Efficient long-term open-access data archiving in mining industries

Saulius Gražulis & the SOLSA consortium

Amsterdam, RTM Conference, 2017

Vilnius University Institute of Biotechnology

This work is licensed under a Creative Commons Attribution 4.0 International License
Data importance

Hipparchus (c. 190 – c. 120 BCE)

- measured the longitude of Spica and Regulus and other bright stars
- compared his measurements with data from his predecessors, Timocharis and Aristillus, who lived $\approx 100$ years before him,
- discovered what is now called the precession of the equinoxes

(Wikipedia, see also articles on Timocharis and Aristyllus)
Data and AI systems for geology

[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
The PROSPECTOR network of inference

[Hart and Duda, 1977]
Data kinds in the SOLSA project

- Crystal structures (COD)
- Raman spectra (ROD)
- Hyperspectral spectra (HOD)

http://solsa-mining.eu/
Requirements for long-term data archiving and reuse

- Platform independence
  - Text-based formats (ASCII, UTF-8)
- Software independence
- Network-transparency
  - Standard, open protocols (W3C http)
  - Standard, open data carrier formats (JSON, XML, CIF).
  - RESTful servers
- Machine-readable semantics
  - Dictionaries, schemas
- Durability
  - Persistent identifiers
  - Open data principles
  - FAIR principles
Data exchange in crystallography

The Crystallographic Interchange File/Framework (CIF):

- Provides standard means for data publishing and exchange;
- Is suitable for archiving;
- Is maintained by the IUCr;

[Hall et al., 1991]
CIF for scientific data

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:KMoO$_4$:BaTiO$_3$'
  _chemical_formula_sum 'Ba O$_3$ 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)
Controlled vocabularies

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

```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 _dифfrn_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 _dифfrn_reflns_transf_matrix_.
```

Crystallographic data

The Crystallography Open Database

http://www.crystallography.net/cod

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
A COD crystal structure page example

Sphalerite

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

Crystallography Open Database

Information card for entry 1525302

1525301 << 1525302 >> 1525303

Preview

 Coordinates 1525302.cif

<table>
<thead>
<tr>
<th>Chemical name</th>
<th>(FeO.2 MnO.05 ZnO.75) Si</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formula</td>
<td>FeO.2 MnO.05 S ZnO.75</td>
</tr>
<tr>
<td>Calculated formula</td>
<td>FeO.2 MnO.05 S ZnO.75</td>
</tr>
<tr>
<td>Title of publication</td>
<td>Unit-cell edges of natural and synthetic sphalerites</td>
</tr>
<tr>
<td>Authors of publication</td>
<td>Skinner, RJ</td>
</tr>
<tr>
<td>Journal of publication</td>
<td>American Mineralogist</td>
</tr>
<tr>
<td>Year of publication</td>
<td>1961</td>
</tr>
<tr>
<td>Journal volume</td>
<td>46</td>
</tr>
<tr>
<td>Pages of publication</td>
<td>1399 - 1411</td>
</tr>
<tr>
<td>a</td>
<td>5.4272 Å</td>
</tr>
<tr>
<td>b</td>
<td>5.4272 Å</td>
</tr>
<tr>
<td>c</td>
<td>5.4272 Å</td>
</tr>
<tr>
<td>α</td>
<td>90°</td>
</tr>
<tr>
<td>β</td>
<td>90°</td>
</tr>
<tr>
<td>γ</td>
<td>90°</td>
</tr>
<tr>
<td>Cell volume</td>
<td>159.855 Å³</td>
</tr>
<tr>
<td>Number of distinct elements</td>
<td>4</td>
</tr>
<tr>
<td>Hermann-Mauguin symmetry space group</td>
<td>F-4 3 m</td>
</tr>
<tr>
<td>Hall symmetry space group</td>
<td>F-4 3</td>
</tr>
<tr>
<td>Has coordinates</td>
<td>Yes</td>
</tr>
<tr>
<td>Has disorder</td>
<td>No</td>
</tr>
<tr>
<td>Has Fmax</td>
<td>No</td>
</tr>
</tbody>
</table>

Display in Jmol

Coordinates 1525302.cif
COD persistence

COD is on-line for 13 years, increased 7-fold over the last 8 years; currently contains over 380,000 records (October 2017):
Raman spectroscopy data

The Raman Open Database

http://solsa.crystallography.net/rod

Raman Open Database

Information card for entry 3500024

3500023 << 3500024 >> 3500025

Preview

Data records to ROD contributed by Yassine El Mendili
ROD data files

ROD uses CIF syntax

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'
The ROD dictionary

ROD uses controlled vocabulary in CIF DDLm dictionaries

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:

```plaintext
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 dictionaries coded by Antanas Vaitkus
Semantic versioning of the ROD dictionaries

- ROD dictionaries undergo semantic versioning:
  - Bug-fix releases (1.2.x) are compatible backwards and forward;
  - Minor releases (1.x) are backwards compatible;
  - Incompatible changes will be marked by major releases (1.x → 2.x);
SOLSA project, COD and ROD

COD will be used in SOLSA for:

- mineral identification;
- subsequent data dissemination.

SOLSA data flow diagram courtesy Monique Le Guen, ERAMET.
The fun of REST

RESTful queries [Fielding, 2000]:

- Programming language, transfer protocol independent
- GET queries should be null-potent (do not change anything; always provide the same result for the same query);
- POST/PUT queries should be idempotent (the same query executed several times should have the same result as just one query).
COD query examples
Web, REST, SQL

- Via the WWW interface – go for “search” in:
  - http://www.crystallography.net/cod
  - http://www.crystallography.net/tcod
  - http://www.crystallography.net/pcod

- Via the **stable** URLs (REST):
  - http://www.crystallography.net/cod/2000000.cif
  - http://www.crystallography.net/tcod/10000002.cif
  - http://www.crystallography.net/cod/result?text=perovskite

- Via the **views** of the SQL database:
  - `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'`
# Acknowledgements

<table>
<thead>
<tr>
<th>VU Institute of Biotechnology</th>
<th>The SOLSA team</th>
<th>COD Advisory board</th>
</tr>
</thead>
<tbody>
<tr>
<td>Virginijus Siksnys (head of the dept.)</td>
<td>Monique Le Guen</td>
<td>Daniel Chateigner</td>
</tr>
<tr>
<td>Andrius Merkys</td>
<td>Beate Orberger</td>
<td>Robert T. Downs</td>
</tr>
<tr>
<td>Antanas Vaitkus</td>
<td>Daniel Chateigner</td>
<td>Werner Kaminsky</td>
</tr>
<tr>
<td></td>
<td>Henry Pilliere</td>
<td>Armel Le Bail</td>
</tr>
<tr>
<td></td>
<td>and all the team</td>
<td>Luca Lutterotti</td>
</tr>
<tr>
<td></td>
<td>working on the project!</td>
<td>Peter Moeck</td>
</tr>
</tbody>
</table>

This project has received funding from the European Union’s Horizon 2020 research and innovation program under grant agreement No 689868.
Thank you!

http://en.wikipedia.org/wiki/Emerald
http://www.crystallography.net/5000095.html

A path to freedom: GNU → Linux → Ubuntu → MySQL → R → \LaTeX → TikZ → Beamer
References I


Open Crystallographic Databases
COD, TCOD, PCOD, MPOD, ...

http://www.crystallography.net/cod
> 366 000 entries (ready to grow > $10^6$?)

http://www.crystallography.net/tcod
> 2000 entries (ready to grow to > 350 000?)

http://mpod.cimav.edu.mx/
> 300 entries

http://www.crystallography.net/pcod
> $10^6$ entries (ready to grow to > $10^8$?)

A path to freedom: GNU → Linux → Ubuntu → MySQL → R → $\LaTeX$ → TikZ → Beamer
COD accessibility

COD is a **fully open-access database**. All records are available under public domain designation.

Provided access methods are:

- Web search
- URLs constructed from stable identifiers
- RESTful interfaces
- Full data download
Common REST API

- Agreed upon in the 2016 Leiden CECAM workshop;
- Suitable for all structural and QM databases.

https://github.com/Materials-Consortia/API

A path to freedom: GNU → Linux → Ubuntu → MySQL → R → TEX → TikZ → Beamer
Definitions of input and output

(* The top-level 'filter' rule: *)
Filter = Keyword, Expression;

(* Keywords *)
Keyword = "filter=" ;

(* Values *)
Value = Identifier | Number | String ;

(* ... some token definitions skipped for brevity ... *)

(* Expressions *)
Expression = Term, [Spaces], [ OR, [Spaces], Expression ] ;
Term = Comparison, [Spaces], [ AND, [Spaces], Term ] ;

(* Operator Comparison operator tokens: *)
Operator = '<', [ '=' ] | '>' , [ '=' ] | '=' | '|' | '::' ;
Comparison = Value, [Spaces], Operator, [Spaces], Value |
       NOT, [Spaces], Comparison |
       '((', [Spaces], Expression, [Spaces], ')' ) ;
Schemas for return data

Schema’s allow to:

- formally agree on what is right and wrong;
- validate program outputs and documents automatically.

"query": {
  "type": "object",
  "properties": {
    "representation": { "type": "string" },
    "api_version": { "type": "string" },
    "time_stamp": { "type": "string" },
    "data_returned": { "type": "integer" },
    "data_available": { "type": "integer" },
    "last_id": { "type": "string" }
  },
  "required": [ "representation", "api_version", "time_stamp" ]
}
API query examples

http://solsa.crystallography.net/cod/optimade/structures?filter=elements="Si,O"ANDnelements=2&limit=1

{
    "resource": {
        "base_url": "http://www.crystallography.net/cod/optimade/v1.0.0-alpha.1/
    },
    "query": {
        "api_version": "v1.0.0-alpha.1",
        "data_returned": 1,
        "representation": "/structures?filter=elements="Si,O"ANDnelements=2&limit=1",
        "last_id": "1010921",
        "time_stamp": "2017-04-06T05:46:50Z",
        "implementation": {
            "maintainer": {
                "email": "cod-bugs@ibt.lt"
            },
            "title": "Crystallography Open Database",
            "version": "v1.0.0-alpha.11",
            "source_url": "svn://crystallography.net/cod/trunk/cod/cgi-bin/optimade.pl@194653"
        },
        "data_available": 344
    },
    "data": [
        {
            "last_modified": "2017-02-28T05:33:56Z",
            "properties": {
                "formula": "O2Si"
            },
            "url": "http://www.crystallography.net/cod/1010921.cif",
            "immutable_id": "http://www.crystallography.net/cod/1010921.cif@130149",
        }
    ]
}
Common pattern of self-describing data definitions

A path to freedom: GNU → Linux → Ubuntu → MySQL → R → \LaTeX\ → TikZ → Beamer