Depositing and managing data in the COD and the TCOD

Antanas Vaitkus

Institute of Biotechnology
Life Sciences Center
Vilnius University

USNC/Cr Database Workshop
2022-04-06

Workshop learning objectives

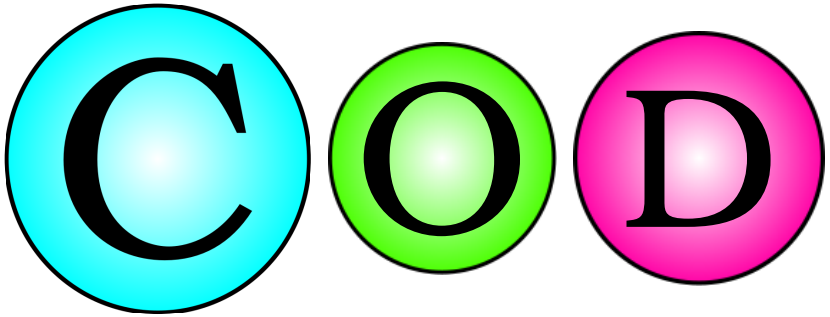
Participants will learn:

- How to deposit their data into the database including what information is needed and how to deposit raw data.
- How to maintain their data records in the (T)COD.
- About the (T)COD data management policy including the way personal data is used, how on-hold records are released to the public and how records are peer reviewed.

Useful resources

- **Workshop slides:** `https://tinyurl.com/4vj9xerp`.
- **COD wiki:** `https://wiki.crystallography.net/`.
- **Peer-reviewed publication:** `https://wiki.crystallography.net/cod/citing/`.
- **Contact email:** `cod-bugs@ibt.lt`.

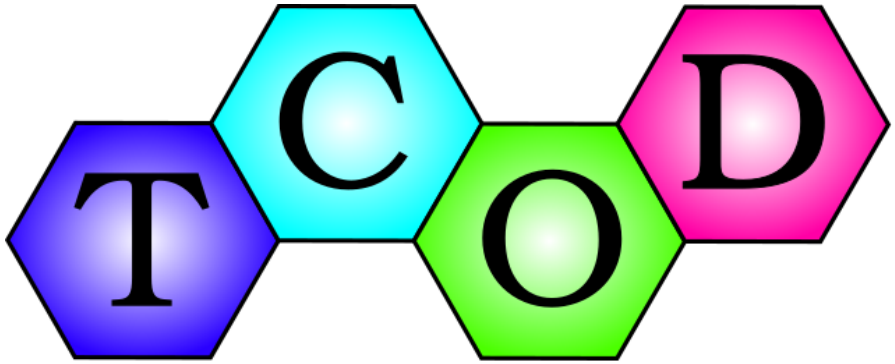
Crystallography Open Database (COD)



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

- Collects experimentally determined crystal structures.
- Includes organic, inorganic, metal-organic compounds and minerals.
- Contains over 485 000 entries.
- Adheres to the FAIR data principles.
- Distributes data under the CC0 licence.

Theoretical Crystallography Open Database (TCOD)



https://www.crystallography.net/tcod/

- Collects theoretically calculated or refined crystal structures.
- Includes organic, inorganic, metal-organic compounds and minerals.
- Contains over 2 908 entries.
- Adheres to the FAIR data principles.
- Distributes data under the CC0 licence.

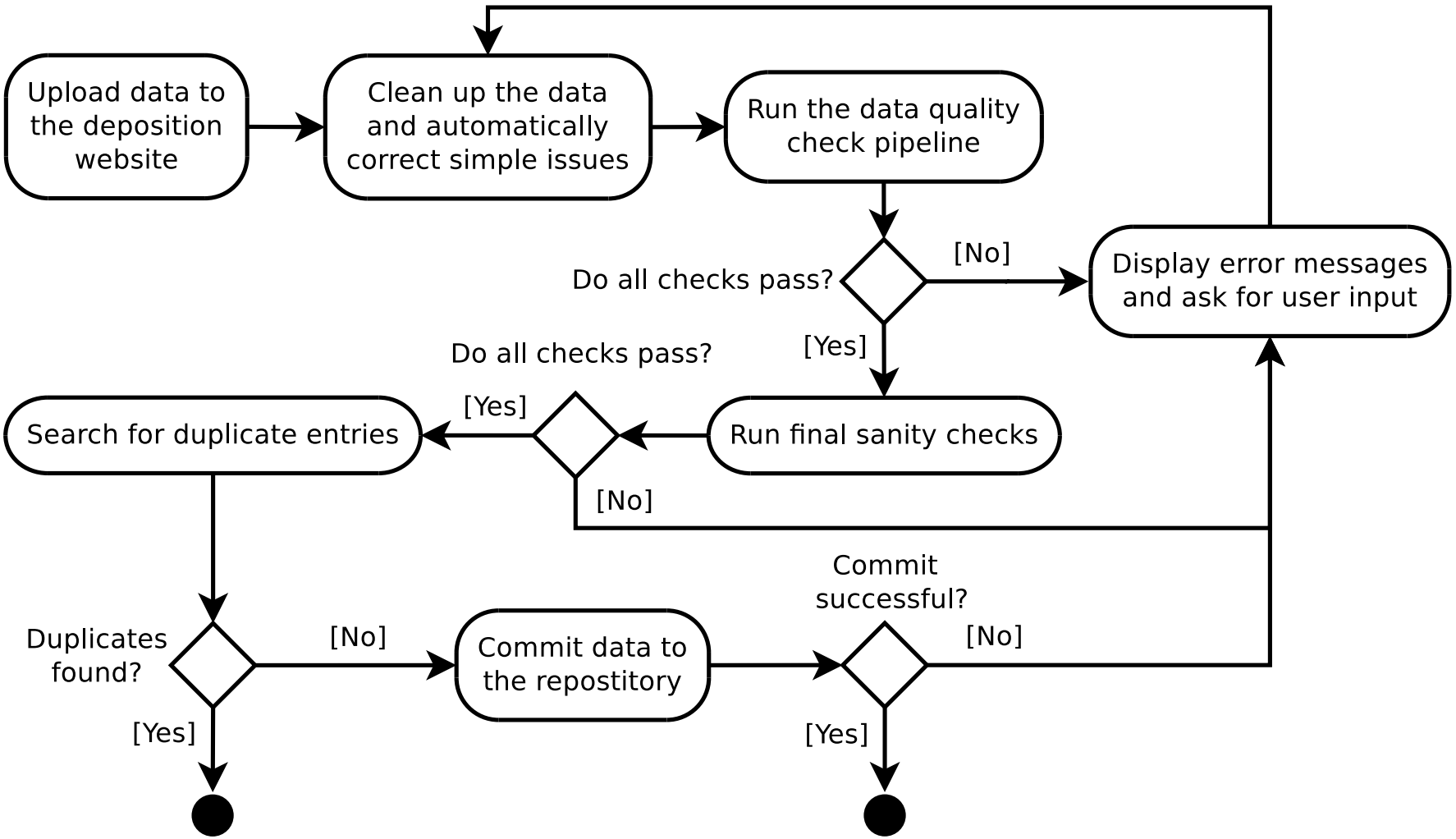
Database organisation

- Uses CIF 1.1 as a carrier format.
- 1 entry – 1 CIF – 1 data block.
- CIF files maybe accompanied by reflection data files (HKL).
- Assigns unique immutable identifiers (COD IDs).
- Stores all files in a world-readable Subversion repository.

Data curation practices

- Routine automated checks.
- Manual checks.
- Validation issue database.
- Data versioning.
- Change logging:
  - Subversion log.
  - Free form `_cod_depositor_comments` data item.
  - Structured `COD_CHANGELOG_ENTRY` loop.

Data deposition workflow



Data curation software

Software available from the `cod-tools` package:

- `COD::CIF::Parser`.
- `cif_filter`.
- `cif_correct_tags`.
- `cif_fix_values`.
- `cif_cod_check`.
- `cif_validate`.

Deposition types: published data

The published data deposition type is intended for data that has been published in peer-reviewed publications.

Mandatory data:

- Authorship (`_publ_author_name`).
- Bibliographic information (`JOURNAL`, `_publ_section_title`).
- Lattice parameters (`_cell_length_*`, `_cell_angle_*`).
- Space group operation list (`SPACE_GROUP_SYMOP`).
- Atomic coordinates (`ATOM_SITE`).
- Chemical formula (`_chemical_formula_sum`).
- Z number (`_cell_formula_units_Z`) or sufficient data to derive it (`_exptl_crystal_density_diffrn`, `_chemical_formula_weight`).

Deposition types: personal communication (1)

The personal communication deposition type is intended for data that was provided directly by volunteer depositors.

Mandatory data:

- Authorship (`_publ_author_name`).
- Lattice parameters (`_cell_length_*`, `_cell_angle_*`).
- Space group operation list (`SPACE_GROUP_SYMOP`).
- Atomic coordinates (`ATOM_SITE`).
- Chemical formula (`_chemical_formula_sum`).
- Z number (`_cell_formula_units_Z`) or sufficient data to derive it (`_exptl_crystal_density_diffrn`, `_chemical_formula_weight`).

Additional requirements:

- Name of the depositing user must appear in the author list.

Deposition types: personal communication (2)

Measures of structure quality:

Data name	Value range
`_refine_ls_R_factor_obs`	< 0.15
`_refine_ls_wR_factor_obs`	< 0.35
`_refine_ls_goodness_of_fit_obs`	$[0.6, 4]$
`_refine_ls_shift/esd_max`	< 0.10

Deposition types: prepublication data (1)

The prepublication data deposition type is intended for data that is in the process of being reviewed or published in a peer-reviewed publication.

- COD ID is assigned immediately after deposition.
- Initially only a minimal set of data is publicly available.
- User controls the full release date.
- Can be released as published data or as personal communication.

Deposition types: prepublication data (2)

Mandatory data:

- Authorship (`_publ_author_name`).
- Space group operation list (`SPACE_GROUP_SYMOP`).
- Lattice parameters (`_cell_length_*`, `_cell_angle_*`).
- Atomic coordinates (`ATOM_SITE`).
- Chemical formula (`_chemical_formula_sum`).
- Z number (`_cell_formula_units_Z`) or sufficient data to derive it (`_exptl_crystal_density_diffrn`, `_chemical_formula_weight`).

Additional requirements:

- Name of the depositing user must appear in the author list.

Deposition types: prepublication data (3)

Measures of structure quality:

Data name	Value range
`_refine_ls_R_factor_obs`	< 0.15
`_refine_ls_wR_factor_obs`	< 0.35
`_refine_ls_goodness_of_fit_obs`	$[0.6, 4]$
`_refine_ls_shift/esd_max`	< 0.10

Recommendations for the deposited data (1)

Recommended data:

- Reflection data (embedded or as a separate file).
- Atomic displacement parameters (ADP).
- R-factor values.
- Molecular geometry values.
- *Any piece of information that the author deems significant.*

Recommendations for the deposited data (2)

Measures of structure quality:

Data name	Value range
_refine_ls_R_factor_obs	< 0.10
_refine_ls_wR_factor_obs	< 0.25
_refine_ls_goodness_of_fit_obs	$[0.8, 2]$
_refine_ls_shift/esd_max	< 0.05

COD user account

- Used only for data deposition and maintenance.
- Requires the minimal disclosure of personal data.
- Collected data will never be forwarded to third parties.
- May be deleted by contacting `cod-bugs@ibt.lt`.

The Test COD database

- `https://www.crystallography.net/cod-test`.
- Uses a different user account than the COD.
- Public.
- May be reset at any point.

RESTful API

- Accepts HTTP POST requests.
- Endpoint:

  `https://www.crystallography.net/`***`cod-test`***`/cgi-bin/cif-deposit.pl`.
- Description:

  `https://wiki.crystallography.net/RESTful_API/#index3h1`.
- Circumvents most of the automated fixes.

RESTful API fields (1)

User authentication fields:

Field	Description
`username`	Depositor's username.
`password`	Depositor's password.
`user_email`	Depositor's e-mail address.

Common deposition fields:

Field	Description
`deposition_type`	`published`, `personal`, `prepublication`.
`cif`	Uploaded CIF file.
`hkl`	Uploaded HKL file.
`output_mode`	Output format: **`html`**, `stdout`.
`progress`	Set to `1` for a more verbose output.

RESTful API fields (2)

Mandatory for `personal` and `prepublication` type depositions:

Field	Description
`author_name`	Name of the author as given in the CIF file.
`author_email`	Authors email address.

Optional for `prepublication` type depositions:

Field	Description
`journal`	Name of the planned publication journal.
`hold_period`	Hold period in months (`0-12`). Default: `6`.
`release`	Set to `1` to release the data into the public domain. Must appear together with the `replace` field.

RESTful API fields (3)

Optional for `published` type deposition:

Field	Description
`doi_only`	Set to `1` to treat digital object identifier (DOI) as sufficient bibliographic reference.

Mandatory when modifying existing data:

Field	Description
`replace`	Set to `1` to replace an existing entry with the given file.
`message`	Log message describing the performed changes.

cif_cod_deposit (1)

The cif_cod_deposit script provides a simple command line interface to the COD deposition API.

Usage example:

```
cif_cod_deposit \
   -c .cod_deposit.cfg \
   --no-print-timestamps \
   --output-mode stdout \
   --cif inputs/xantheose.cif \
   --hkl inputs/xantheose.hkl \
   --log-message \
    'Initial␣deposition␣of␣the␣structural␣data.' \
   --url \
    https://www.crystallography.net/cod-test/cgi-bin/cif-deposit.pl \
   --script \
;
```

cif_cod_deposit (2)

Output of the `cif_cod_deposit` script when run with the `--script` option (folded for readability):

```
curl \
   --silent \
   --show-error \
   -F message=</tmp/tmp-cif_cod_deposit-247839/message \
   -F strict=1 \
   -F output_mode=stdout \
   -F username=</tmp/tmp-cif_cod_deposit-247839/username \
   -F password=</tmp/tmp-cif_cod_deposit-247839/password \
   -F user_email=</tmp/tmp-cif_cod_deposit-247839/user_email \
   --user-agent cif_cod_deposit \
   -F "cif=@inputs/xantheose.cif;filename=xantheose.cif" \
   -F "hkl=@inputs/xantheose.hkl;filename=xantheose.hkl" \
   -F deposition_type=published \
   "https://www.crystallography.net/cod-test/cgi-bin/cif-deposit.pl"
```

cif_cod_deposit (3)

Example of the cif_cod_deposit configuration file (e.g. ".cod.deposit.cfg"):

```
# Comment lines start with '#'
username=cod_user_42
password=p4ssw0rd
user_email=mailbox@domain.com
web_client_ip=127.0. 0.1
journal=Crystallographic Databases
author_name=Name Surname
author_email=mailbox@domain.com
message=Added atomic displacement parameters.
```

Workshop material

- **Workshop material:** `https://tinyurl.com/3t23rs7w`.

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