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

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

Turinys

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

Turinys

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

Paprastos užklausos

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

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

Paprastos užklausos

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

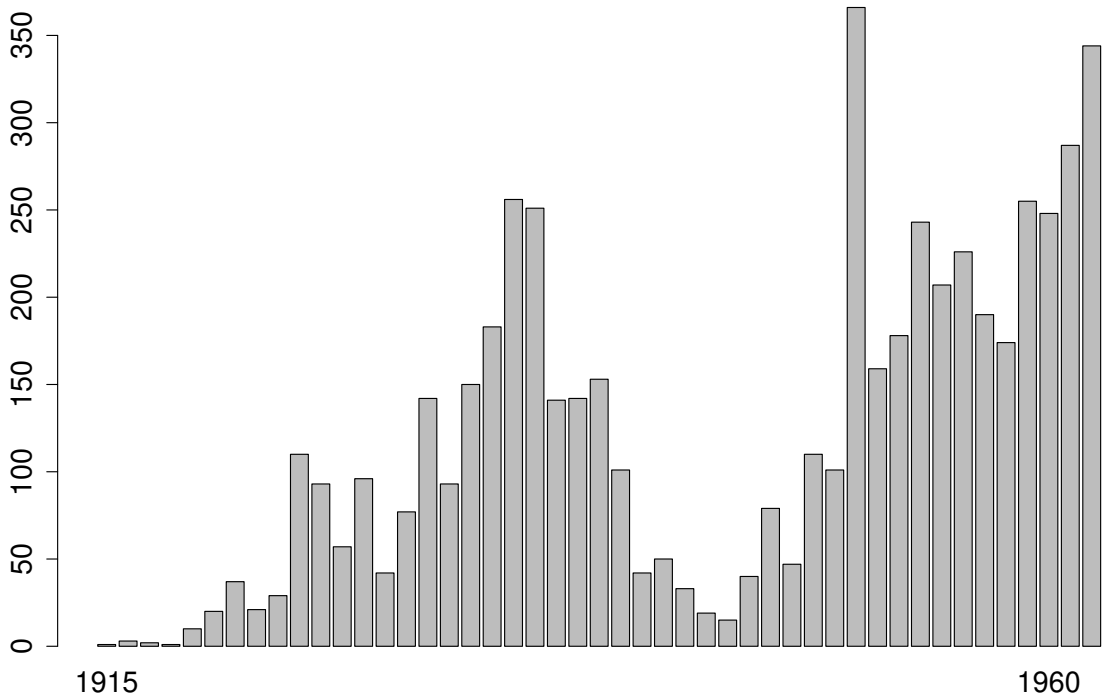
```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```

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

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
```

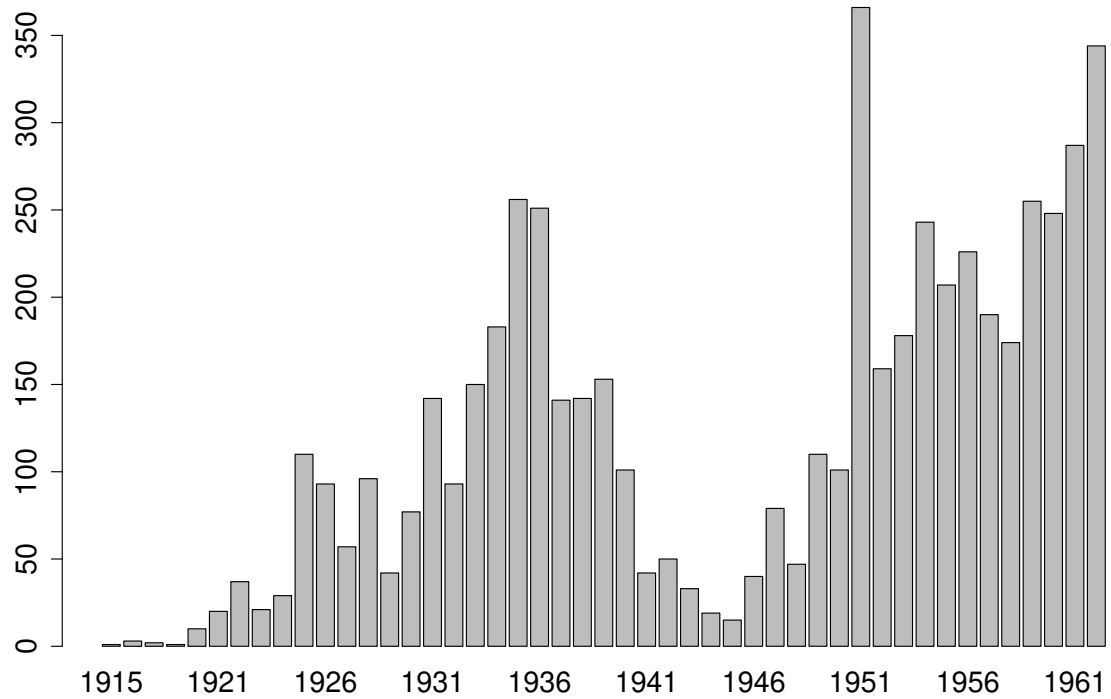


RDA Node Lithuania yra projekto „RDA Europe 4.0“, finansuojamo ES bendrosios mokslinių tyrimų ir inovacijų programos „Horizontas 2020“, dalis (sutarties nr. 777388)

Paprastos užklausos

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

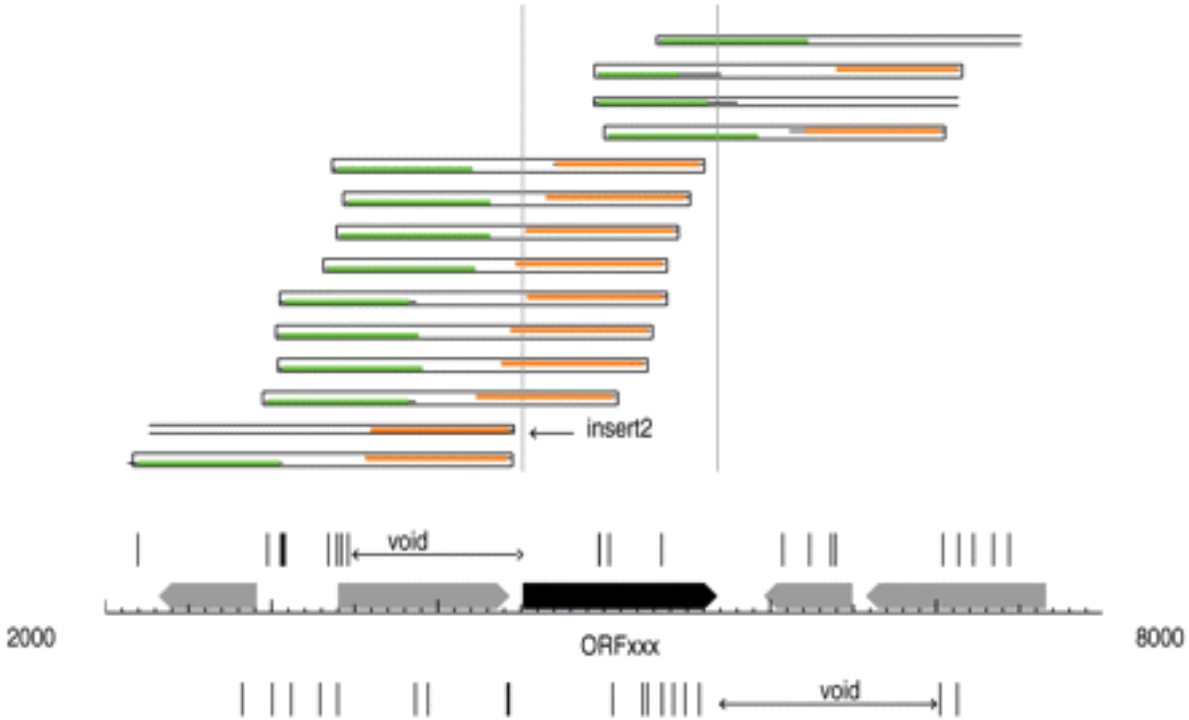
```
SELECT count(*) AS nr, year FROM data
WHERE year IS NOT NULL AND
GROUP BY year ORDER BY year DESC
```



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

Atradimai „žaliuose“ duomenyse

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



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

Duomenų svarba

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

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

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

Turinys (2)

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

FAIR duomenų mainų principai

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

- **F**indable: Randami (automatinėmis priemonėmis!)
- **A**cessible: Prieinami (automatinėmis priemonėmis!)
- **I**nteroperable: Suderinami (su įvairiomis programomis!)
- **R**eusable: Pakartotinai panaudojami (įvairiems, nenumatytiems tikslams!)

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

Turinys (3)

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

Įvairūs duomenų mainų sprendimai

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

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

Įvairūs 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.
   -
   -
   -
   -

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

Įvairūs duomenų mainų sprendimai

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

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

Įvairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693 – stendinis pranešimas (PDF)...
   - Zenodo https://zenodo.org/record/3841841 – susieti COVID-19 duomenys (TTL)...
2.
   -
   -
   -
   -

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

Įvairūs duomenų mainų sprendimai

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

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

Įvairūs duomenų mainų sprendimai

1. Bendri duomenų archyvai (Zenodo, Data Dryad, MIDAS, ...);
   - Zenodo https://doi.org/10.5281/zenodo.3560693 – stendinis 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

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

Turinys (4)

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

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;

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

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.

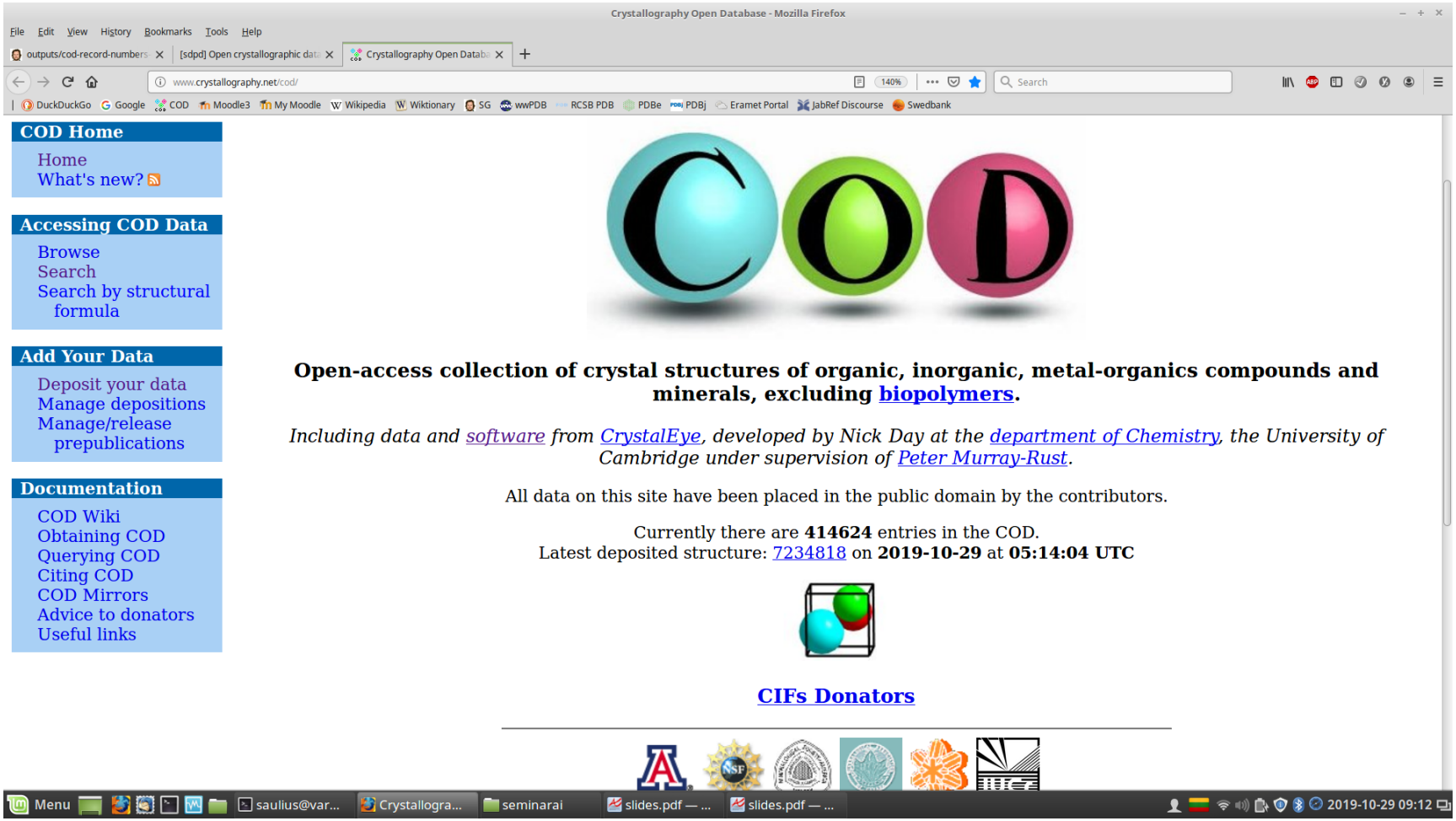
gemstonede (Dr. Michael BERNDT) Fri Feb 14, 2003 1:26 pm

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

Po 17 metų ... :)

The Crystallography Open Database

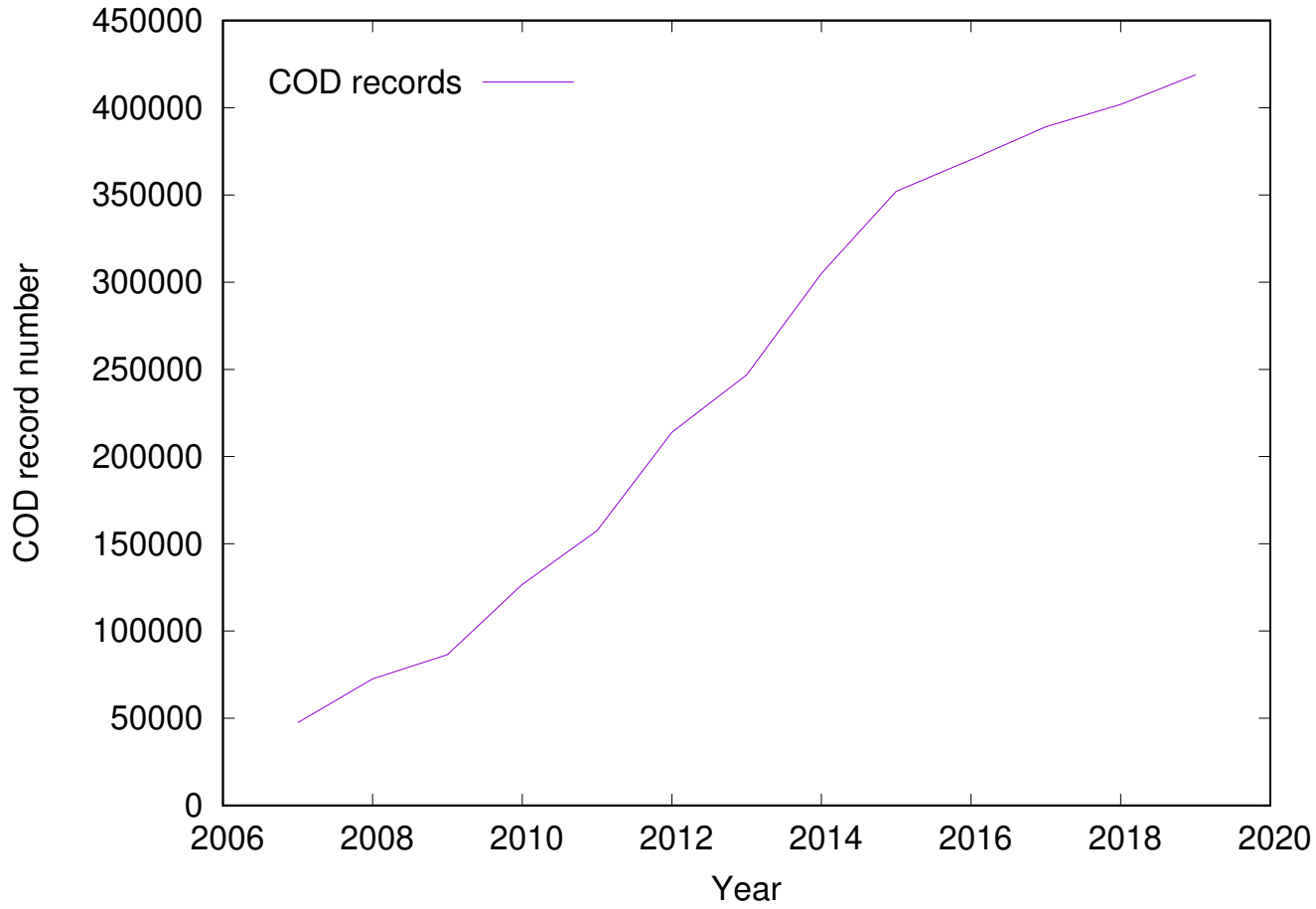
http://www.crystallography.net/cod



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

COD tvarumas ir augimas

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



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

Turinys (5)

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

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

Turinys (5)

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

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

Identifikatorių tipai

1. Bendras (centralizuotas arba paskirstytas) registras:
   - 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);

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

Turinys (6)

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

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

Metaduomenys

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

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

Turinys (7)

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

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

Homogeniškumas

- Įrašai 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) # ...
```

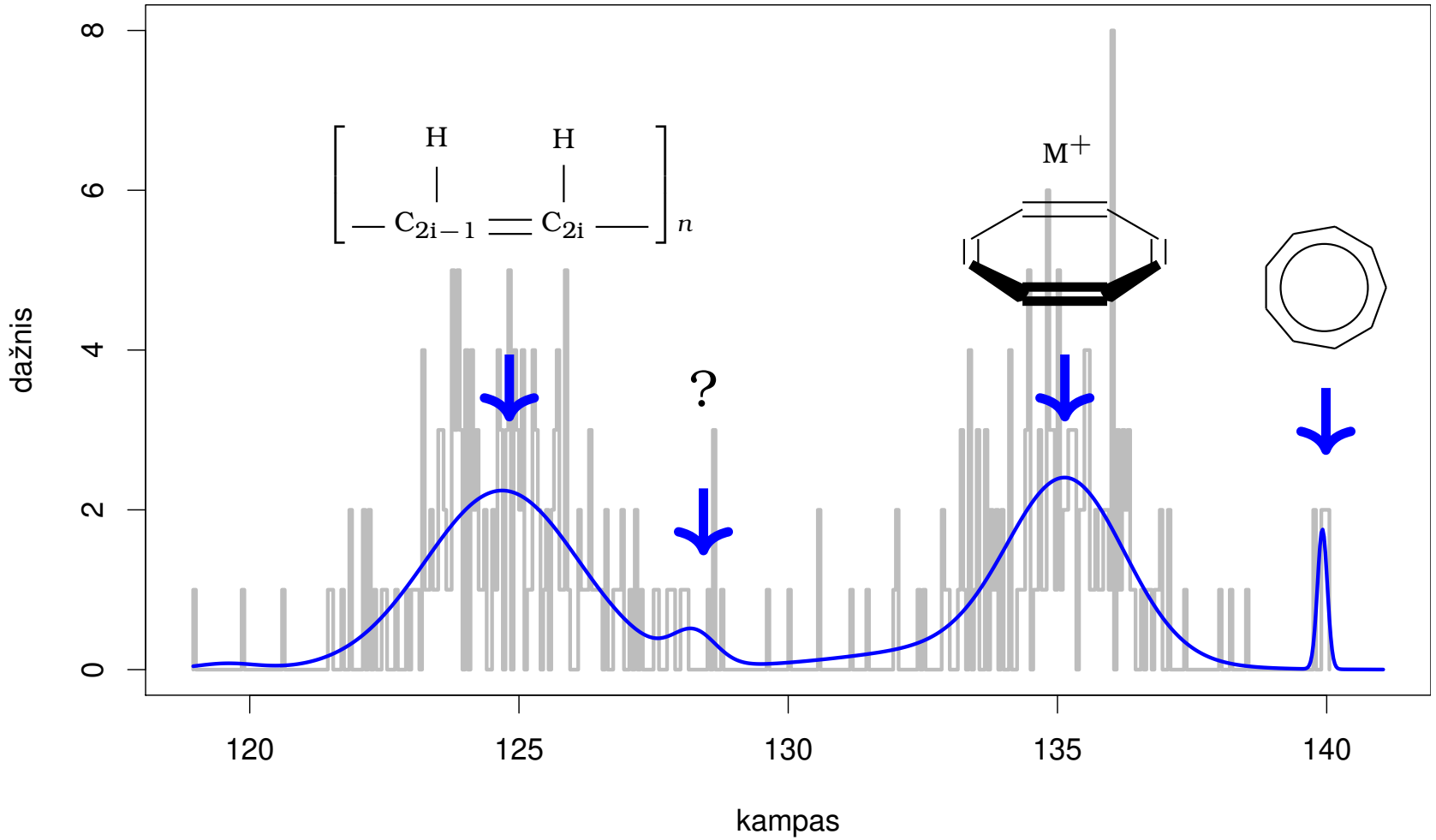
cod/2103669-head.cif:

```
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) # ...
```

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

Sudėtingų bruožų atpažinimas

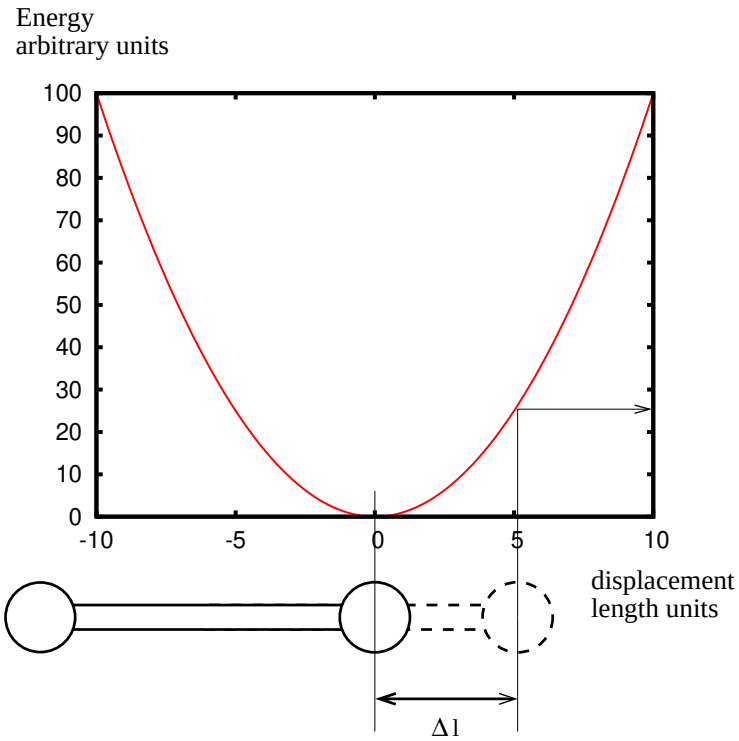
Polienų grandinės kampai iš COD

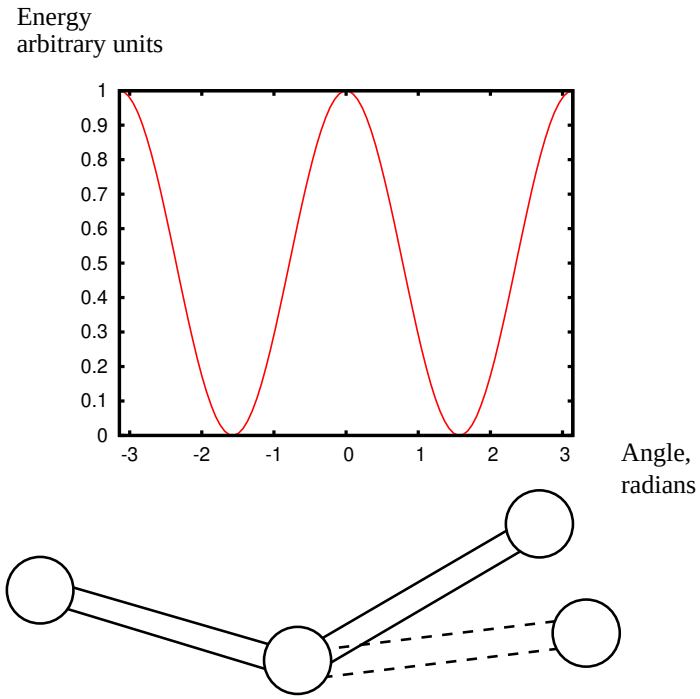


A. Merkys, daktaro disertacija

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

Molekulių mechaninės savybės



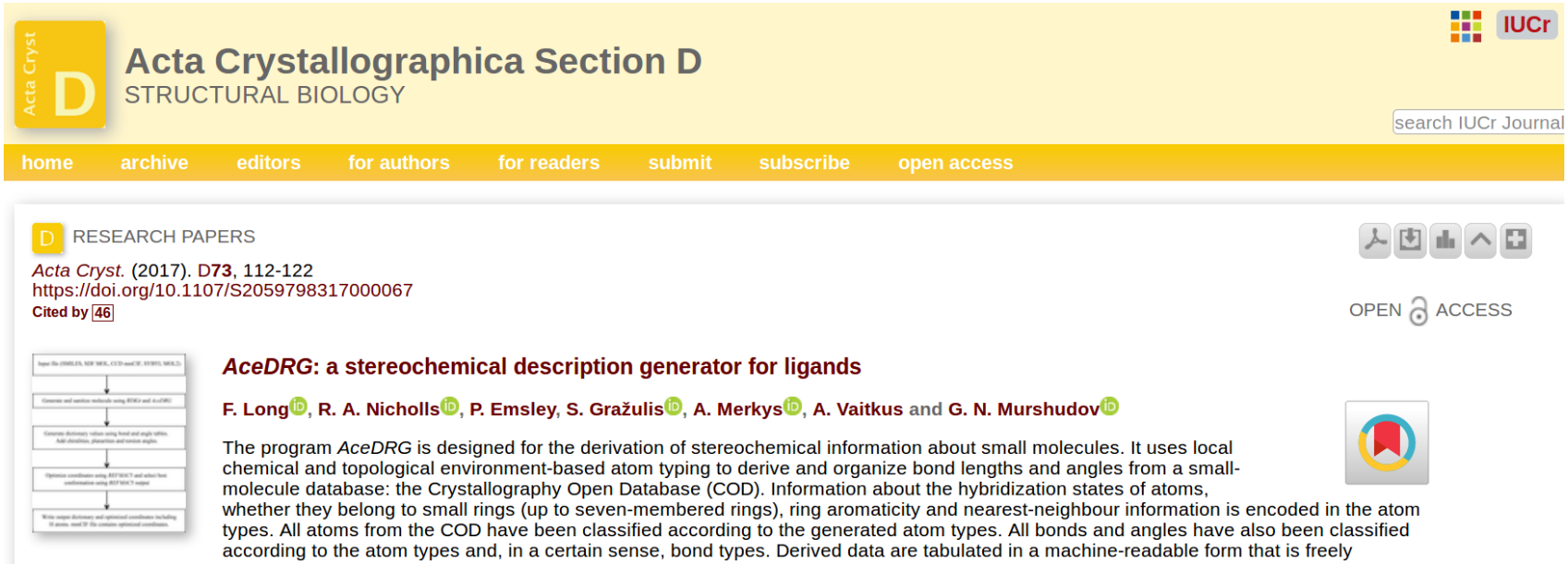


$$E = \frac{1}{2}k\Delta l^2 \qquad p(\Delta l) \sim e^{-\frac{E}{k_B T}} = e^{-\frac{\Delta l^2}{2\sigma^2}} \qquad E = k(1+\cos(n\Delta\varphi)), \text{if } n > 0$$

- Kristalų statistika leidžia nustatyti jėgas, veikiančias atomus molekulėse.

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

Naujos apribojimų bibliotekos



Acta Crystallographica Section D
STRUCTURAL BIOLOGY

IUCr

search IUCr Journal

home archive editors for authors for readers submit subscribe open access

RESEARCH PAPERS

Acta Cryst. (2017). D**73**, 112-122
https://doi.org/10.1107/S2059798317000067
Cited by 46

OPEN ACCESS

***AceDRG*: a stereochemical description generator for ligands**

F. Long, R. A. Nicholls, P. Emsley, S. Gražulis, A. Merkys, A. Vaitkus and G. N. Murshudov

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)]

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

Turinys (8)

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

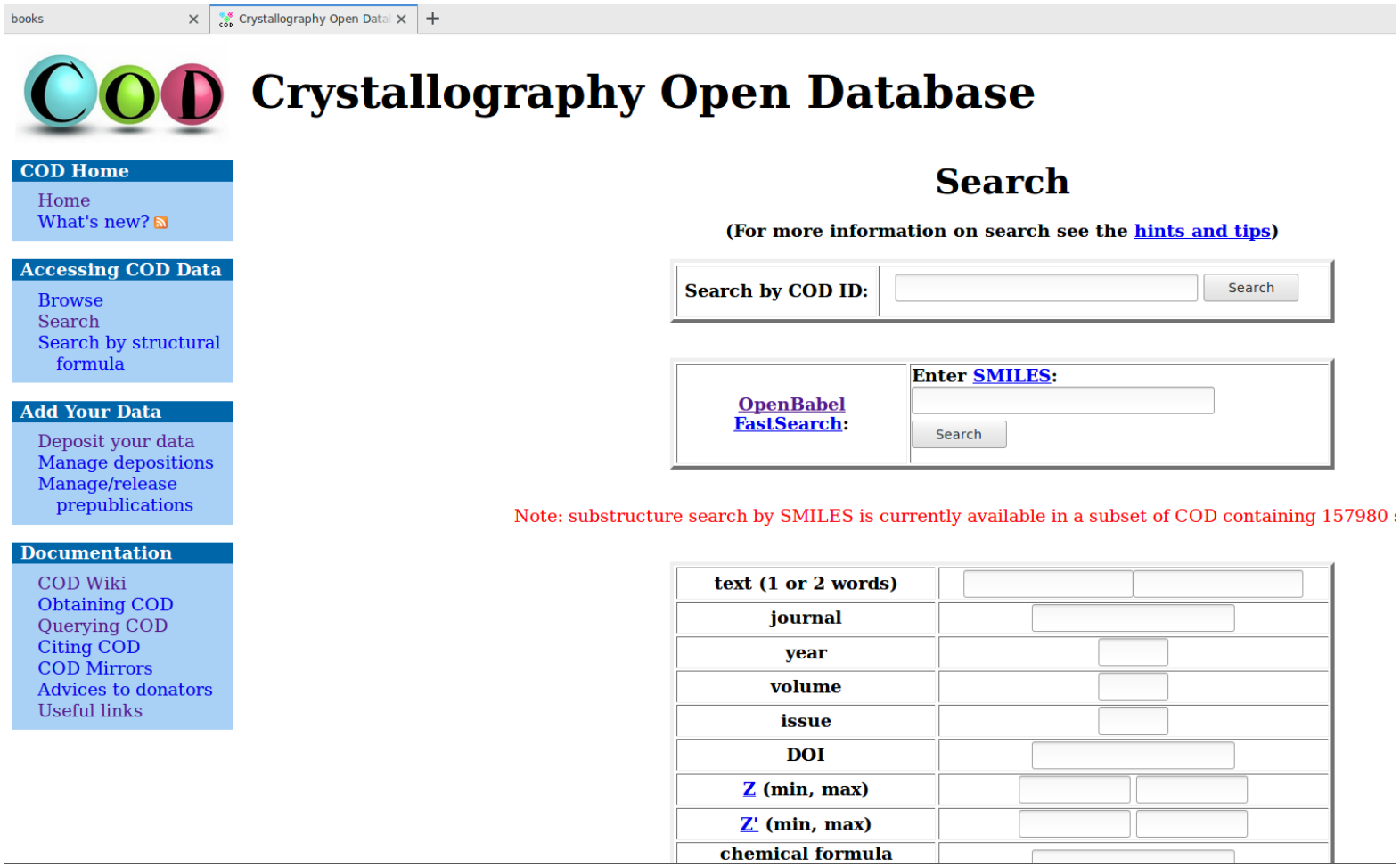
Pasiekiamumas

- 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)]

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

COD paieškos forma

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



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

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 kalbą** 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'
    ```

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

Turinys (9)

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

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), ...;

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

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:KMoO4: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)
```

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

Turinys (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

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

Semantika

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

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

Kontroliuojami žodynai

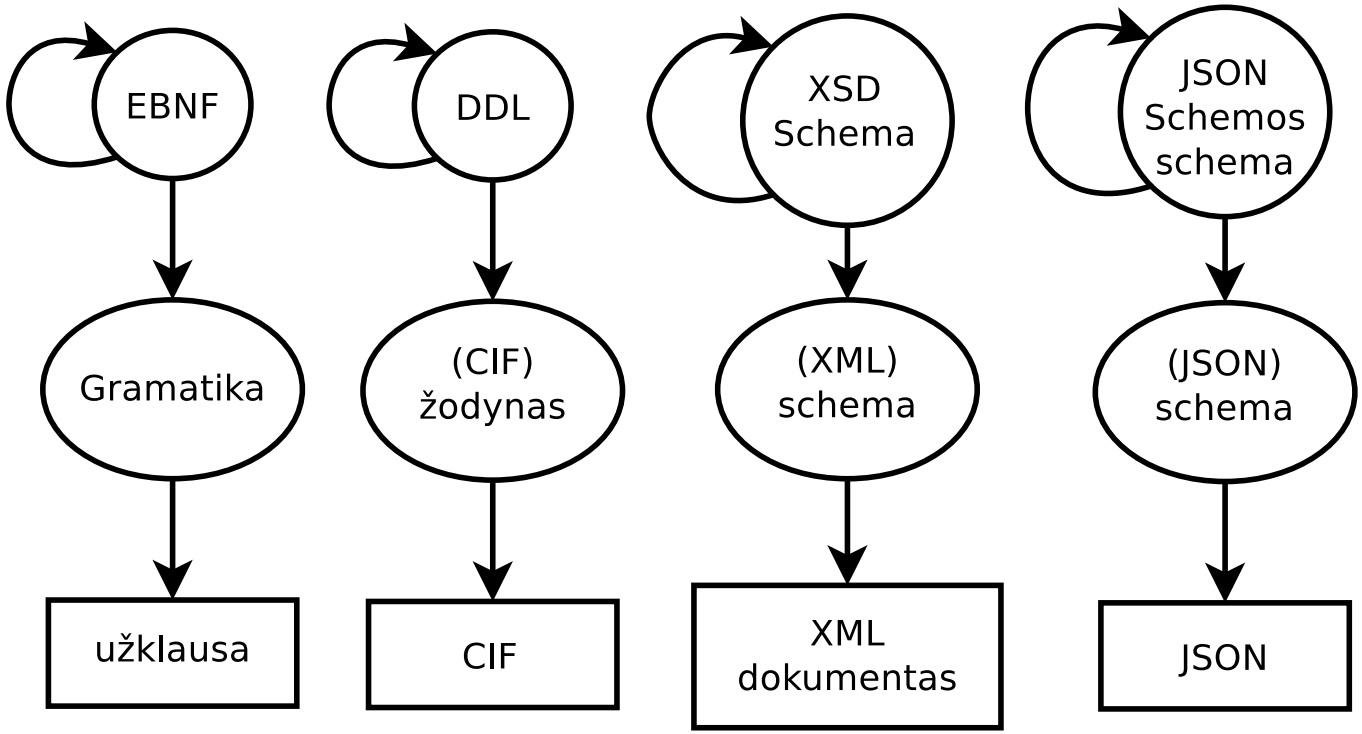
examples/dictionaries/cif-core-example.cif:

```
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_.
;
```

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

Save aprašantys duomenys

Visada užtenka trijų lygių duomenų apribojimams aprašyti!



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

Turinys (11)

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

Pilnumas

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

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

Turinys (12)

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

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

Kokybės kriterijai

- 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;

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

COD duomenų tikrinimas

COD duomenų tikrinimo politika:

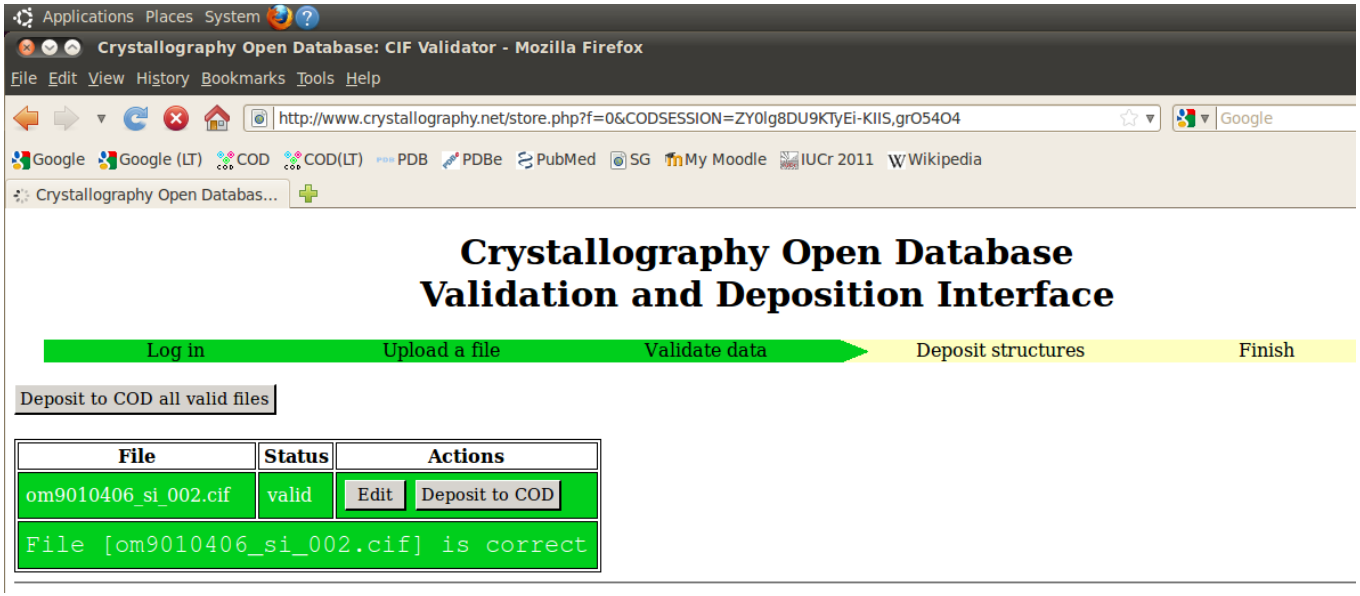
1. Sintaksės patikrinimas:
   ```
   $ cifparse 7234818.cif
   ```
2. Semantinis patikrinimas (pagal žodynus):
   ```
   $ cif_validate -D cif_core.dic 7234818.cif
   ```
3. Specifiniai COD duomenų bazės testai:
   ```
   $ cif_cod_check 7234818.cif
   ```

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

COD validavimo ir deponavimo svetainė

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

COD validavimo ir deponavimo svetainė



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

Turinys (13)

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

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

Programos duomenų apdorojimui

Atkartojami skaitmeniniai tyrimai reikalauja dokumentuotų, prieinamų kompiuterinių programų.

- Pilną 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, ...

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

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** pavidale;

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

Padėkos

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)

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

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

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

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

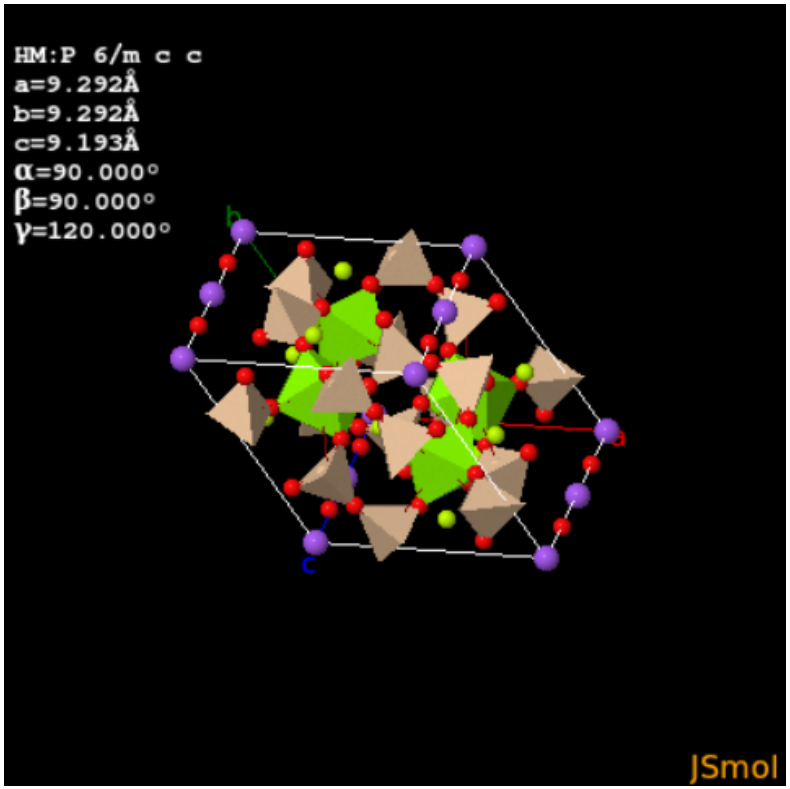
Nuorodos III

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

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

Dėkoju už dėmesį!

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



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

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