

Atvira kristalografinė duomenų bazė COD

FAIR duomenų pateikimas gamtos moksluose

Saulius Gražulis

Atvirojo mokslo ir tyrimų duomenų aktualijos
Vilnius, 2020

Vilniaus universitetas, Gyvybės mokslų centras

Biotechnologijos institutas



Ši skaidrių rinkinių galima kopijuoti, kaip nurodyta Creative Commons Attribution-ShareAlike 4.0 International licencijoje



1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

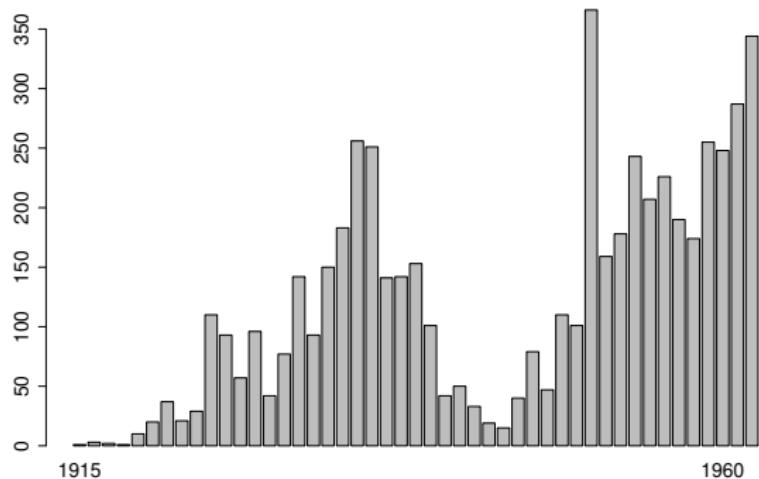
Kiek kristalų struktūrų buvo publikuota kiekvienais metais? Užklausa [COD duomenų bazėje](#):

Kiek kristalų struktūrų buvo publikuota kiekvienais metais? Užklausa [COD duomenų bazėje](#):

```
SELECT count(*) AS nr, year FROM data  
WHERE year IS NOT NULL AND  
GROUP BY year ORDER BY year DESC
```

Kiek kristalų struktūrų buvo publikuota kiekvienais metais? Užklausa **COD** duomenų bazėje:

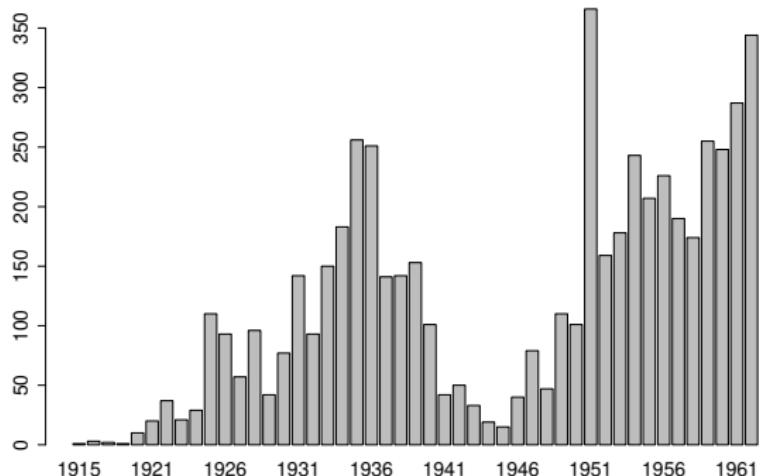
```
SELECT count(*) AS nr, year FROM data  
WHERE year IS NOT NULL AND  
GROUP BY year ORDER BY year DESC
```



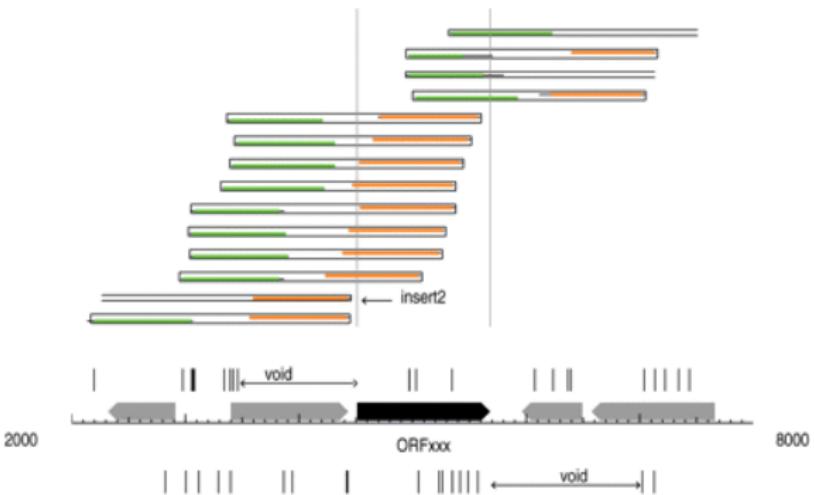
Paprastos užklausos

Kiek kristalų struktūrų buvo publikuota kiekvienais metais? Užklausa **COD** duomenų bazėje:

```
SELECT count(*) AS nr, year FROM data  
WHERE year IS NOT NULL AND  
GROUP BY year ORDER BY year DESC
```



Zheng iš Robertso NEB komandos panaudojo „žalius“ sekoskaitos duomenis *aktyviems* genus karpantiems baltymams aptikti [Zheng et al. (2008)]:



Kinijos tyrėjų grupė padarė atradimą biochemijos srityje be eksperimentinių tyrimų [Li et al. (2008)]:

The screenshot shows a PLOS Computational Biology article page. At the top, there are navigation links for Browse, Publish, and About, along with a search bar. Below the header, it says "OPEN ACCESS" and "PEER-REVIEWED". The article title is "Genes and (Common) Pathways Underlying Drug Addiction" by Chuan-Yun Li, Xizeng Mao, Liping Wei, et al. It was published on January 4, 2008, with a DOI of <http://dx.doi.org/10.1371/journal.pcbi.0040002>. To the right of the article title are two green boxes containing metrics: "159 Save" and "93 Citation" in the top box, and "52,365 View" and "2 Share" in the bottom box. Below these are buttons for "Download PDF", "Print", and "Share". On the left side, there's a sidebar with "Abstract", "Author Summary", "Introduction", "Results", and "Discussion" sections. The main content area has an "Abstract" section with the following text: "Drug addiction is a serious worldwide problem with strong genetic and environmental influences. Different technologies have revealed a variety of genes and pathways underlying addiction; however, each individual technology can be biased and incomplete. We integrated".

<http://slidegur.com/doc/3077570/introducing-bioinformatics-databases>

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

Siekiame, kad duomenys būtų [Wilkinson et al. (2016)]:

- ▶ **Findable:** Randami (automatinėmis priemonėmis!)
- ▶ **Accessible:** Prieinami (automatinėmis priemonėmis!)
- ▶ **Interoperable:** Suderinami (su įvairiomis programomis!)
- ▶ **Reusable:** Pakartotinai panaudojami (ivairiems, nenumatytiems tikslams!)

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

Įvairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
 - ▶
 - ▶
2.
 - ▶
 - ▶
 - ▶
 - ▶

Ivairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
 - ▶ Zenodo <https://doi.org/10.5281/zenodo.3560693>
 - ▶ Zenodo <https://zenodo.org/record/3841841>
2.
 - ▶
 - ▶
 - ▶
 - ▶

Ivairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
 - ▶ Zenodo <https://doi.org/10.5281/zenodo.3560693>
– stendinės pranešimas (PDF)...
 - ▶ Zenodo <https://zenodo.org/record/3841841>
2.
 - ▶
 - ▶
 - ▶
 - ▶

Ivairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
 - ▶ Zenodo <https://doi.org/10.5281/zenodo.3560693>
 - stendinės pranešimas (PDF)...
 - ▶ Zenodo <https://zenodo.org/record/3841841>
 - susieti COVID-19 duomenys (TTL)...

2.

- ▶
- ▶
- ▶
- ▶

Ivairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
 - ▶ Zenodo <https://doi.org/10.5281/zenodo.3560693>
 - stendinės pranešimas (PDF)...
 - ▶ Zenodo <https://zenodo.org/record/3841841>
 - susieti COVID-19 duomenys (TTL)...
2. Tematiniai duomenų archyvai (PDB, **COD**, NCBI, SwissProt, EuropePMC, PubMed (!));
 - ▶
 - ▶
 - ▶
 - ▶

Ivairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
 - ▶ Zenodo <https://doi.org/10.5281/zenodo.3560693>
 - stendinės pranešimas (PDF)...
 - ▶ Zenodo <https://zenodo.org/record/3841841>
 - susieti COVID-19 duomenys (TTL)...
2. Tematiniai duomenų archyvai (PDB, COD, NCBI, SwissProt, EuropePMC, PubMed (!));
 - ▶ <https://www.crystallography.net/cod/1557684.cif>
 - ▶ <https://www.crystallography.net/cod/1544162.html>
 - ▶ <https://www.pdb.org/pdb/files/1KNV.cif>
 - ▶ <https://www.rcsb.org/structure/2IXS>

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

COD duomenų bazė

Crystallography Open Database

(COD, <https://www.crystallography.net>)

[Gražulis et al. (2009), Gražulis et al. (2012)]:

COD turinys:

- ▶ Kiek įmanoma, visos mažų molekulių (organinių, mineralų) struktūros;
- ▶ Atvira duomenų bazė, galima paieška Žiniatinklio puslapyje, automatinėmis priemonėmis ir visos DB nukėlimas;

COD projektas

But what if crystallographers work together to establish a public domain database with all relevant crystallographic data? This would not only overcome the current situation with 'fragmented' databases, it would also prevent for becoming dependent from monopolists.

What would be needed?

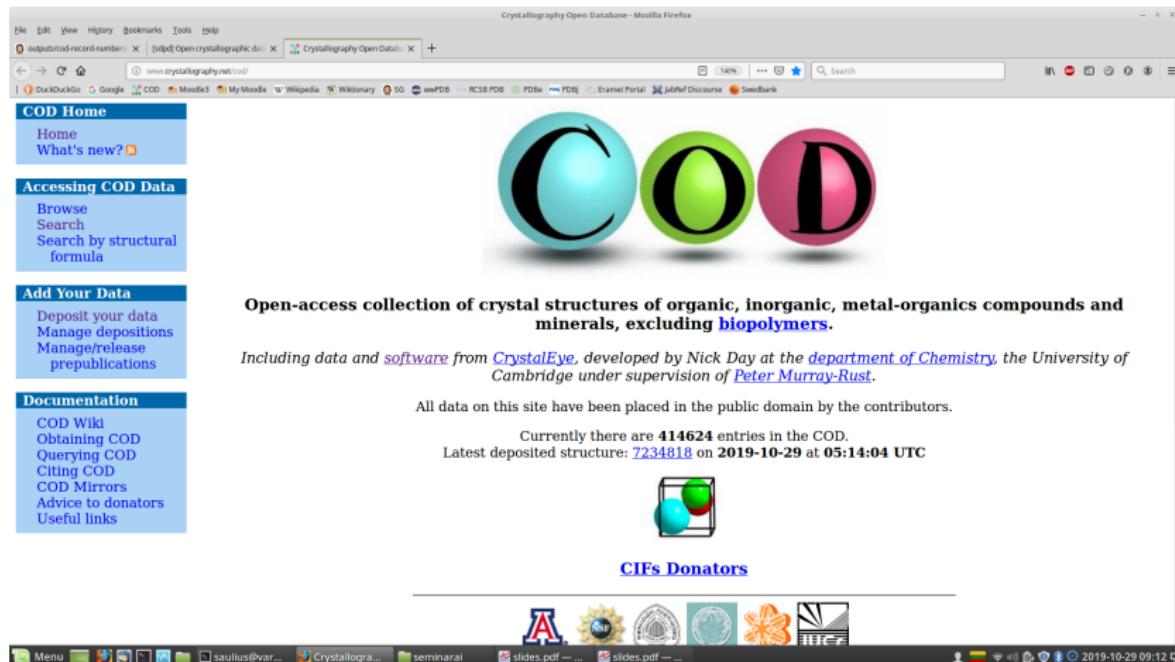
1. A small team of engaged scientists with some experience in database and software design to coordinate the project.
2. The authors (i.e. the scientific community = YOU) who provides the project with database entries (note, that if you have'nt sold your experimental results exclusively, you are free to distribute the data to such a database, even if they have already been part of a publication - and a lot of good data have never been published).
3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval.

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

Po 17 metų ... :)

The Crystallography Open Database

<http://www.crystallography.net/cod>



File Edit View History Bookmarks Tools Help

outputted record number | (sdpd) Open crystallographic data | Crystallography Open Database - Mozilla Firefox

← → ⌛ ⌚ www.crystallography.net/cod +

DuckDuckGo Google CDD Moodle3 My Moodle Wikipedia Wikidictionary SG WebPDB RCSB PDB PDBe PDBj Brahma Portal Jmol Discourse Seedbank

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

Crystallography Open Database - Mozilla Firefox

14% 100% ⌛ ⌚

Crystallography Open Database

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.

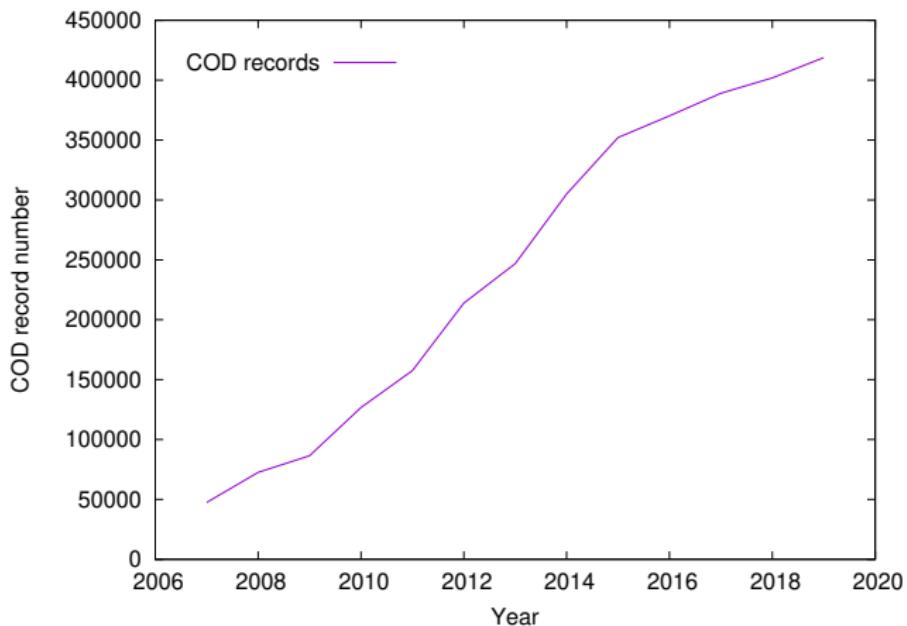
Currently there are **414624** entries in the COD.
Latest deposited structure: [7234818](#) on **2019-10-29 at 05:14:04 UTC**

CIFs Donators

saulius@var... Crystallography seminarai slides.pdf ... slides.pdf ...

COD tvarumas ir augimas

COD gyvuoja jau 17 metų, išaugo 8 kartus per paskutinius 10 metus; šiuo metu talpina virš 450 000 įrašų (2020):



1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

1. Bendras (centralizuotas arba paskirstytas) registratorius:
 - ▶ COD identifikatoriai (pvz. COD 2000000);
 - ▶ PDB identifikatoriai (pvz. PDB 1KNV);
 - ▶ DOI (pvz. 10.1093/nar/gkn883);
 - ▶ URI (pvz. <https://www.w3.org/Provider/Style/URI.html>)
 - ▶ ISSN, ISBN, PMID, PMCID, ...
2. Randomizuoti identifikatoriai:
 - ▶ UUID (pvz. 90376010-a315-11ea-adba-6bb1c61159af)
 - ▶ Kriptografinės kontrolinės sumos (pvz. git commit 42a03a255612b8d43ecd77bb0acc02def888f688, 42a03a2);

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ **Metaduomenys**
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

- ▶ Bendrieji metaduomenys:
 - ▶ Bibliografija
 - ▶ Duomenų bazės tvarkymo įrašai (revizija, įvedimo data, ...)
- ▶ Dalykinės srities metaduomenys:
 - ▶ Eksperimento salygos
 - ▶ Kokybės kriterijai
 - ▶ COD: cheminė formulė
 - ▶ COD: kristalo auginimo salygos
 - ▶ Sąsajos su kitomis duomenų bazėmis
 - ▶ ir t.t. (metaduomenų sąrašas potencialiai atviras!)

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

► Irašai tarpusavyje palyginami

cod/2014638-head.cif:

```

data_2014638
loop_
_publ_author_name
'U\,car, \.Ibrahim'
# et al.
_publ_section_title
;
3-Acetoxy-2-(acetylamino)pyridinium ...
;
_journal_issue          3
_journal_name_full      'Acta Cryst. C'
_journal_paper_doi
10.1107/S0108270104031841
#...
_chemical_formula_sum    'C13 H10 N2 O6'
_chemical_formula_weight 290.23
#...
_cell_length_a           8.8959(16)
#...
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
#...
C1 0.1598(5) 0.4645(4) 0.7452(4) # ...
C2 0.1712(5) 0.4275(5) 0.9028(5) # ...

```

cod/2103669-head.cif:

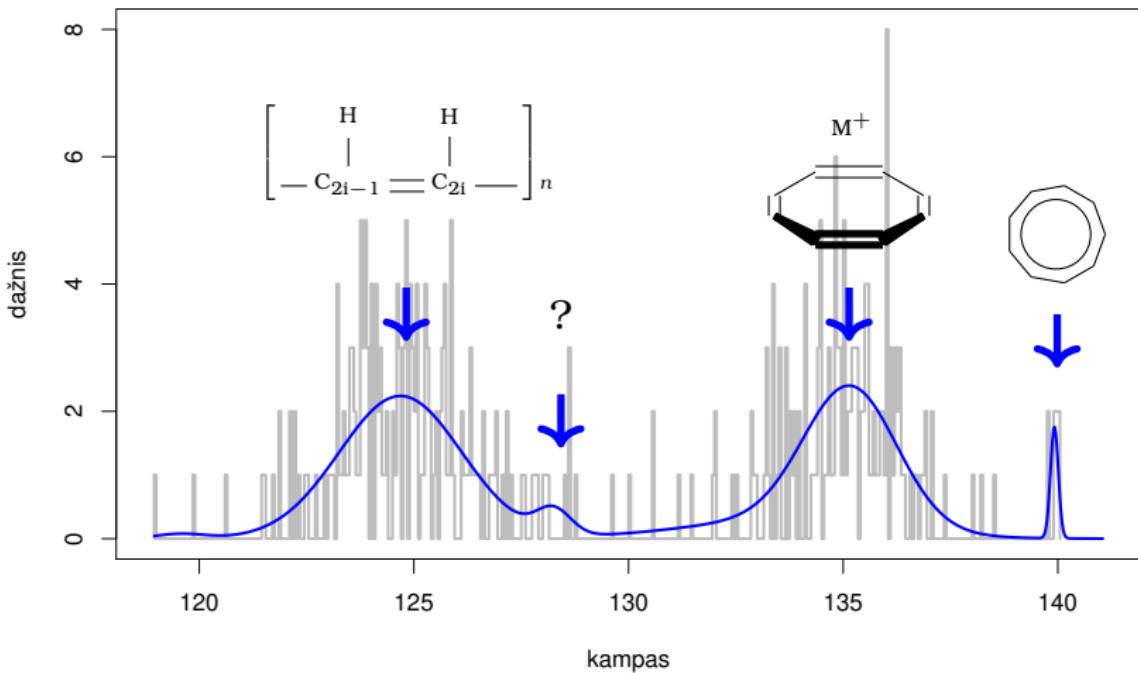
```

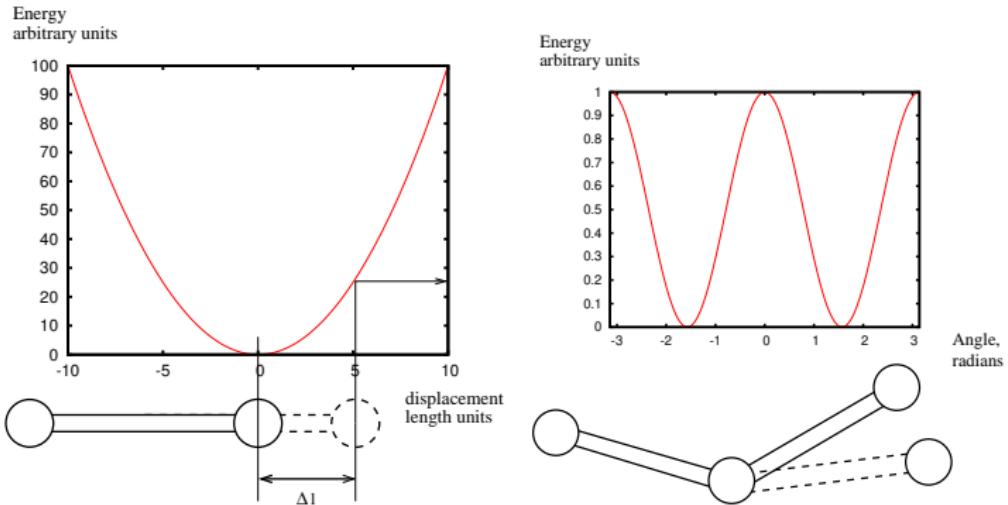
data_2103669
loop_
_publ_author_name
'Rychlewska, Urszula'
'War\.zajtis, Beata'
_publ_section_title
;
... dibenzoyltartaric acid
;
_journal_issue          3
_journal_name_full      'Acta Cryst. B'
_journal_paper_doi
10.1107/S010876810100430X
# ...
_chemical_formula_sum    'C20 H20 N2 O6'
_chemical_formula_weight 384.38
# ...
_cell_length_a           8.9153(6)
# ...
loop_
_atom_site_label
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
#...
C1 .1554(3) -.9495(2) -1.00258(16) # ...
C2 -.0044(3) -.9472(2) -1.03195(16) # ...

```

Sudėtingų bruozų atpažinimas

Polienų grandinės kampai iš COD





$$E = \frac{1}{2}k\Delta l^2 \quad p(\Delta l) \sim e^{-\frac{E}{k_B T}} = e^{-\frac{\Delta l^2}{2\sigma^2}} \quad E = k(1 + \cos(n\Delta\varphi)), \text{ if } n > 0$$

- ▶ Kristalų statistika leidžia nustatyti jėgas, veikiančias atomus molekulėse.



Acta Crystallographica Section D STRUCTURAL BIOLOGY



search IUCr Journal

home archive editors for authors for readers submit subscribe open access

D RESEARCH PAPERS

Acta Cryst. (2017), D73, 112-122
<https://doi.org/10.1107/S2059798317000067>
Cited by 46

AceDRG: a stereochemical description generator for ligands

F. Long^①, R. A. Nicholls^②, P. Emsley, S. Gražulis^③, A. Merkys^④, A. Vaikus and G. N. Murshudov^⑤

The program AceDRG is designed for the derivation of stereochemical information about small molecules. It uses local chemical and topological environment-based atom typing to derive and organize bond lengths and angles from a small-molecule database: the Crystallography Open Database (COD). Information about the hybridization states of atoms, whether they belong to small rings (up to seven-membered rings), ring aromaticity and nearest-neighbour information is encoded in the atom types. All atoms from the COD have been classified according to the generated atom types. All bonds and angles have also been classified according to the atom types and, in a certain sense, bond types. Derived data are tabulated in a machine-readable form that is freely



[Long et al. (2017a), Long et al. (2017b)]

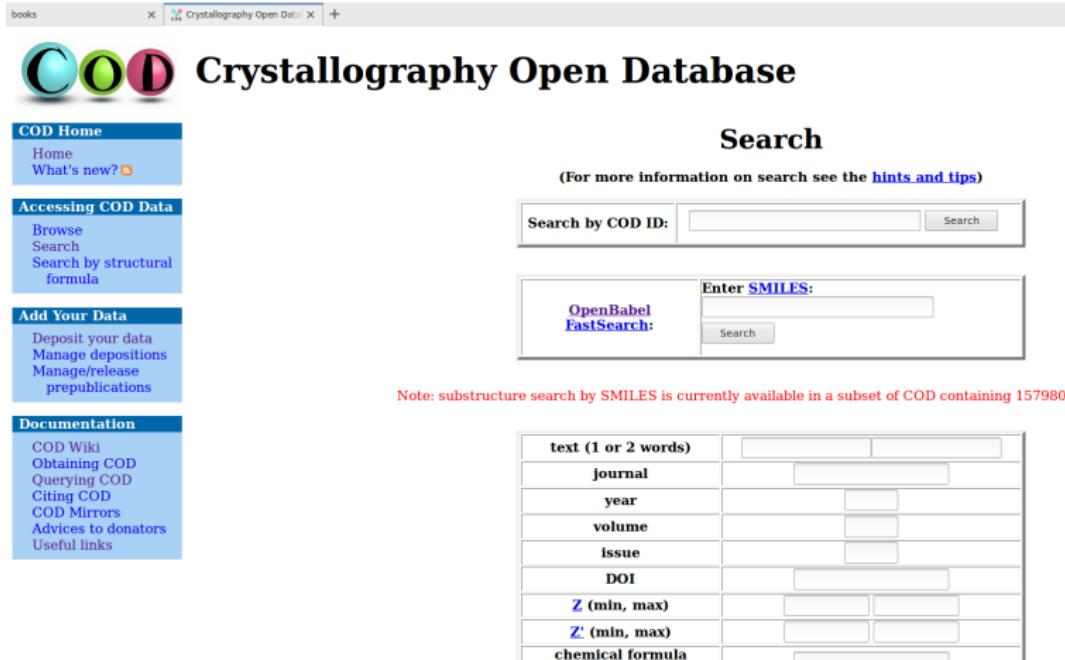
1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

Pasiekiamumas

- ▶ Duomenys turėtų būti pasiekiami URI (URL) pagalba;
- ▶ Duomenys turėtų būti mašina skaitomu pavidalu;
- ▶ Duomenys turėtų surandami **automatinių užklausų** pagalba.
- ▶ Neblogai turėti ir Žiniatinklio paieškos formą ;).
- ▶ Reikėtų siekti atkartojamų užklausų
[Rauber et al. (2016)]

COD paieškos forma

<http://www.crystallography.net/cod/search.html>



The screenshot shows the homepage of the Crystallography Open Database (COD). The top navigation bar includes links for books, COD, and Crystallography Open Data. The main header features the COD logo and the text "Crystallography Open Database". On the left, there's a sidebar with sections for 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), and Documentation (COD Wiki, Obtaining COD, Querying COD, Citing COD, COD Mirrors, Advices to donators, Useful links).

Search
(For more information on search see the [hints and tips](#))

Search by COD ID:

Enter SMILES:

OpenBabel FastSearch:

Note: substructure search by SMILES is currently available in a subset of COD containing 157980 :

text (1 or 2 words)	<input type="text"/>
journal	<input type="text"/>
year	<input type="text"/>
volume	<input type="text"/>
issue	<input type="text"/>
DOI	<input type="text"/>
Z (min, max)	<input type="text"/> <input type="text"/>
Z' (min, max)	<input type="text"/> <input type="text"/>
chemical formula	<input type="text"/>

COD užklausų pavyzdžiai

- ▶ Naudojant **stabilius** URL'us (REST):
 - ▶ <http://www.crystallography.net/cod/2100202.html>
 - ▶ <http://www.crystallography.net/cod/2100202.cif>
 - ▶ <http://www.crystallography.net/cod/result?text=caffeine>
- ▶ Naudojant SQL **užklausų kalba** duomenų bazėje:
 - ▶ `mysql -u cod_reader cod -h www.crystallography.net \
-e 'select file, a, b, c, vol, formula \
from data where \
year between 2013 and \
2014 and \
formula regexp " C[0-9]* " \
order by vol desc limit 10'`

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

Formatai

- ▶ Naudokime atvirus, gerai dokumentuotus formatus;
- ▶ Pageidautina laikyti duomenis teksto pavidale, jei leidžia vieta;
- ▶ Super formatai :)
 - ▶ CSV, CIF, JSON (su schema!), XML (su schema!), PDB, FASTA – tekstu paremti;
 - ▶ HDF5 (su žodynais!), EXI (su schemomis!), BSON – dvejetainiai;
 - ▶ TXT (UTF-8 Unikodo koduotės tekstas);
- ▶ Šiaip sau formatai:
 - ▶ XLS(X), DOC(X), ...;

CIF karkasas kristalografinių duomenų mainams

Sukurtas ir palaikomas Tarptautinės kristalografų draugijos (International Union of Crystallography, IUCr) [Hall et al. (1991)].

examples/data/2100858-head.cif:

```
data_2100858
loop_
  _publ_author_name
  'Buttner, R. H.'
  'Maslen, E. N.'
  _publ_section_title
;
  Structural parameters and electron difference density in BaTiO~3~
;
  _journal_issue          6
  _journal_name_full      'Acta Crystallographica Section B'
  _journal_page_first     764
  _journal_page_last      769
  _journal_volume         48
  _journal_year           1992
  _chemical_compound_source 'synthetic, from a mixture of KF:KMnO4:BaTiO3'
  _chemical_formula_sum   'Ba O3 Ti'
  _chemical_formula_weight 233.24
  _symmetry_cell_setting  tetragonal
  _symmetry_space_group_name_Hall 'P 4 -2'
  _symmetry_space_group_name_H-M  'P 4 m m'
  _cell_length_a          3.9998(8)
  _cell_length_b          3.9998(8)
  _cell_length_c          4.0180(8)
```

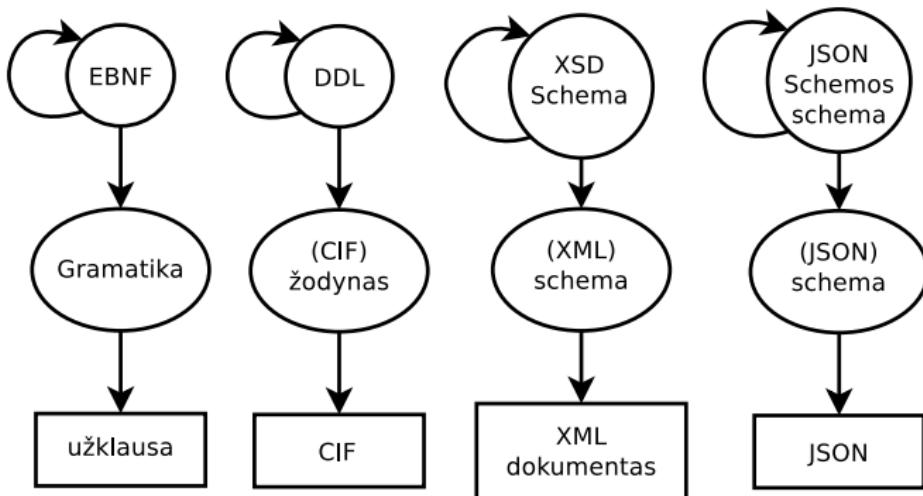
1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ **Semantika**
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

- ▶ CIF – žodynai!
- ▶ XML, JSON – schemas!
- ▶ SQL – schemas!
- ▶ Semantiniai tinklai (RDF);

examples/dictionaries/cif-core-example.cif:

```
data_cell_length_
loop_ _name           '_cell_length_a'
                           '_cell_length_b'
                           '_cell_length_c'
    _category          cell
    _type              numb
    _type_conditions  esd
    _enumeration_range 0.0:
    _units             A
    _units_detail     'angstroms'
    _definition
;
    Unit-cell lengths in angstroms corresponding to the structure
    reported. The values of _refln_index_h, *_k, *_l must
    correspond to the cell defined by these values and _cell_angle_
    values. The values of _diffrn_refln_index_h, *_k, *_l may not
    correspond to these values if a cell transformation took place
    following the measurement of the diffraction intensities. See
    also _diffrn_reflns_transf_matrix_.
;
```

Visada užtenka trijų lygių duomenų apribojimams aprašyti!



1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

Pilnumas

- ▶ COD įrašų skaičius: > 450 000;
- ▶ Žinomų paskelbtų struktūrų skaičius (pagal DataCite): > 818 000 (2019-12-31);

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

- ▶ IUCr kokybės kriterijai
 - ▶ IUCr kriterijų sąrašas <ftp://ftp.iucr.org/pub/dvntests>
 - ▶ IUCr Publikacijų kokybės kriterijai
- ▶ COD kokybės kriterijai
 - ▶ ✓✓ Teisinga sintaksė;
 - ▶ ✓ Validacija pagal žodynus;
 - ▶ ✓ Validacija pagal duomenų statistiką;
 - ▶ ✗ Validacija pagal pamatinius fizikinius principus;

COD duomenų tikrinimo politika:

1. Sintaksės patikrinimas:

```
§ cifparse 7234818.cif
```

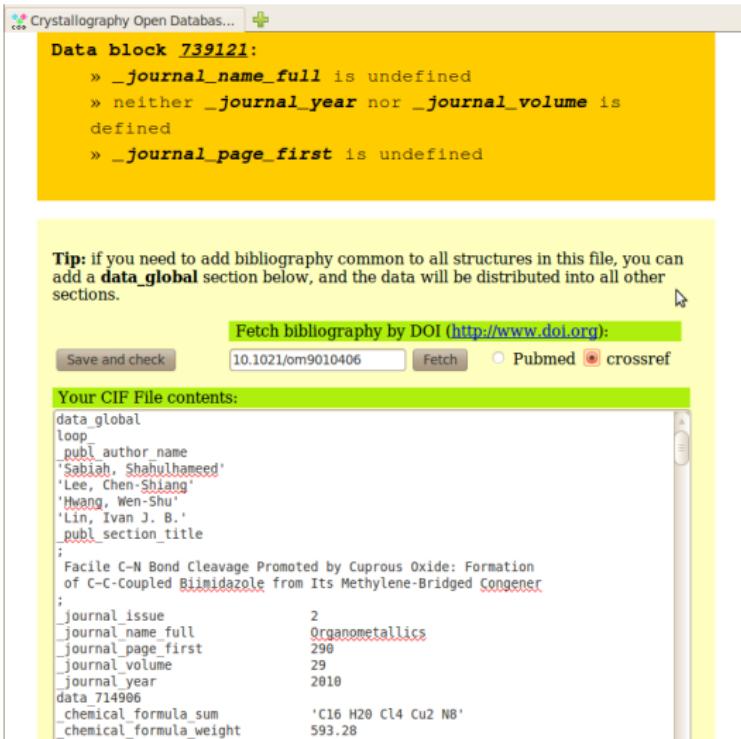
2. Semantinis patikrinimas (pagal žodynus):

```
§ cif_validate -D cif_core.dic 7234818.cif
```

3. Specifiniai COD duomenų bazės testai:

```
§ cif_cod_check 7234818.cif
```

COD validavimo ir deponavimo svetainė



The screenshot shows a web-based application for managing crystallographic data. At the top, a yellow box displays validation errors:

```
Data block 739121:  
» _journal_name_full is undefined  
» neither _journal_year nor _journal_volume is defined  
» _journal_page_first is undefined
```

Below this, a tip is provided:

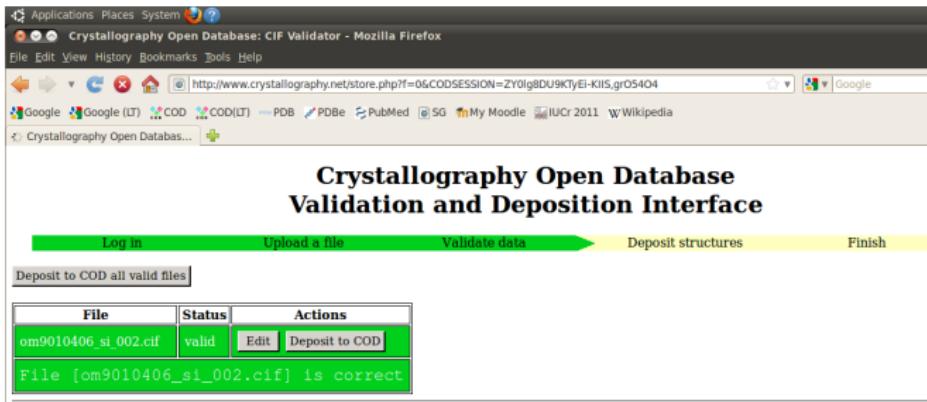
Tip: if you need to add bibliography common to all structures in this file, you can add a **data_global** section below, and the data will be distributed into all other sections.

At the top right, there are buttons for "Save and check", "Fetch bibliography by DOI (<http://www.doi.org/>)", and "Fetch". There are also radio buttons for "Pubmed" and "crossref".

The main area shows the "Your CIF File contents:" section, which contains the following CIF code:

```
data_global  
loop  
publ_author_name  
'Sabah, Shahulhameed'  
'Lee, Chen-Shiang'  
'Hwang, Wen-Shu'  
'Lin, Ivan J. B.'  
publ_section_title  
;  
Facile C-N Bond Cleavage Promoted by Cuprous Oxide: Formation  
of C-C-Coupled Bimidazole from Its Methylene-Bridged Congener  
;  
journal_issue 2  
journal_name_full Organometallics  
journal_page_first 290  
journal_volume 29  
journal_year 2010  
data_714906  
chemical_formula_sum 'C16 H20 Cl4 Cu2 N8'  
chemical_formula_weight 593.28
```

COD validavimo ir deponavimo svetainė



The screenshot shows a Mozilla Firefox window with the title "Crystallography Open Database: CIF Validator - Mozilla Firefox". The address bar displays the URL <http://www.crystallography.net/store.php?f=0&CODSESSION=ZY0lg8DU9KTyEl-KIIS.gr05404>. The page content is the "Crystallography Open Database Validation and Deposition Interface". A navigation menu at the top includes "Log in", "Upload a file", "Validate data", "Deposit structures", and "Finish". Below the menu, a button labeled "Deposit to COD all valid files" is highlighted. A table lists a single file entry:

File	Status	Actions
om9010406_si_002.cif	valid	Edit Deposit to COD

A message below the table states: "File [om9010406_si_002.cif] is correct".

RDA Node Lithuania yra projekto „RDA Europe 4.0“ finansuojamo ES bendroios mokslo tyrimų ir inovacijų programos „Horizontas 2020“ dėl (suriarie nr. 777388)

1. Motyvacija – kodėl duomenų bazės?
2. FAIR principai
3. Atvirų mokslo resursų tipai
4. COD duomenų bazė
5. Mokslo duomenų pateikimas
 - ▶ Identifikatoriai
 - ▶ Metaduomenys
 - ▶ Homogeniškumas
 - ▶ Pasiekiamumas
 - ▶ Formatai
 - ▶ Semantika
 - ▶ Pilnumas
 - ▶ Kokybės kriterijai
 - ▶ Programos

Atkartojami skaitmeniniai tyrimai reikalauja dokumentuotų, prieinamų kompiuterinių programų.

- ▶ Pilna atkartojamumą galima pasiekti **tik** naudojant atviro kodo programas (F/LOSS);
- ▶ Daugybė duomenų apdorojimo programų prieinamos su Atviro kodo licencijomis”
 - ▶ Octave, R, Perl, Python, Make, MySQL, MariaDB, cod-tools, ...

Reziume

“Take-home message”

- ▶ Mokslo ir visuomenės labui – skelbkime duomenis **atvirai** ir pagal **FAIR** principus;
- ▶ Užtikrinkime **stabilius unikalius identifikatorius** (juk žinome, kaip!);
- ▶ Užtikrinkime **atvirus, suderinamus formatus**;
- ▶ Užtikrinkime **mokslo srities kokybę**;
- ▶ Siekime **pilno** duomenų pateikimo **mašina skaitomame** pavidaile;

VU GMC
Biotechnologijos i-tas

Virginijus Šikšnys
(*skyriaus vedėjas*)

Andrius Merkys
Antanas Vaitkus
Algirdas Grybauskas
Aleksandras
Konovalovas

COD Patarėjų taryba

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

RDA Node Lithuania yra projekto „RDA Europe 4.0“, finansuojamo ES bendrosios mokslinių tyrimų ir inovacijų programos „Horizontas 2020“ dalis (sutarties nr. 777388)

Nuorodos I



Gražulis S, Chateigner D, Downs RT, Yokochi AFT, Quirós M, Lutterotti L, et al. (2009) Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography* 42:726–729, DOI 10.1107/S0021889809016690, URL <http://dx.doi.org/10.1107/S0021889809016690>



Gražulis S, Daškevič A, Merkys A, Chateigner D, Lutterotti L, Quirós M, et al. (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, URL <http://nar.oxfordjournals.org/content/40/D1/D420.abstract>



Hall SR, Allen FH, Brown ID (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A* 47:655–685, DOI 10.1107/S010876739101067X, URL <http://dx.doi.org/10.1107/S010876739101067X>



Li CY, Mao X, Wei L (2008) Genes and (common) pathways underlying drug addiction. *PLoS Computational Biology* 4(1):e2, DOI 10.1371/journal.pcbi.0040002, URL <http://dx.doi.org/10.1371/journal.pcbi.0040002>

Nuorodos II

-  Long F, Nicholls RA, Emsley P, Gražulis S, Merkys A, Vaitkus A, et al. (2017a) ACEDRG: A stereo-chemical description generator for ligands. *Acta Crystallographica Section D* 73(2):112–122, DOI 10.1107/S2059798317000067, URL <https://doi.org/10.1107/S2059798317000067>
-  Long F, Nicholls RA, Emsley P, Gražulis S, Merkys A, Vaitkus A, et al. (2017b) Validation and extraction of stereochemical information from small molecular databases. *Acta Crystallographica Section D* 73(2):103–111, DOI 10.1107/S2059798317000079, URL <https://doi.org/10.1107/S2059798317000079>
-  Rauber A, Asmi A, van Uytvanck D, Pröll S (2016) Identification of reproducible subsets for data citation, sharing and re-use. URL <https://tinyurl.com/y7o7o8m4>
-  Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. (2016) The FAIR guiding principles for scientific data management and stewardship. *Scientific Data* 3(1), DOI 10.1038/sdata.2016.18, URL <https://doi.org/10.1038/sdata.2016.18>

Nuorodos III

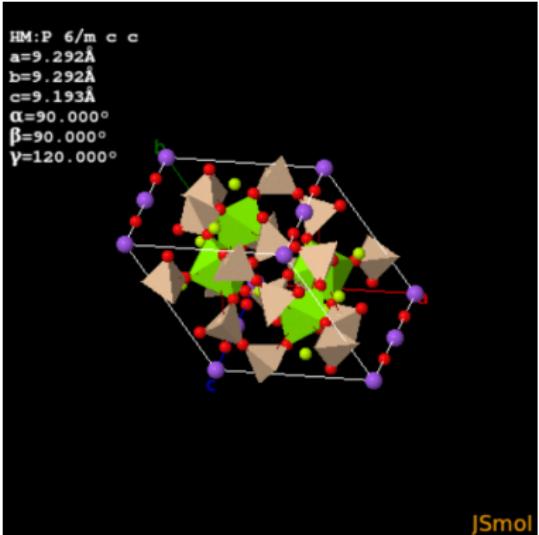


Zheng Y, Posfai J, Morgan RD, Vincze T, Roberts RJ (2008) Using shotgun sequence data to find active restriction enzyme genes. Nucleic Acids Research 37(1):e1–e1, DOI 10.1093/nar/gkn883, URL <https://doi.org/10.1093/nar/gkn883>

Dėkoju už dėmesį!



<http://en.wikipedia.org/wiki/Emerald>



<http://www.crystallography.net/5000095.html>