17 years ago...

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

The Crystallography Open Database (COD) project was started by the late Michael Berndt and the current Advisory Board members Daniel Chatteigner, Robert T. Downs, Armel Le Bail, Luca Lutterotti, Peter Moect, 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.

The COD

COD is the largest FAIR crystal structure collection.

https://www.crystallography.net/cod

Crystallography Open Database

Open-access collection of crystal structures of organic, inorganic, metal-organic compounds and intermetallic compounds.

Including data and software from CrystalDek developed by Nick Day of the Department of Chemistry, the University of Cambridge under supervision of Sir Winfrid Maslen.

All data entries have been placed in the public domain by the contributors.

Currently there are 450,000 entries in the database.

The F AIRness of the COD

The COD database is:

- Findable:
  - 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#r12345

- Interoperable:
  - The data stream is in the standard CIF [2] format and is suitable for immediate automated processing:
    curl -sL http://crystallography.net/cod/2006278.cif | cif_molecule -i --p1 | buffer jmol 2> /dev/null &

- Reusable:
  - 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 [6];
    - 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).

Ways to access data

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

Crystallography Open Database

Information card for entry 2006278

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 FAIR?);
- Data from some publications are behind a paywall (even less FAIR?);
- 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


License

This poster is distributed under Creative Commons Attribution 4.0 International License.

Created with Ubuntu → MySQL → R Sweave → EEP/beamerposter

grazulis commercial-at ibt dot lt

COD::CIF::Parser: an error-correcting CIF parser for the Perl language.

Journal of Applied Crystallography

Data in many publications are not resolvable automatically (landing page instead of a data stream – not FAIR?).

Failure of some publishers to adhere to F AIR principles hampers the construction of scientifically valuable databases and obstructs the advance of science in general.

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.

The COD is the largest FAIR crystal structure collection.

https://www.crystallography.net/cod