



Atviros susietos duomenų bazės kalnakasybos pramonėje

Saulius Gražulis

Kaunas, OpenCon 2017

Vilniaus universiteto Biotechnologijos institutas



Šį darbą galite naudoti pagal
Attribution-ShareAlike 4.0 tarptautinės licenzijos sąlygas

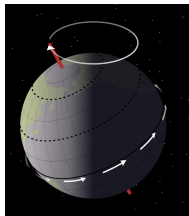




Duomenų svarba

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 reiškinių, kurių mes dabar vadiname *lygiadienių precesija*



NASA, Viešoji nuosavybė

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



Duomenys ir dirbtinio intelekto sistemos geologijoje

[Hart and Duda, 1977]

October 20, 1977

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

by

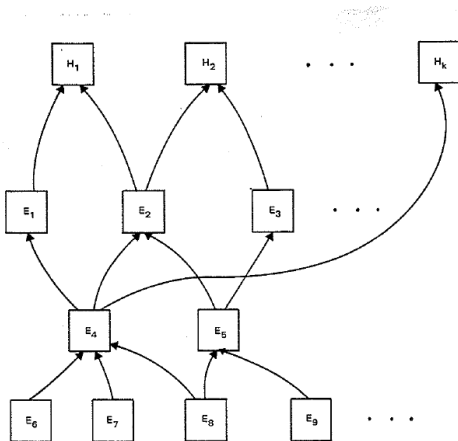
P. E. Hart and R. O. Duda

Artificial Intelligence Center
SRI International
Menlo Park, California 94025



Programos PROSPECTOR išvadų generavimo tinklas

[Hart and Duda, 1977]





Duomenys SOLSA projekte



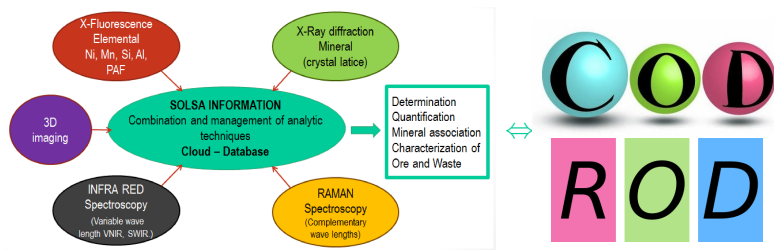
Discover SOLSA

<http://solsa-mining.eu/>

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



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.



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), schemas (XML, JSON)
- ▶ Tvarumas
 - ▶ Nuolatiniai identifikatoriai
 - ▶ Atvirų duomenų principai
 - ▶ Aptinkamumas, suderinamumas, sąveikumas, tinkamumas naudoti (FAIR)



Duomenų mainai kristalografijoje

The screenshot shows the IUCr website's navigation menu with categories like 'journals', 'books', 'news', 'education', 'people', 'resources', and 'outreach'. The main content area is titled 'Specifications' and includes a large 'CIF' logo. The text explains that the pages provide the formal specification of the CIF file format, mentioning versions 1.1 and 2.0, and references to Hall, Allen & Brown (1991) and COMCIFS (1997). It also notes that ancillary notes describe conventions and guidelines for CIF data items.

[Hall et al., 1991]

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

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



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)
```



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         Å
  _units_detail  'angstroms'
;
  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_.
;
```



Duomenys kristalografijoje

Atvira kristalografinė duomenų bazė COD

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

Crystallography Open Database

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
Advises to donators
Useful links

Open-access collection of crystal structures of organic, inorganic, metal-organic 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 **385190** entries in the COD.
Latest deposited structure: [1547638](#) on **2017-10-07** at **23:51:11 UTC**



Kristalo struktūros pavyzdys

COD Sfleritas

<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](#)

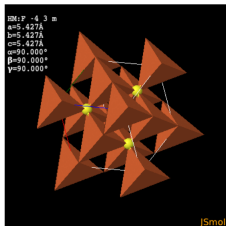
Documentation

[COD Wiki](#)
[Obtaining COD](#)
[Querying COD](#)
[Citing COD](#)
[COD Mirrors](#)
[Advice to donors](#)
[Useful links](#)

Information card for entry 1525302

[1525301](#) << [1525302](#) >> [1525303](#)

Preview



[Display in Jmol](#)

Coordinates [1525302.cif](#)

Coordinates [1525302.cif](#)

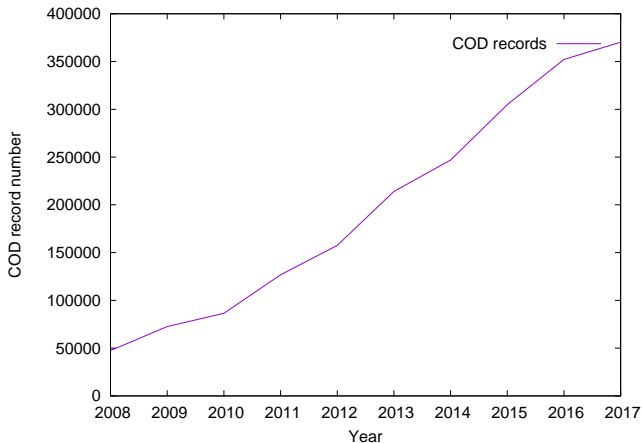
Structure parameters

Chemical name	Fe _{0.2} Mn _{0.05} S _{2n0.75} 5
Formula	Fe _{0.2} Mn _{0.05} S _{2n0.75}
Calculated formula	Fe _{0.2} Mn _{0.05} S _{2n0.75}
Title of publication	Unit-cell edges of natural and synthetic sflerites
Authors of publication	Sämer, 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 Fdata	No



COD duomenų bazės tvarumas

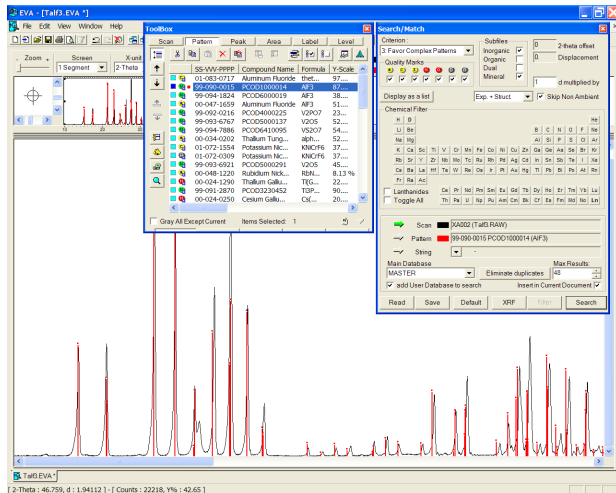
COD prieinama Internete jau 13 metų, per paskutinius 8 metus išaugo 7 kartus; šiuo metu talpina daugiau negu 385 000 įrašų (2017 spalį):



COD ir PCOD duomenų bazių panaudojimas

Kristalinės medžiagos identifikavimas

Medžiagos identifikavimas pagal Rentgeno spindulių sklaidymo intensyvumus.



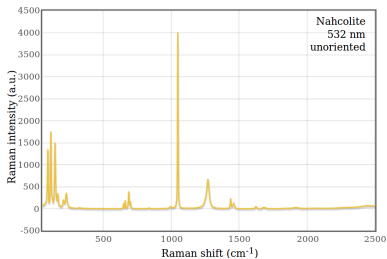
Paveiksliuką parengė Armelis Le Bail ((Le Bail, 2008))



Ramano spektroskopija



us-tech.co.za



ROD 3500101

- ▶ labai greitas metodas
- ▶ reikalauja geros išsamios duomenų bazės



Ramano spektroskopijos duomenys

Atvira Ramano spektrų duomenų bazė ROD

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



Raman Open Database

ROD Home

Home
What's new?

Accessing ROD Data

Search

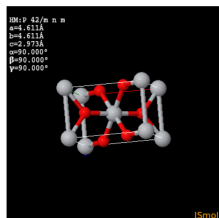
Add Your Data

Deposit your data
Manage depositions
Manage/release
prepublications

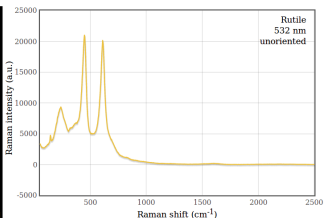
Information card for entry 3500024

3500023 << 3500024 >> 3500025

Preview



[Display in Jmol](#)



Duomenų įrašai įkelti į ROD Yassine El Mendili



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'
```



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 omitted 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



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 → 2.x);



COD užklausių 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 užklausių 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'
```



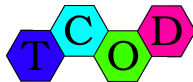
Atviros kristalografinės duomenų bazės

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



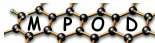
<http://www.crystallography.net/cod>

> 385 000 įrašų (augš iki > 10^6 ?)



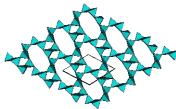
<http://www.crystallography.net/tcod>

> 2500 įrašų (augš iki > 10^7 ?)



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

> 300 įrašų



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

> 10^6 įrašų (augš iki > 10^8 ?)



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

> 120 įrašų

HOD

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

ruošiamą...



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



Hiperspektrinių vaizdų duomenų bazė (HOD)

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

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

- ▶ Metaduomenys ir paveiksluko antraštės saugomos CIF formatu;
- ▶ Paveiksluko rastras saugomas kaip „žali“ dvejetainiai žinomo tipo duomenys;



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
```



[HOD Home](#)

[Home](#)
[What's new?](#)

[Accessing HOD Data](#)

[Search](#)

[Add Your Data](#)

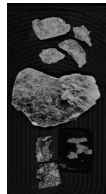
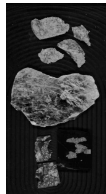
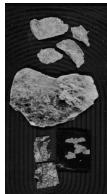
[Deposit your data](#)
[Manage deposits](#)
[Manage publications](#)

Test Hyperspectral Open Database

Information card for entry 1000000

4000001 << 1000000 >> 4000000

Preview



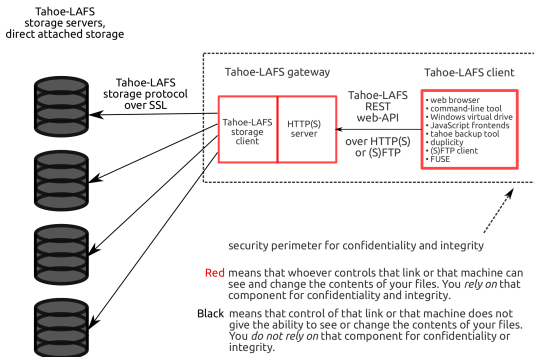


SOLSA didelių duomenų talpykla

Tinkama, pvz., vaizdams

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

Tahoe-LAFS architecture



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



Tahoe LAFS „debasis“ SOLSA projektui

Tahoe-LAFS

Nickname: public_client
Node ID: v0-fyepv3shuodkq3x6utvgjydk15lan77kdory6t7sm6gkzq

Grid Status

✓ 2 introducers connected

○ Helper
None

Services

- Not running storage server
- Not running helper

Connected to 6 of 6 known storage servers

Nickname	Connection	Last RX	Version	Available
✓ balandis v0-45qps0v3wq3k3wdbent44oaf8722uotefv0f9t0z	Connected to tcp:172.17.170.119:53026 via tcp	15h 33m 28s	1m 5s	tahoe-lafs/1.12.1 1867.64GB
✓ delfinas3 v0-5wz0z0m3r0z0f6a7ar5scapke2nwmqg42y0xawwqz0744z	Connected to tcp:172.17.170.129:51898 via tcp	15h 33m 28s	1m 4s	tahoe-lafs/1.12.1 469.92GB
✓ orka v0-ekvpa0q56zqpw0v240duyhoeb4q3kerret0wq72hw6a	Connected to tcp:172.17.170.122:47977 via tcp	15h 33m 29s	1m 5s	tahoe-lafs/1.12.1 2965.21GB
✓ stumbras v0-rcos07y6b0vna03m64s02ng3umngq0p70zjv9k046kq	Connected to tcp:172.17.170.121:47082 via tcp	15h 33m 28s	1m 4s	tahoe-lafs/1.12.1 2965.21GB
✓ delfinas v0-07yhr0z0p0z0z0v0f0g0x0ar0v0e0z0z0x0e0k0tr0n03q	Connected to tcp:172.17.170.129:52200 via tcp	15h 33m 28s	1m 4s	tahoe-lafs/1.12.1 466.02GB
✓ delfinas2 v0-0p0w0q0z0z070em0v0z0z0y0p0a040f0p07y0c0ee0m0q	Connected to tcp:172.17.170.129:34498 via tcp	15h 33m 28s	1m 4s	tahoe-lafs/1.12.1 469.92GB

Connected to 2 of 2 introducers

Connection	Last RX
✓ Connected to tcp:172.17.170.121:54295 via tcp	15h 34m 10s 1m 29s
✓ Connected to tcp:172.17.170.122:57127 via tcp	15h 34m 12s 1m 47s

OPEN TAHOE-URI:

DOWNLOAD TAHOE-URI:

URI

Filename

UPLOAD FILE

 No file selected.

Immutible

SDMF

MDMF (experimental)

CREATE DIRECTORY

SDMF

MDMF (experimental)

TOOLS

[Recent and Active Operations](#)

[Operational Statistics](#)

SAVE INCIDENT REPORT

What went wrong?

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

Tahoe-LAFS talpyklą SOLSA projektui parengė Erikas Raginis



HOD failai Tahoe LAFS „debesyje“

Return to Welcome page
Refresh
More info on this directory
Read-Only Version

Tahoe-LAFS Directory SI=eckfk

Type	Filename	Size	Times			
FILE	DARKREF_scan_bibu.raw	22118400	lcr: 2017-10-10 14:41:44 lmo: 2017-10-10 14:41:44	unlink	rename/relink	More info
FILE	WHITEREF_scan_bibu.raw	47996928	lcr: 2017-10-10 14:39:52 lmo: 2017-10-10 14:39:52	unlink	rename/relink	More info
FILE	refl avec roi.jpg	52864	lcr: 2017-10-10 14:59:06 lmo: 2017-10-10 14:59:06	unlink	rename/relink	More info
FILE	refl.jpg	52678	lcr: 2017-10-10 14:59:49 lmo: 2017-10-10 14:59:49	unlink	rename/relink	More info
FILE	scan_bibu.raw	207249408	lcr: 2017-10-10 14:21:52 lmo: 2017-10-10 14:21:52	unlink	rename/relink	More info
FILE	subset refl	382835712	lcr: 2017-10-10 14:59:28 lmo: 2017-10-10 14:59:28	unlink	rename/relink	More info

Create a new directory in this directory



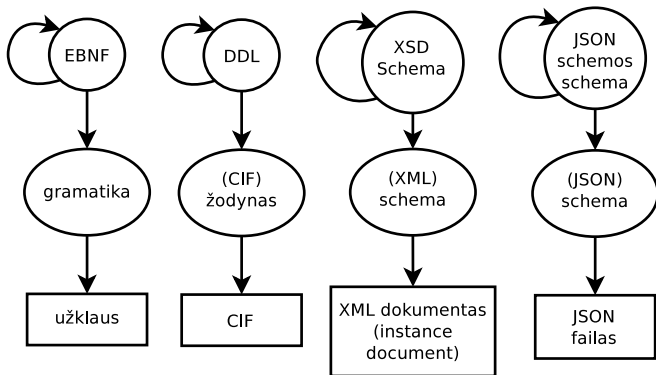
HOD didelių duomenų išsaugojimo politika

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. PI 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.



Save aprašantys duomenys





Padėkos

**VU Biotechnologijos SOLSA komanda
institutas**

Virginijus Šikšnys
(*skyriaus vedėjas*)

Andrius Merkys
Antanas Vaitkus
Erikas Raginis

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 grantą sutartį Nr. 689868.**

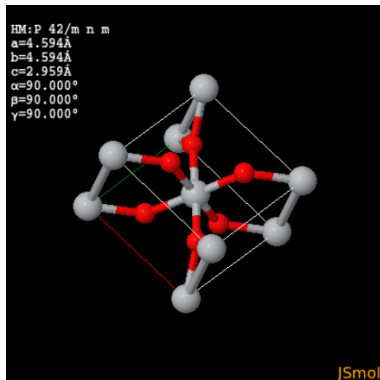


Ačiū už dāmes!






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

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





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

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.

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.



Kuo geras REST

REST užklauso [Fielding, 2000]:

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