Using COD and TCOD, searching and getting data

Andrius Merkys
andrius.merkys@gmc.vu.lt

USNC/Cr Database Workshop 2022

Family of open-access structural databases

Database	Records	License	URL	Est.
COD	490 000	public domain	https://www.crystallography.net/cod/	2003
PCOD	1 000 000	public domain	https://www.crystallography.net/pcod/	2003
MPOD	300	public domain	http://mpod.cimav.edu.mx/	2010
TCOD	2 600	public domain	https://www.crystallography.net/tcod/	2013
ROD	1 100	public domain	https://solsa.crystallography.net/rod/	2017

COD entry

- Describes a single crystal structure
- Has provenance record
- May be accompanied by diffraction data
- Is assigned permanent 7-digit identifier

Deposition types in the COD

- Published material:
  - harvested journal supplements
  - donated collections
  - individual depositors
- Prepublication material
- Personal communications

Entry information card

Contents:

- 3D structure preview
- links to download files
- links to other databases
- bibliographic data
- cell parameters
- contents (name, formula, SMILES)
- space group information
- experiment details
- version history

Example:

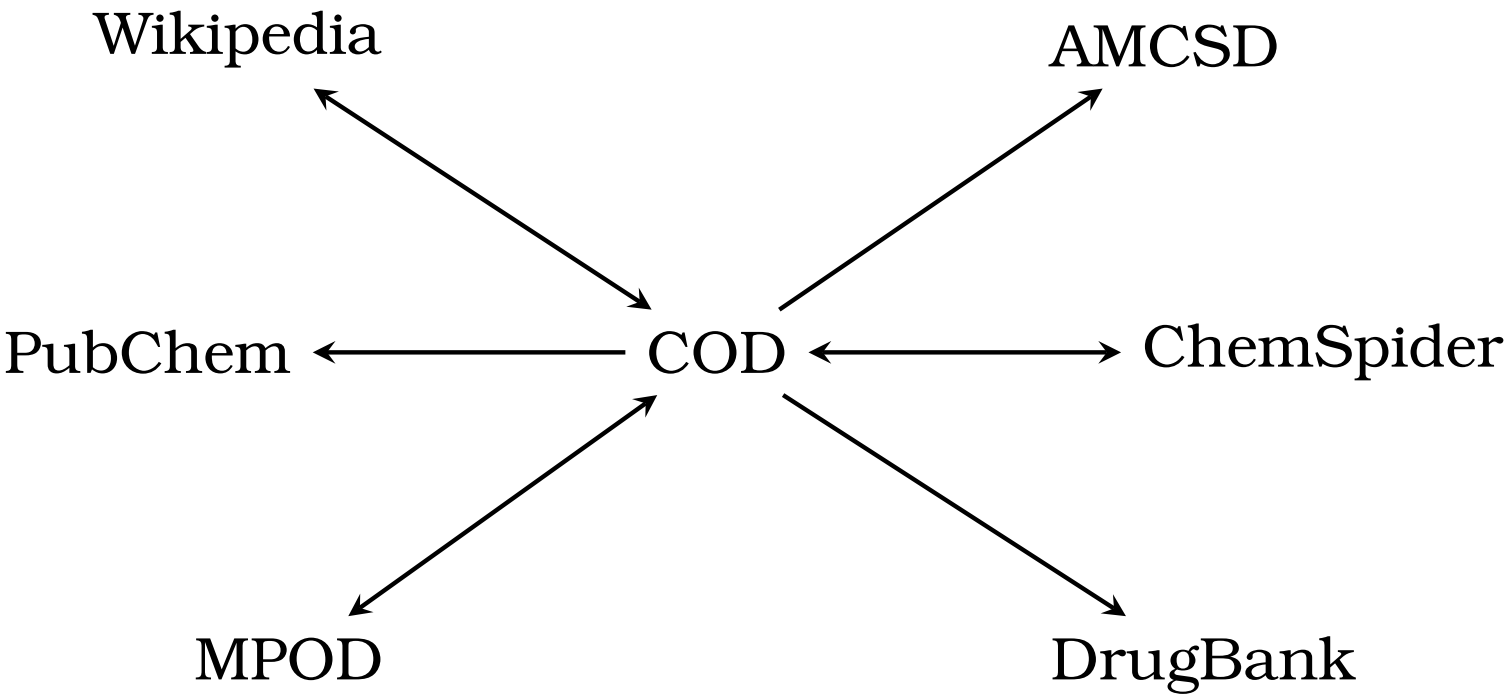
`https://www.crystallography.net/cod/4308000.html`

Crystallographic Information File/Framework (CIF)

COD and TCOD uses CIF v1.1

- ASCII-based human-readable text
- Data storage:
  - key-value pairs
  - tables
- 3D viewers:
  - Avogadro
  - Jmol
- I/O software libraries:
  - codcif, CIF API (C)
  - ciftools-java (Java)
  - PyCifRW, pycodcif (Python)

Links with other databases



Data versioning

Changes:

- are recorded
- get stable sequential identifiers
- are accompanied by log messages
- can be viewed

Curation

Curation

- Deposition time fixes
- (Semi-)automated fixes
- Manual curation

Deposition time fixes

- Fix CIF syntax errors
- Reformat or estimate space group
- Reformat summary chemical formula
- Calculate cell volume
- Exclude unknown or placeholder CIF data items
- Convert temperature values
- Perform various other fixes

Unusual entries

COD entries regarded unusual:

- duplicate entries (≈ 4500, e.g., 1000018)
- theoretical structures (≈ 600, e.g., 2100167)
- entries without coordinates (≈ 240, e.g., 1000195)
- retracted structures (≈ 150, e.g., 2015946)

Validation database

- Uses CIF dictionaries to check data conformity
- Updated periodically
- Provides insight into most frequent issues

Table of validation issues:

`https://sql.crystallography.net/db/cod_validation/validation_issue`

Vaitkus et al., 2021

Querying COD

Search methods:

- bibliography
- cell parameters
- cell contents
- substructure by SMILES
- substructure by drawn fragment

Response formats:

- human-readable HTML
- database ID list
- CIF URL list
- CSV
- archive of matching files

Download the whole COD. Quarterly releases

Links

- Link for the most recent release:

  `https://www.crystallography.net/cod/archives/cod-cifs-mysql.tgz`

- Older releases:

  `https://www.crystallography.net/cod/archives/<YEAR>/data/`

Contents

- coordinate files
- SQL table data

Sizes of 2021-10-16 release

- TAR GZ – 17 GB
- TAR XZ – 12 GB
- ZIP – 18 GB

Individual entry access

Whole file tree for browsing:

`https://www.crystallography.net/cod/cif`

Download individual entry

- coordinate file

  `https://www.crystallography.net/cod/<ID>.cif`
- diffraction data

  `https://www.crystallography.net/cod/<ID>.hkl`
- coordinate file at specified revision

  `https://www.crystallography.net/cod/<ID>.cif@<revision>`

Download the whole COD. *Rsync*

Rsync access

```
$ mkdir cif hkl
$ rsync -aq --delete rsync://www.crystallography.net/cif/ cif/
$ rsync -aq --delete rsync://www.crystallography.net/hkl/ hkl/
```

(for verbose mode, replace `-q` with `-v`)

Advantages:

- Subsequent `rsync` executions will fetch updates
- Network traffic is kept minimal

Access with *Subversion*

Checkout (downloads uncompressed COD CIF collection)

```
$ export LC_TIME=en_US.UTF-8 # Fix locale
$ svn checkout svn://www.crystallography.net/cod/cif
```

Fetch updates

```
$ svn up cif/
```

Go back to specific revision (here 1234)

```
$ svn up cif/ -r1234
```

Examining changes with *Subversion*

Get contents of CIF file in revision 12345

```
$ svn cat svn://www.crystallography.net/cod/cif/1/00/00/1000000.cif -
    r12345
```

Get change of CIF file in revision 91932

```
$ svn diff svn://www.crystallography.net/cod/cif/1/00/00/1000000.cif -
    c91932
Index: 1000000.cif
===================================================================
--- 1000000.cif (revision 91931)
+++ 1000000.cif (revision 91932)
@@ -348,3 +348,4 @@
 N2 H2C O6 1_655 .89 1.88 2.750(5) 164.8
 N2 H2D O4 2_645 .89 2.05 2.869(5) 153.3
 N2 H2E O7 1_655 .89 1.98 2.861(6) 170.9
+_journal_paper_doi 10.1107/S0108270100008532
```

RESTful API for querying the COD

Documentation:

`https://wiki.crystallography.net/RESTful_API/`

Fetch entries describing cucurbiturils from year 2017, in CSV format

```
$ curl https://www.crystallography.net/cod/result -F text=cucurbituril
    -F format=csv -F year=2017
```

SQL access

Connection details:

- Server: `sql.crystallography.net`
- Database: `cod`
- User: `cod_reader`

Explanation of fields in `data` table:

`https://wiki.crystallography.net/cod_mysql_schema/`

SQL access (*cont.*)

Find five most voluminous MOFs

```
$ mysql -h sql.crystallography.net cod -u cod_reader -e "select file,
    commonname, vol from data where commonname like \"%MOF%\" order by
    vol desc limit 5"
+---------+-------------+--------+
| file    | commonname  | vol    |
+---------+-------------+--------+
| 4120255 | bio-MOF-102 | 425902 |
| 4120254 | bio-MOF-101 | 238778 |
| 4109100 | MOF-HTB'    | 148818 |
| 4111295 | mesoMOF-1   | 122163 |
| 4109099 | MOF-HTB     | 110988 |
+---------+-------------+--------+
```

OPTIMADE interface

Select ternary structures having C, Si, Ge or Sn, but not having Pb

```
$ curl -L https://www.crystallography.net/cod/optimade/structures -F '
    filter=elements HAS ANY "C", "Si", "Ge", "Sn" AND NOT elements HAS
    "Pb" AND elements LENGTH 3'
```

Andersen et al., 2021

Theoretical Crystallography Open Database (TCOD)

- Has 8-digit identifiers
- Contains computational details
- Embeds files used in computations

Merkys et al., 2017

Acknowledgments

COD Advisory Board

Saulius Gražulis
Andrius Merkys
Daniel Chateigner
Robert T. Downs
Werner Kaminsky
Armel Le Bail
Luca Lutterotti
Peter Moeck
Peter Murray-Rust
Miguel Quirós