- Data in many publications are not resolvable automatically (landing page instead of a data stream – not **F**AIR?);
- Data from some publications are behind a paywall (even less F**A**IR?);
- Data formats and **semantics** not formally specified (can it still be FAIR?).

Conclusions

- Using F/LOSS, Unix-architecture based tools allow one to build, curate and maintain an open scientific data collection;
- Version control systems, traditionally used for software development, are instrumental for ensuring **reproducible data (research)**;
- Organised, curated databases with well-defined data semantics add value to the published data;
- Failure of some publishers to adhere to FAIR principles hampers the construction of scientifically valuable databases and obstructs the advance of science in general.

Bibliography

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

[2] Hall et al. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47:655–685, 1991, http://dx.doi.org/10.1107/S010876739101067X.

[3] Hanson. Jmol – a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography*, 43:1250–1260, 2010, http://dx.doi.org/10.1107/S0021889810030256.

[4] Long et al. ACEDRG: A stereo-chemical description generator for ligands. *Acta Crystallographica Section D*, 73(2):112–122, Feb 2017, https://doi.org/10.1107/S2059798317000067.

[5] Mendili et al. Raman open database: first interconnected raman–x-ray diffraction open-access resource for material identification. *Journal of Applied Crystallography*, 52(3):618–625, may 2019, https://doi.org/10.1107/s1600576719004229.

[6] Merkys et al. COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *Journal of Applied Crystallography*, 49(1), Feb 2016, http://dx.doi.org/10.1107/S1600576715022396.

[7] Mounet et al. Two-dimensional materials from high-throughput computational exfoliation of experimentally known compounds. *Nature Nanotechnology*, Feb 2018, https://doi.org/10.1038/s41565-017-0035-5.

[8] Shankland et al. Refinement of Ibuprofen at 100 K by single-crystal pulsed neutron diffraction. *The Crystallography Open Database*, rev. 247336, 1997, https://www.crystallography.net/cod/2006278.cif.

[9] The SOLSA project. An innovative project for sustainable exploration technologies and geomodels: On-line-on-mine-real-time, 2016. http://solsa-mining.eu.

License

Created with Ubuntu → MySQL → R Sweave → LaTeXbeamerposter

RDA Node Lithuania

grazulis commercial-at ibt dot lt