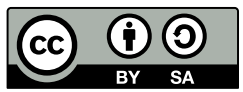
Atviros susietos duomenų bazės kalnakasybos pramonėje

Saulius Gražulis

Kaunas, OpenCon 2017

Vilniaus universiteto Biotechnologijos institutas

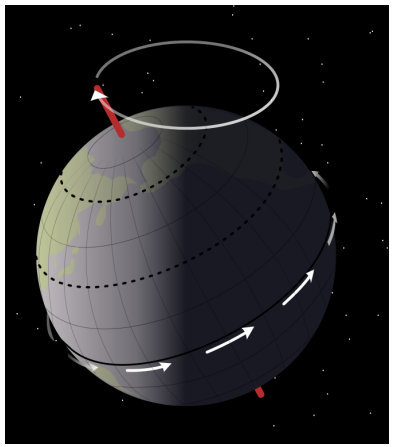


Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Duomenų svarba

Hiparchas (apie 190 – 120 p. m. e)

- išmatavo Spikos, Regulo ir kitų ryškių žvaigždžių koordinates
- Palygino rezultatus su savo pirmtakų Timoarcho ir Aristilo, gyvenusių ≈**100** metų anksčiau, rezultatais
- aptiko reiškinį, kurį mes dabar vadiname *lygiadienių precesija*



NASA, Viešoji nuosavybė

(Wikipedia, žr taip pat straipsnius apie Timoarchą ir Aristlą)

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Duomenys ir dirbtinio intelekto sistemos geologijoje

[Hart and Duda, 1977]

October 20, 1977

PROSPECTOR -- A Computer-Based Consultation
System for Mineral Exploration

by

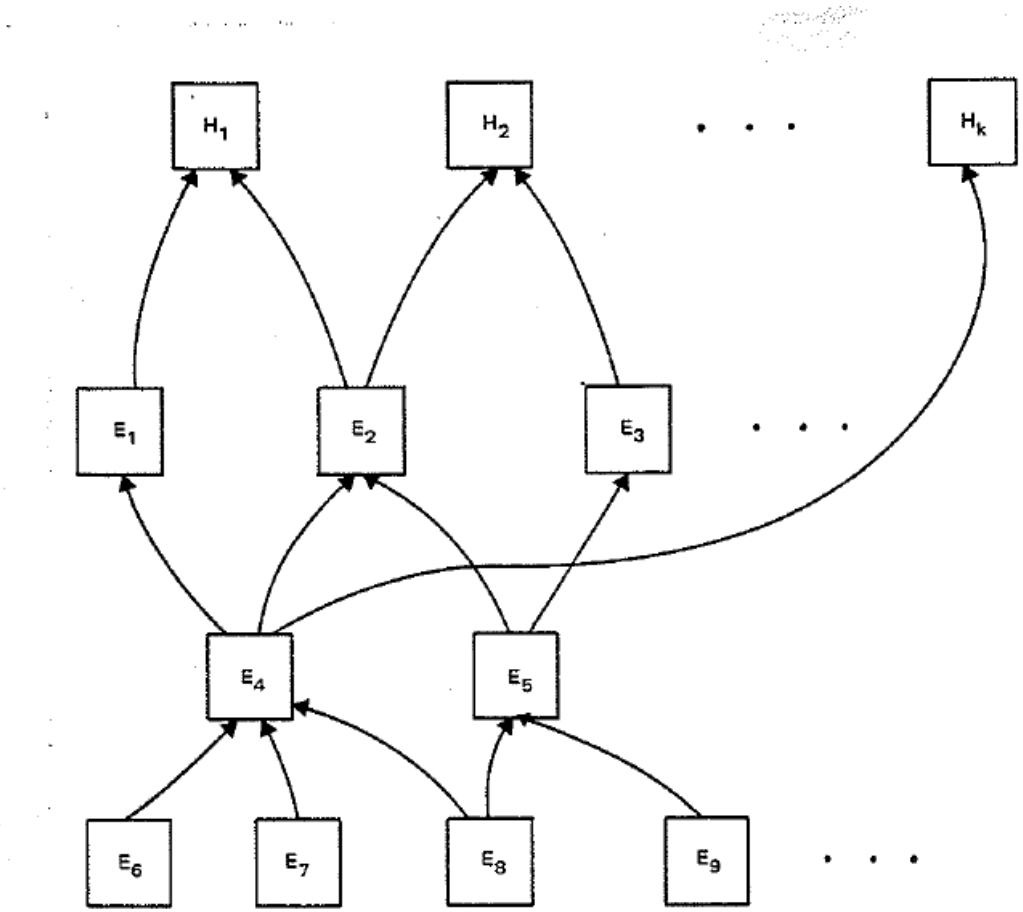
P. E. Hart and R. O. Duda

Artificial Intelligence Center
SRI International
Menlo Park, California 94025

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Programos PROSPECTOR išvadų generavimo tinklas

[Hart and Duda, 1977]



Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Duomenys SOLSA projekte

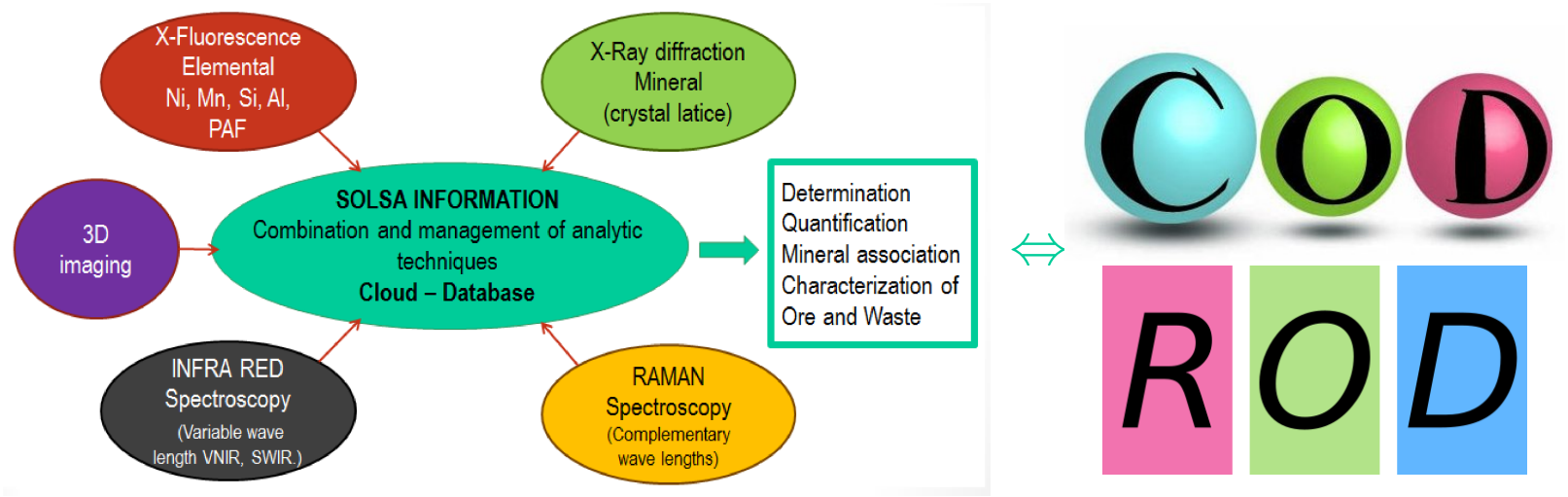
Discover SOLSA

http://solsa-mining.eu/

- Kristalų struktūros (COD)
- Ramano spektrai (ROD)
- Hiperspektriniai vaizdai (HOD)

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

SOLSA projektas, COD ir ROD



COD ir kitos atviros duomenų bazės bus naudojamos:

- mineralų identifikacijai;
- projekto duomenų publikavimui.

SOLSA duomenų srauto diagramą parengė Monique Le Guen, ERAMET.

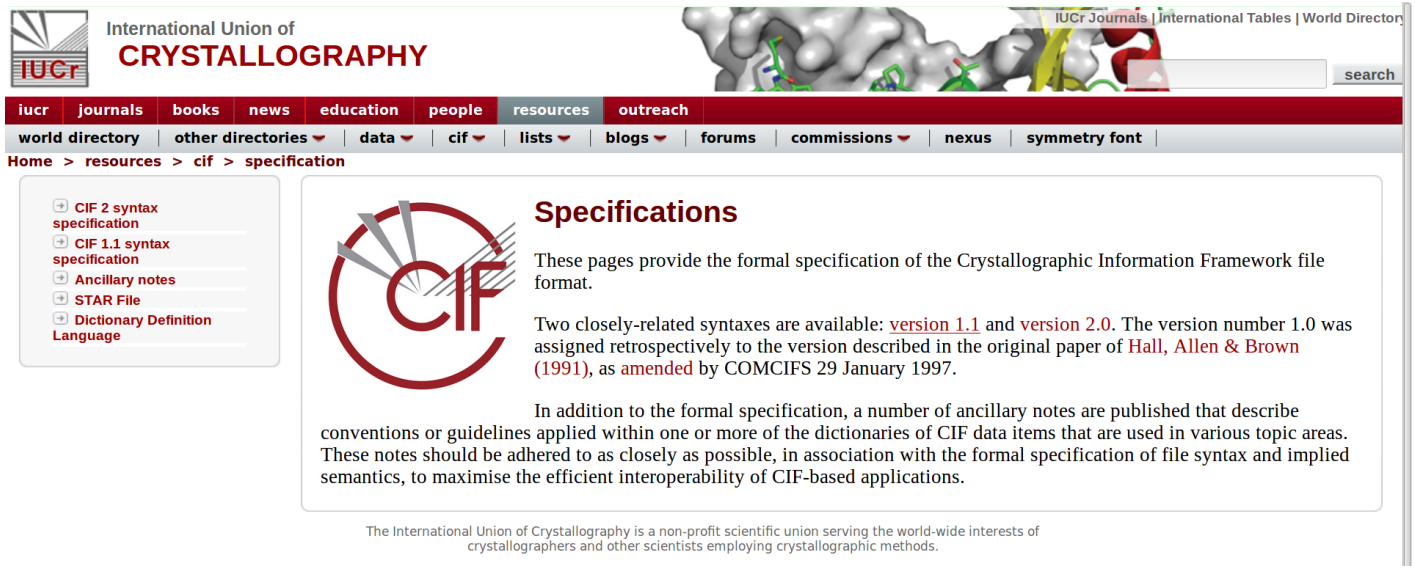
Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Reikalavimai ilgalaikiam duomenų archyvavimui ir pakartotiniam panaudojimui

- Nepriklausymas nuo konkrečios platformos
  - Tekstu paremti formatai (pageidautina standartinių koduočių: ASCII, UTF-8)
- Nepriklausymas nuo konkrečios programinės įrangos
- Galimybė naudoti tinkle
  - Standartiniai atviri protokolai (W3C http)
  - Standartiniai atviri failų formatai (JSON, XML, CIF).
  - REST architektūros serveriai
- Mašina skaitoma semantika
  - Žodynai (CIF), schemos (XML, JSON)
- Tvarumas
  - Nuolatiniai identifikatoriai
  - Atvirų duomenų principai
  - Aptinkamumas, suderinamumas, sąveikumas, tinkamumas naudoti (FAIR)

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Duomenų mainai kristalografijoje



[Hall et al., 1991]

Kristalografinis duomenų mainų failas/karkasas (The Crystallographic Interchange File/Framework, CIF):

- Suteikia standartines galimybes keistis duomenimis;
- Tinkamas duomenų archyvavimui ir publikavimui;
- Vystomas Tarptautinės kristalografų sąjungos (International Union of Crystallography, IUCr);

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

CIF panaudojimas mokslo duomenims

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_angle_alpha                90.0
_cell_angle_beta                 90.0
_cell_angle_gamma                90.0
_cell_formula_units_Z            1
_cell_length_a                   3.9998(8)
_cell_length_b                   3.9998(8)
_cell_length_c                   4.0180(8)
```

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Prižiūrimi žodynai

Controlled vocabularies

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

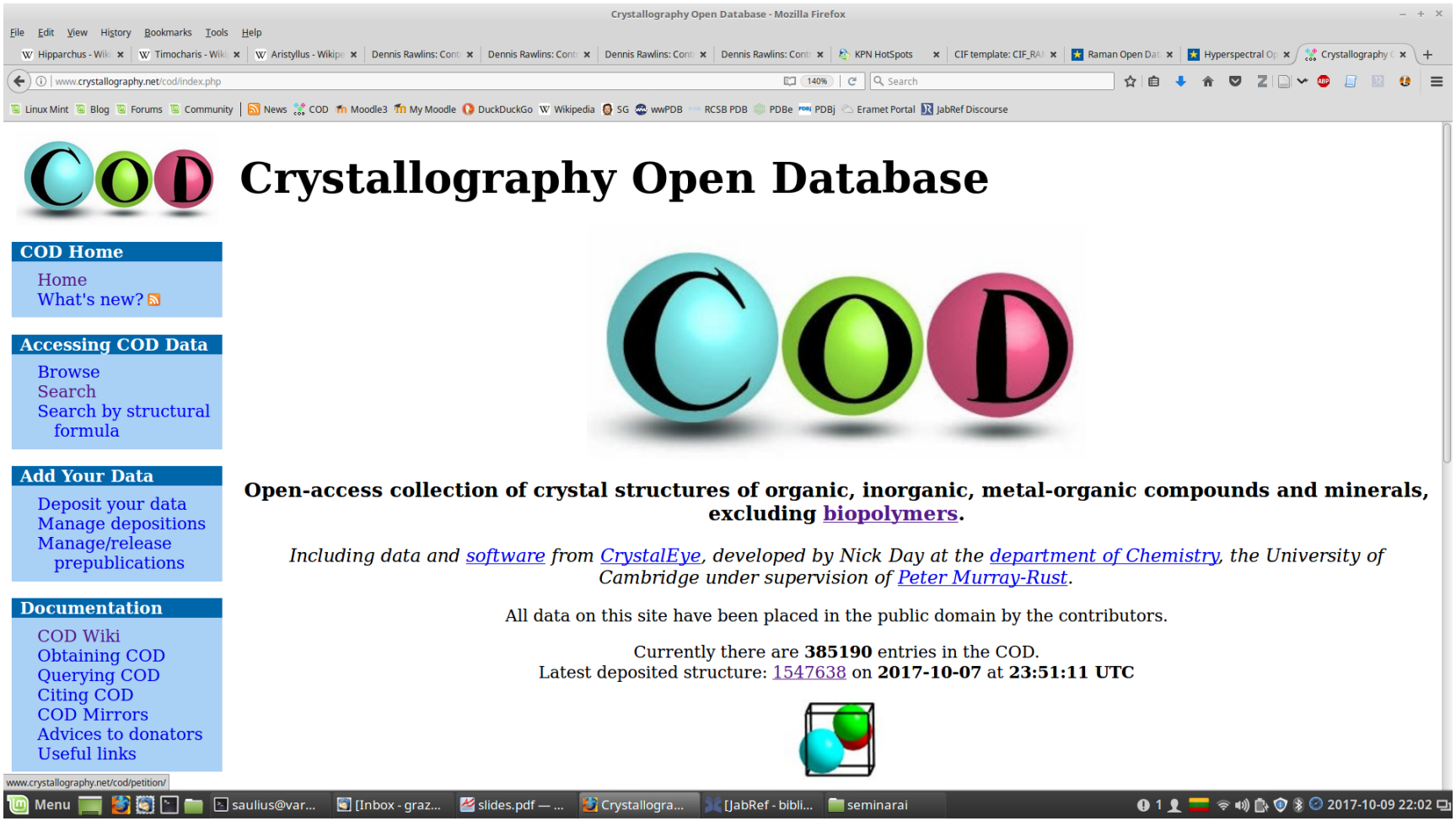
```
data_cell_length_
    loop_ _name                  '_cell_length_a'
                                 '_cell_length_b'
                                 '_cell_length_c'
    _category                      cell
    _type                          numb
    _type_conditions               esd
    _enumeration_range             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_.
;
```

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Duomenys kristalografijoje

Atvira kristalografinė duomenų bazė COD

http://www.crystallography.net/cod



Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Kristalo struktūros pavyzdys
COD
Sfaleritas

http://www.crystallography.net/cod/1525302.html

Crystallography Open Database

COD Home
- Home
- What's new?

Accessing COD Data
- Browse
- Search
- Search by structural formula

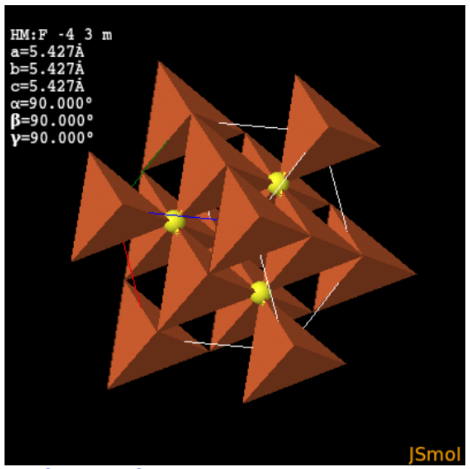
Add Your Data
- Deposit your data
- Manage depositions
- Manage/release prepublications

Documentation
- COD Wiki
- Obtaining COD
- Querying COD
- Citing COD
- COD Mirrors
- Advices to donators
- Useful links

Information card for entry 1525302

1525301 << **1525302** >> 1525303

Preview



Display in Jmol

Coordinates 1525302.cif

Coordinates 1525302.cif

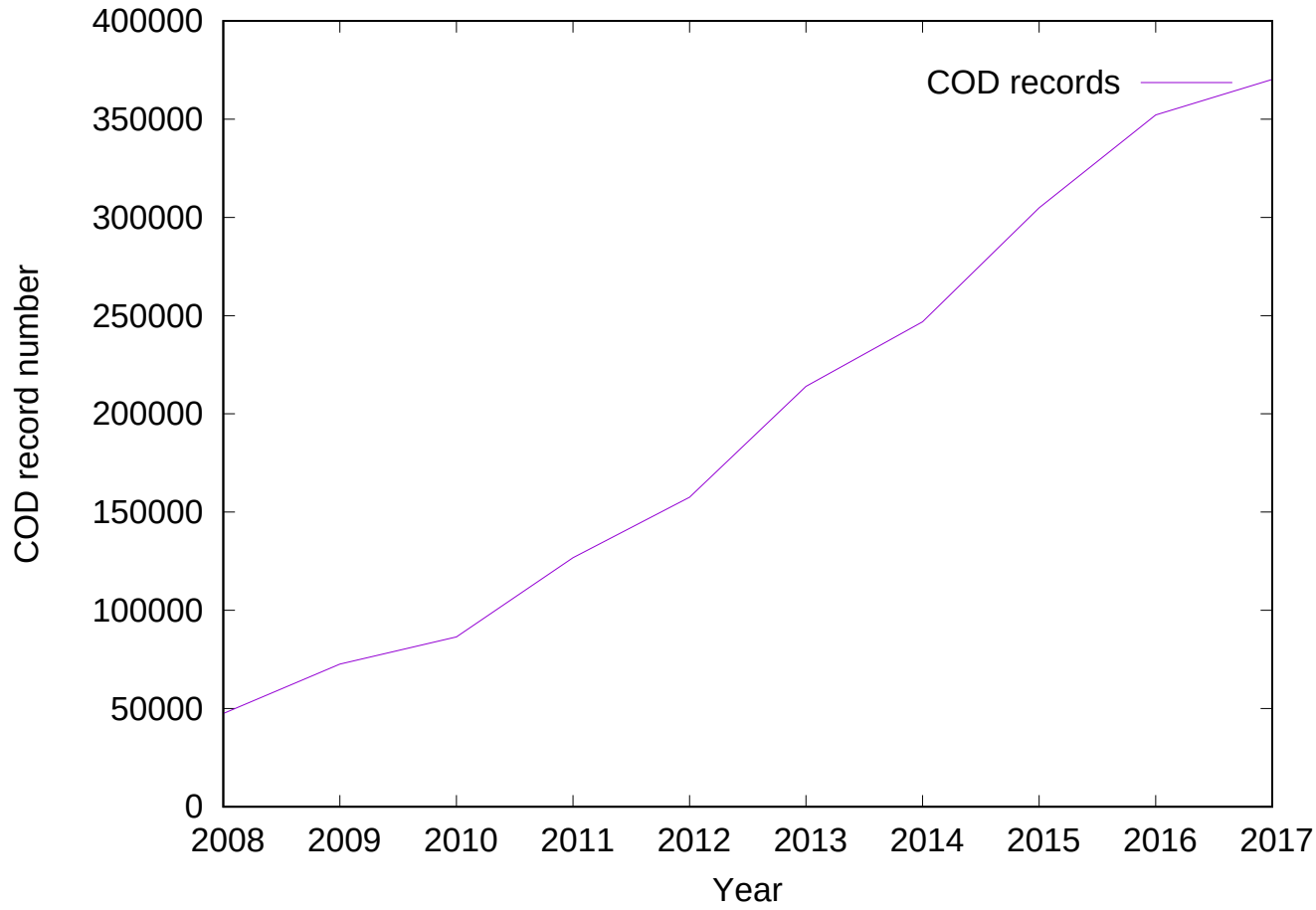
Structure parameters

Chemical name	(Fe0.2 Mn0.05 Zn0.75) S
Formula	Fe0.2 Mn0.05 S Zn0.75
Calculated formula	Fe0.2 Mn0.05 S Zn0.75
Title of publication	Unit-cell edges of natural and synthetic sphalerites
Authors of publication	Skinner, B.J.
Journal of publication	American Mineralogist
Year of publication	1961
Journal volume	46
Pages of publication	1399 - 1411
a	5.4272 Å
b	5.4272 Å
c	5.4272 Å
α	90°
β	90°
γ	90°
Cell volume	159.855 Å^3
Number of distinct elements	4
Hermann-Mauguin symmetry space group	F -4 3 m
Hall symmetry space group	F -4 2 3
Has coordinates	Yes
Has disorder	No
Has F_{obs}	No

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

COD duomenų bazės tvarumas

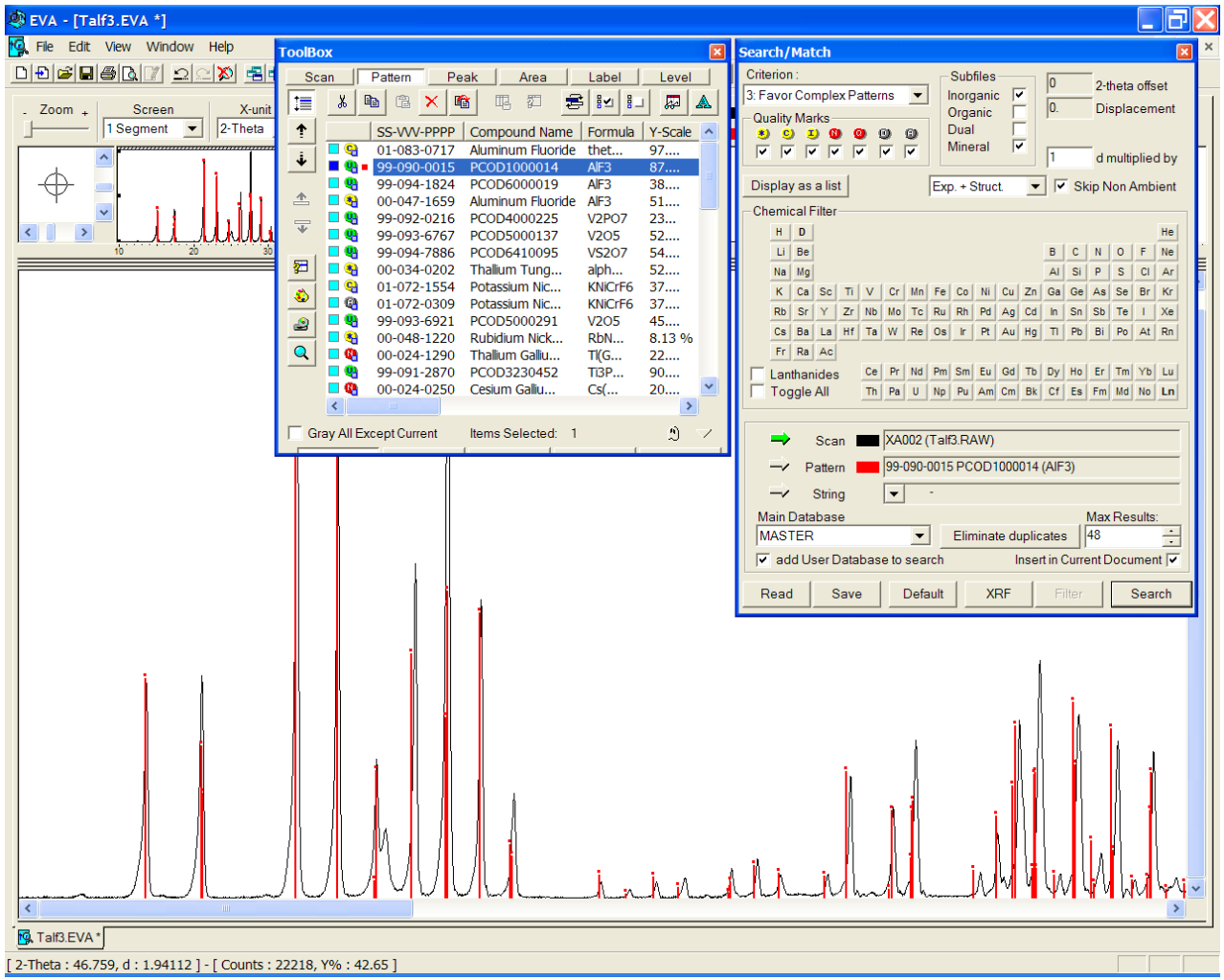
COD prieinama Internete jau 13 metų, per paskutinius 8 metus išaugo 7 kartus; šiuo metu talpina daugiau negu 385 000 įrašų (2017 spalis):



Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

COD ir PCOD duomenų bazių panaudojimas

Kristalinės medžiagos identifikavimas



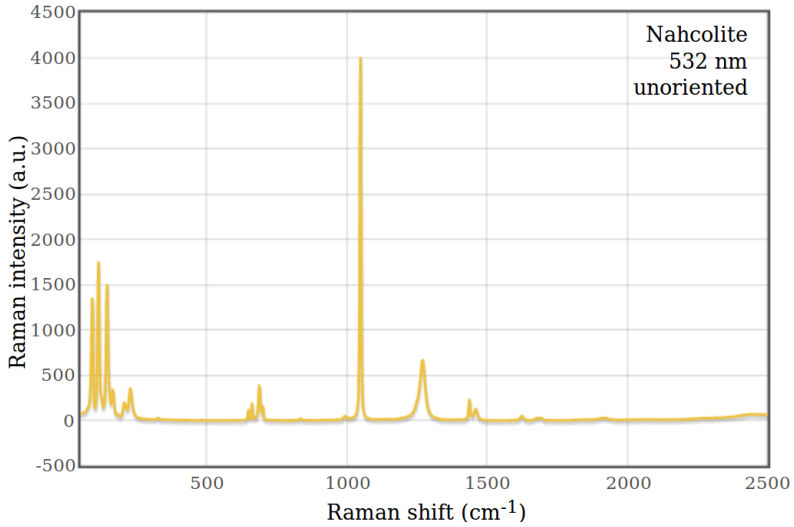
Medžiagos identifikavimas pagal Rentgeno spindulių sklaidymo intensyvumus.

Paveikslių parengė Armelis Le Bail ([Le Bail, 2008])

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Ramano spectroskopija

us-tech.co.za



ROD 3500101

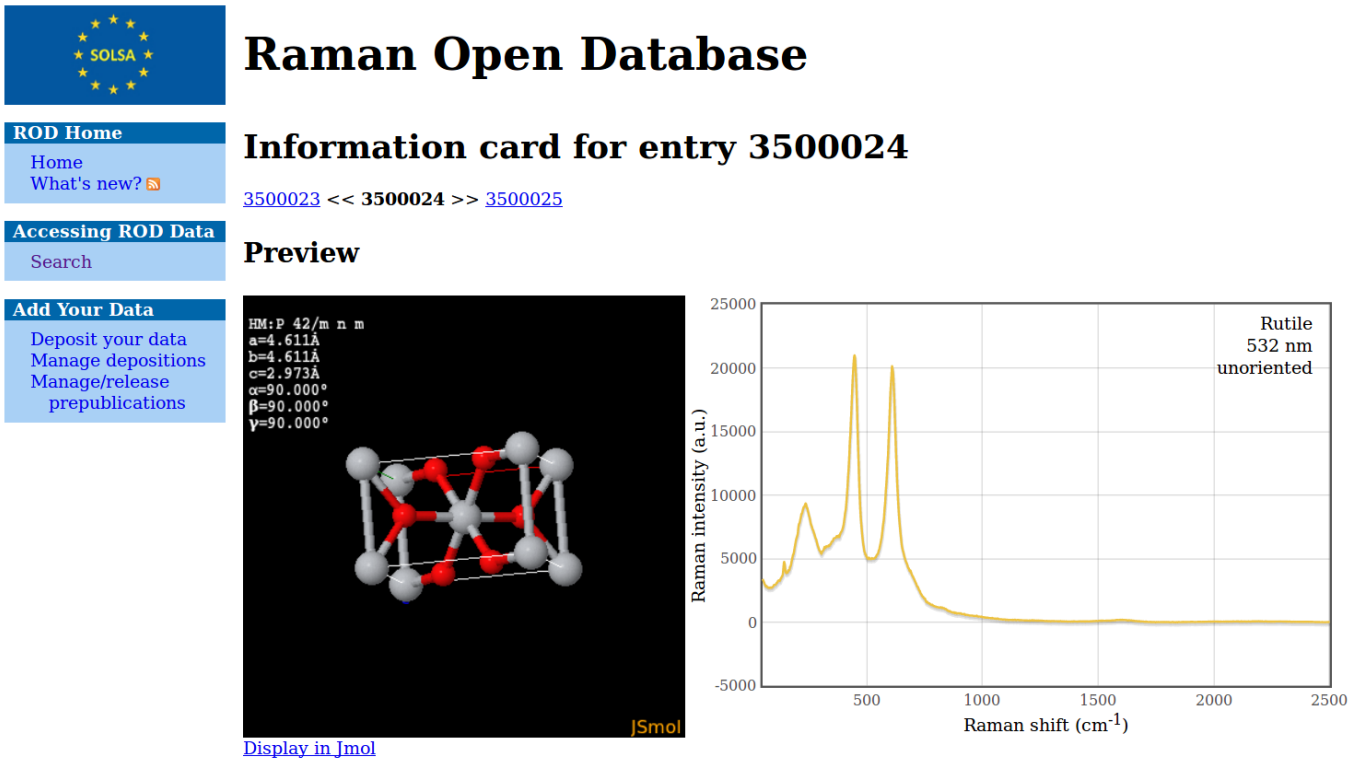
- labai greitas metodas
- reikalauja geros išsamios duomenų bazės

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Ramano spektroskopijos duomenys

Atvira Ramano spektrų duomenų bazė ROD

http://solsa.crystallography.net/rod



Duomenų įrašai įkelti į ROD Yassine El Mendili

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

ROD duomenų failai

ROD naudoja CIF sintaksę

examples/data/3500024-head.rod:

```
#------------------------------------------------------------------------------
#$Date: 2017-10-05 18:15:36 +0300 (Thu, 05 Oct 2017) $
#$Revision: 219 $
#$URL: svn://172.16.1.102/rod/cif/3/50/00/3500024.rod $
#------------------------------------------------------------------------------
#
# This file is available in the Raman Open Database (ROD),
# http://solsa.crystallography.net/rod/
#
# All data on this site have been placed in the public domain by the
# contributors.
#
data_3500024
loop_
_publ_author_name
'El Mendili, Y'
_publ_section_title
;
 SOLSA communication to ROD
;
_journal_name_full               'Personal communication to ROD'
_journal_year                    2017
_chemical_compound_source        'commercial powder Prolabo pur'
_chemical_formula_structural     'O2 Ti'
```

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

ROD žodynas

ROD naudoja prižiūrimą CIF žodyną

http://solsa.crystallography.net/rod/cif/dictionaries/cif_raman_0.1.1.dic
http://solsa.crystallography.net/rod/cif/dictionaries/cif_rod_0.1.0.dic

examples/dictionaries/raman-example.dic:

```
save__raman_measurement_device.direction_polarization
    _definition.id                '_raman_measurement_device.direction_polarization'
# ... some text omited for brevity ...
    _definition.update            2017-04-10
    _description.text
;
    The direction polarization of the measurement device.
;
# ...
    loop_
    _enumeration_set.state
    _enumeration_set.detail
    unoriented
;
 Unoriented.
;
    Z(XX)Z
;
 Laser polarized parallel to the x axis; analyzer set to pass the x axis
 polarized light.
;
```

ROD žodyną parengė Antanas Vaitkus

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

ROD žodynų semantinis versijavimas

- ROD žodynams taikomas semantinis versijavimas (http://semver.org/)
  - Klaidas pataisančios laidos (1.2.x) suderinamos tiek su ankstesnėmis, tiek su vėlesnėmis minorinėmis versijomis;
  - Minorinės versijos (1.x) suderinamos su senesnėmis minorinėmis;
  - Nesuderinami pakeitimai pažymimi kaip nauja pagrindinė (major) versija: (1.x $\rightarrow$ 2.x);

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

COD užklausų pavyzdžiai

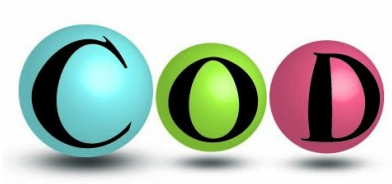
Web, REST, SQL

- Naudojant Žiniatinklio sąsają (nuoroda „search“):
  - http://www.crystallography.net/cod
  - http://solsa.crystallography.net/rod
  - http://solsa.crystallography.net/hod
- Naudojant **stabilias** nuorodas (URL) REST architektūros serveryje:
  - http://www.crystallography.net/cod/2000000.cif
  - http://solsa.crystallography.net/rod/3500021.rod
  - http://solsa.crystallography.net/rod/3500021.html
  - http://www.crystallography.net/cod/result?text=perovskite
- Naudojant SQL užklausų kalbą:
  - ```
    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'
    ```

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.
    ```

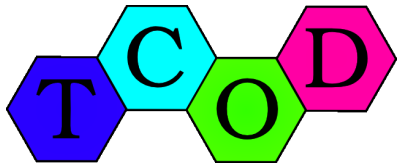
Atviros kristalografinės duomenų bazės

COD, TCOD, PCOD, MPOD, ROD, HOD...



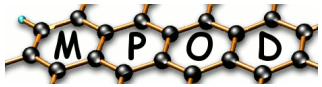
http://www.crystallography.net/cod

$>$ 385 000 įrašų (augs iki $> 10^6$?)



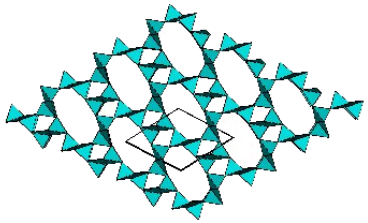
http://www.crystallography.net/tcod

$>$ 2500 įrašų (augs iki $> 10^7$?)



http://mpod.cimav.edu.mx/

$>$ 300 įrašų



http://www.crystallography.net/pcod

$>$ 10^6 įrašų (augs iki $> 10^8$?)

http://solsa.crystallography.net/rod/

$>$ 120 įrašų

HOD

http://solsa.crystallography.net/hod/

ruošiama...

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

COD prieinamumas

COD yra **visiškai atvira duomenų bazė**. Visi įrašai prieinami kaip Viešoji nuosavybė.

Suteikiami tokie prieigos metodai:

- Paieška žiniatinklio formoje;
- Duomenų nukėlimas naudojant stabilius identifikatorius (URI)
- REST sąsaja
- Galimybė nukelti **visus** COD duomenis

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Hiperspektrinių vaizdų duomenų bazė (HOD)

http://solsa.crystallography.net/hod

Dėl didelio rastrinių duomenų kiekio reikalingas „hibridinis“ duomenų saugojimo metodas:

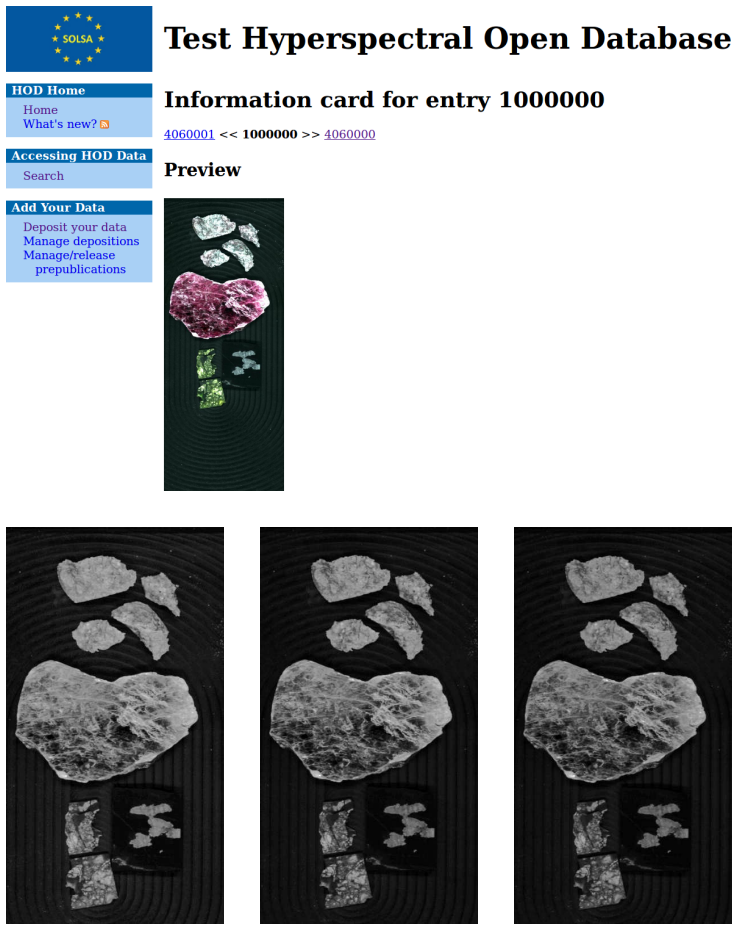
- Metaduomenys ir paveiksliuko antraštės saugomos CIF formatu;
- Paveiksliuko rastras saugomas kaip „žali“ dvejetainiai žinomo tipo duomenys;

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

HOD įrašo pavyzdys

examples/hod/1000000-head.cif:

```
data_1000000
loop_
_[local]_description
'ENVI File'
'Created [Wed Jun 08 12:34:07 2016]'
_[local]_wavelength_units         Nanometers
loop_
_hyper_bands.default
220
227
253
_hyper_bands.lines                937
_hyper_bands.number               288
_hyper_bands.samples              384
_hyper_file.byte_order            0
_hyper_file.data_type             4
_hyper_file.type                  ENVI_Standard
_hyper_header.offset              0
_hyper_header_file.contents
;ENVI
description = {
  ENVI File, Created [Wed Jun 08 12:34:07 2016]}
samples = 384
lines   = 937
```

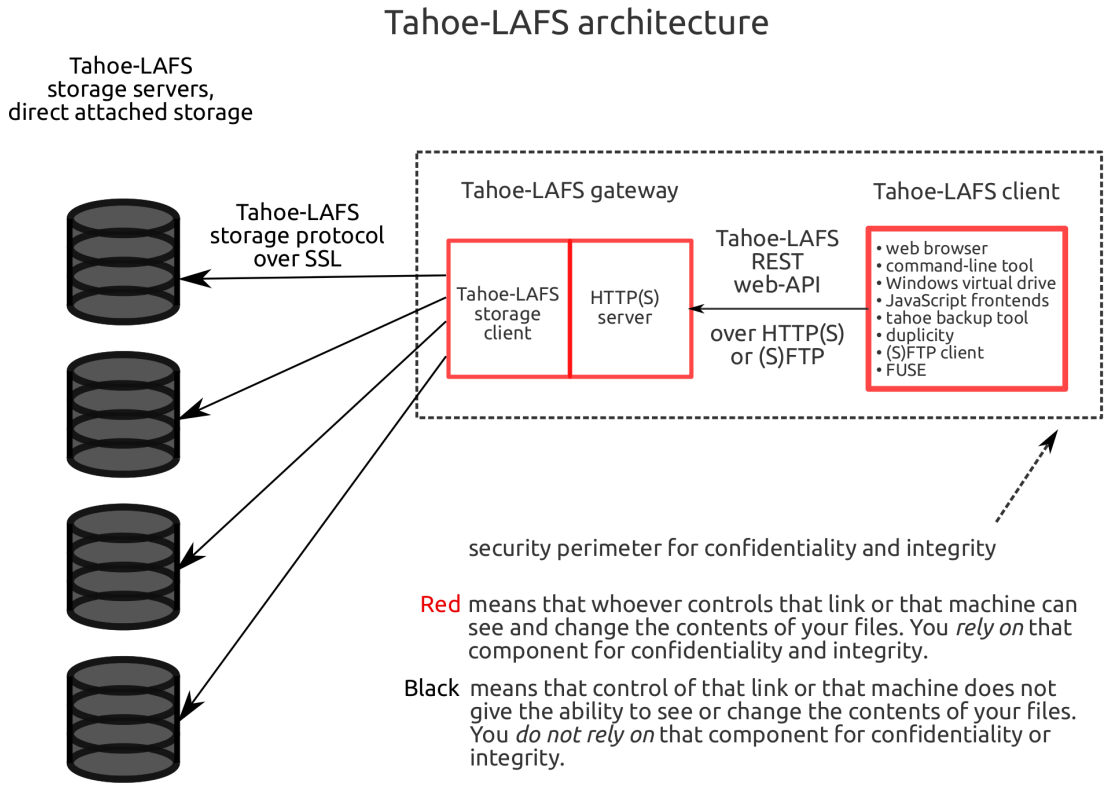


Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

SOLSA didelių duomenų talpykla

Tinkama, pvz., vaizdams

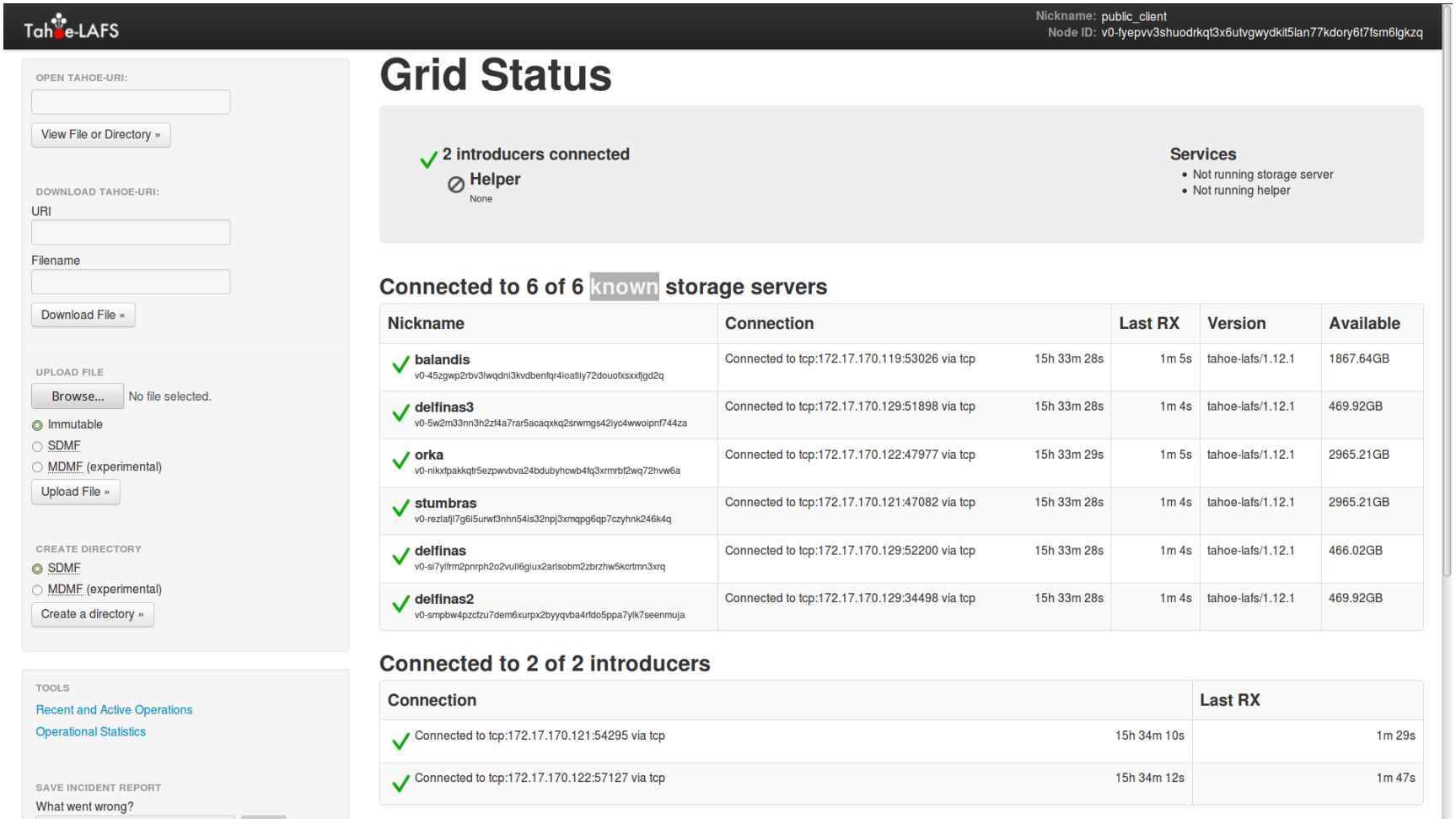
Naudoja Tahoe-LAFS (https://tahoe-lafs.org) kaip duomenų saugojimo „varikliuką“ [Selimi and Freitag, 2014]:



Paimta iš https://tahoe-lafs.org/trac/tahoe-lafs

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

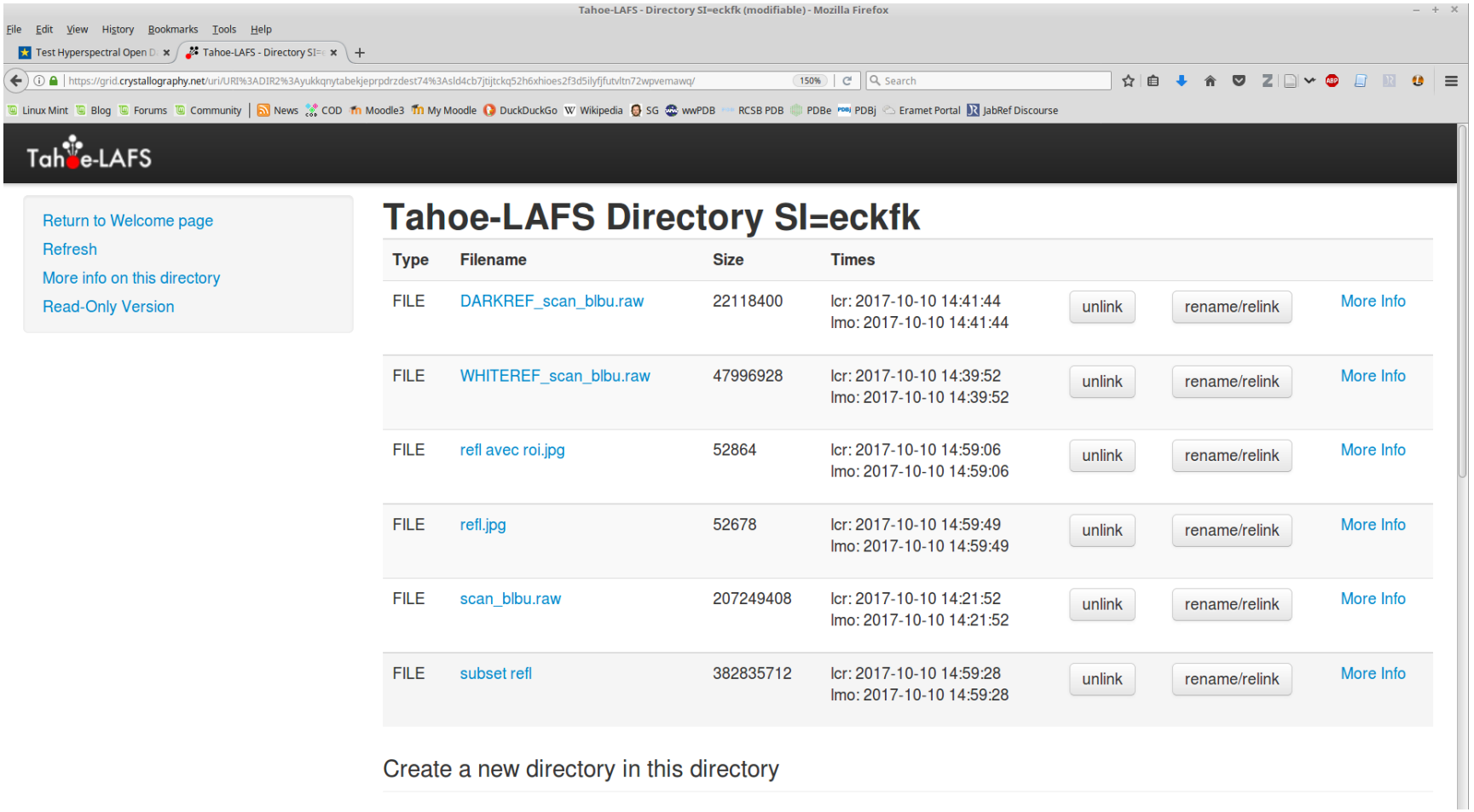
Tahoe LAFS „debesis“ SOLSA projektui



Tahoe-LAFS talpyklą SOLSA projektui parengė Erikas Raginis

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

HOD failai Tahoe LAFS „debesyje“



Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

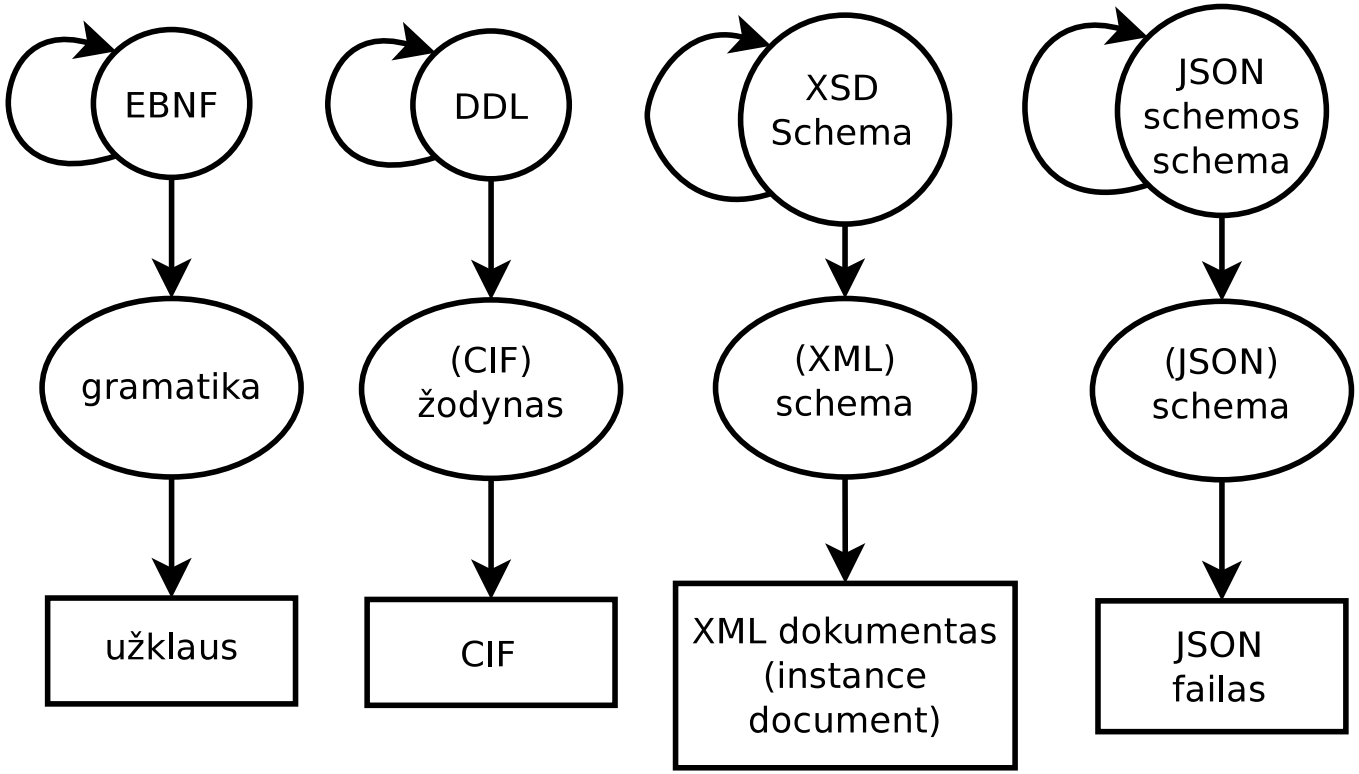
HOD didelių duomenų išsaugojimo politika

Prižiūrima duomenų apykaita:

- Išlaikome duomenis kurie yra:
  - Pirmi gauti tam tikros rūšies duomenys;
  - Geriausi tam tikros rūšies duomenys;
  - Dažniausiai cituojami;
  - Sudaro reprezentatyvią šios rūšies įrašų imtį (pvz. PĮ testavimui);
- Seniems, retai naudojamiems įrašams taikysime prarandančius duomenis glaudinimo algoritmus (įmanomi suspaudimo lygiai apie $\times 20$ kartų);
- Seni nenaudojami duomenys bus ištrinami, paliekant tik (agreguotus) metaduomenis.

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Save aprašantys duomenys



Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Padėkos

VU Biotechnologijos institutas

Virginijus Šikšnys
(*skyriaus vedėjas*)

Andrius Merkys
Antanas Vaitkus
Erikas Raginis

SOLSA komanda

Monique Le Guen
Beate Orberger
Daniel Chateigner
Henry Pilliere
ir visa projekto komanda!

COD Patarėjų taryba

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

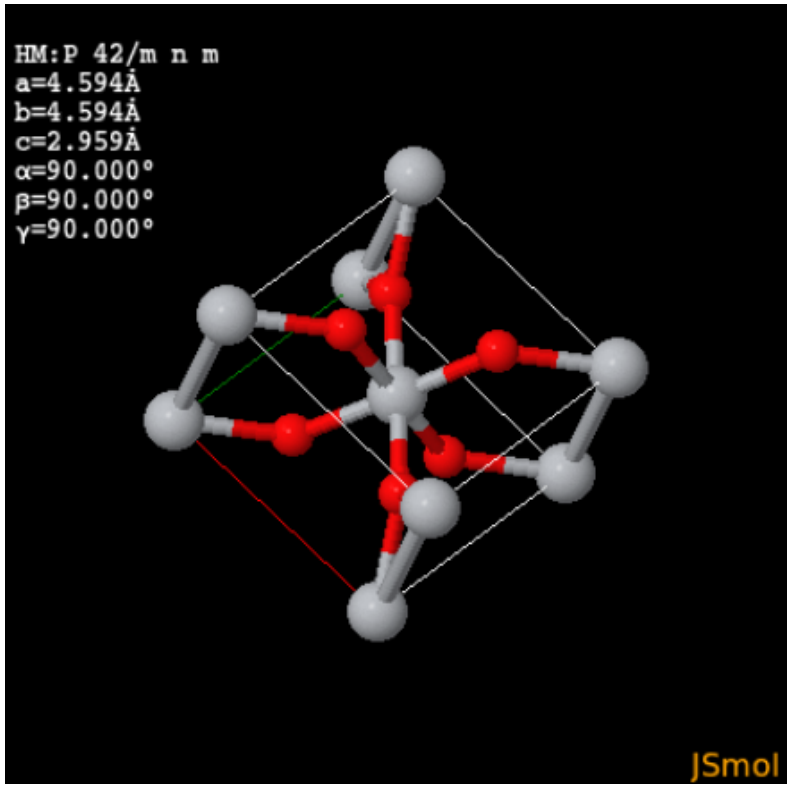
Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Ačiū už dėmesį!

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

Rob Lavinsky, iRocks.com – CC-BY-SA-3.0



http://www.crystallography.net/9015662.html

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Šaltiniai I


Fielding, R. T. (2000).
Architectural Styles and the Design of Network-based Software Architectures.
PhD thesis, University of California, Irvine.


Hall, S. R., Allen, F. H., and Brown, I. D. (1991).
The crystallographic information file (CIF): a new standard archive file for crystallography.
Acta Crystallographica Section A, 47:655–685.


Hart, P. E. and Duda, R. O. (1977).
Prospector – a computer-based consultation system for mineral exploration.
techreport, Artificial Intelligence Center, SRI International, Menlo Park, California 94025.

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Šaltiniai II

Le Bail, A. (2008).
Frontiers between crystal-structure prediction and determination by powder diffractometry.
Powder Diffraction Suppl., pages S5–S12.

Selimi, M. and Freitag, F. (2014).
Tahoe-lafs distributed storage service in community network clouds.
2014 IEEE Fourth International Conference on Big Data and Cloud Computing.

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.

Kuo geras REST

REST užklausos [Fielding, 2000]:

- **nepriklauso** nuo programavimo kalbos ar ryšio protokolo;
- GET užklausos yra null-potentinės (t.y. duomenų skaitymo užklausos nieko nekeičia ir ta pati užklausa kiekvieną kartą grąžina tą patį rezultatą);
- POST/PUT užklausos yra idempotentinės (t.y. duomenų įkėlimo užklausos pakeičia serverį tik vieną kartą, t.y. pakartotos vėl turi tokį pat poveikį, kaip ir vieną kartą atlikta pirmoji užklausa).

Šis projektas finansuojamas ES tyrimo ir inovacijos programos H2020 pagal granto sutartį Nr. 689868.