The Crystallography Open Database – new perspectives

Saulius Gražulis Andrius Merkys Antanas Vaitkus Armel Le Bail Daniel Chateigner Henry Pilliere Robert T. Downs Luca Lutterotti Peter Moeck Peter Murray-Rust Miguel Quirós Olozábal Werner Kaminsky

Denver, SciDataCon2016

Vilnius University Institute of Biotechnology

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

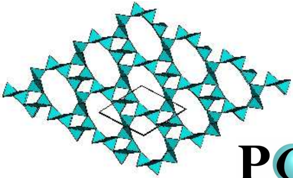
Open Crystallographic Databases

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 $> 10^8$?)

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

The COD project

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).

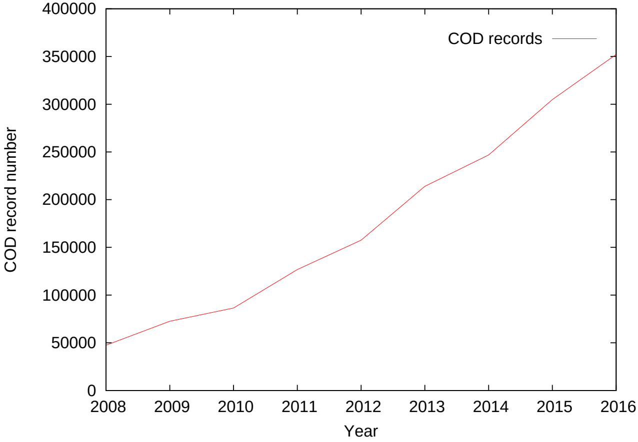
3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval.

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

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

COD 13 years later

COD increased 7-fold; currently contains over 366000 records (Sept. 2016)



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

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

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

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
                date between "2013-01-01" and
                             "2014-12-31" and
                formula regexp " C[0-9]* "
                order by vol desc limit 10'
    ```

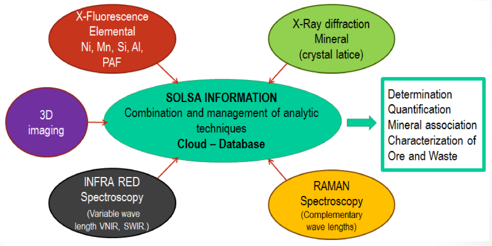
This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868.
    ```

COD applications

- SOLSA
  - http://www.solsa-mining.eu/
- AiiDA [Pizzi et al., 2016]
  - http://www.aiida.net/
- COSMOS [Sadowski and Baldi, 2013]
  - http://cdb.ics.uci.edu/
- FPSM [Boullay et al., 2014], MAUD [Boullay et al., 2012]
  - http://fpsm.radiographema.com/
  - http://maud.radiographema.eu/
- DataWarrior
  - http://www.openmolecules.org/datawarrior/
- MolView
  - http://molview.org/
- search-match (Bruker, PANalytical, Rigaku)
- ... and more!

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

SOLSA project and COD



⇔

COD will be used in SOLSA for:

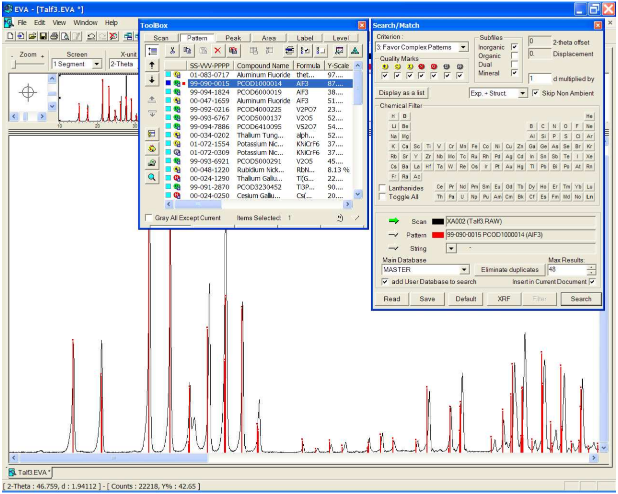
- mineral identification;
- subsequent data dissemination.

SOLSA data flow diagram courtesy Monique Le Guen, ERAMET.

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

Use of *COD databases

Search-match identification of the materials

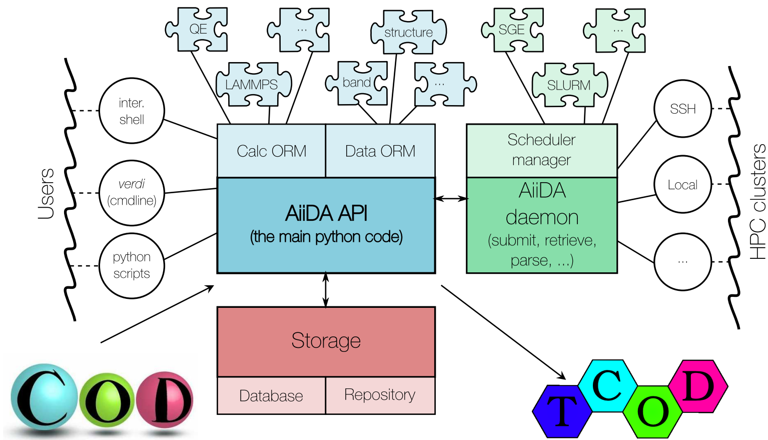


A **predicted** phase from PCOD could be identified in experimental data.

Courtesy Armel Le Bail [Le Bail, 2008]

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

COD, TCOD and AiiDA link

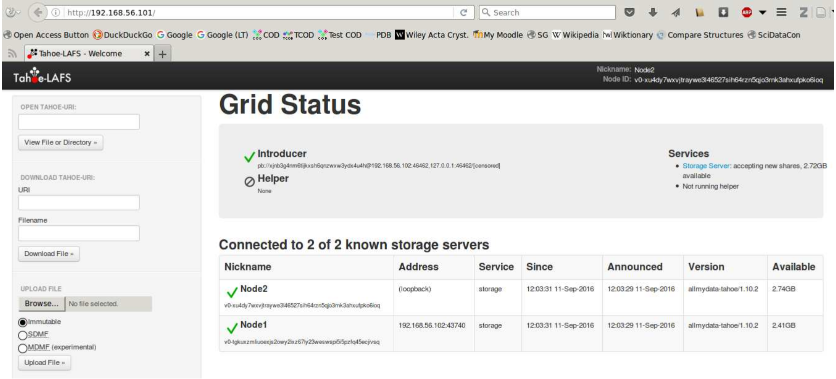


Courtesy AiiDA developers [Pizzi et al., 2016]

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

COD Diffraction Image Store

Uses Tahoe-LAFS (https://tahoe-lafs.org) as a back-end:

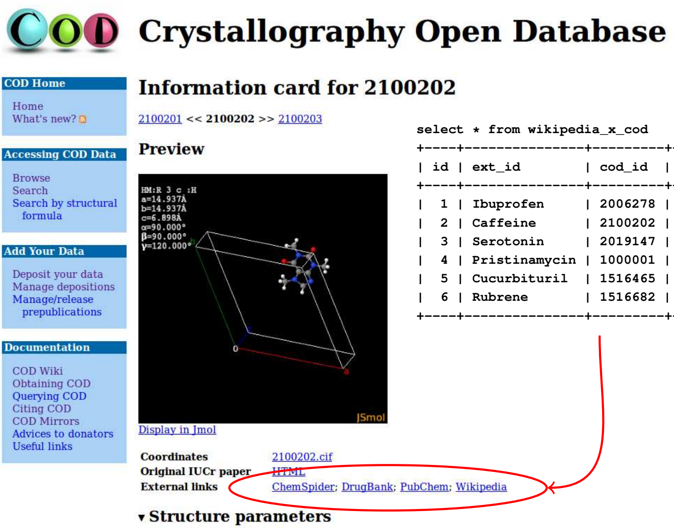


Provides:

- community-backed store ($\geq$1 PB)
- confidentiality through strong encryption
- extreme hardware loss tolerance

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

Interlinked data in COD




```
select * from wikipedia_x_cod
+----+---------------+---------+-------------+
| id | ext_id        | cod_id  | relation_id |
+----+---------------+---------+-------------+
|  1 | Ibuprofen     | 2006278 |           1 |
|  2 | Caffeine      | 2100202 |           1 |
|  3 | Serotonin     | 2019147 |           1 |
|  4 | Pristinamycin | 1000001 |           1 |
|  5 | Cucurbituril  | 1516465 |           1 |
|  6 | Rubrene       | 1516682 |           1 |
+----+---------------+---------+-------------+
```

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

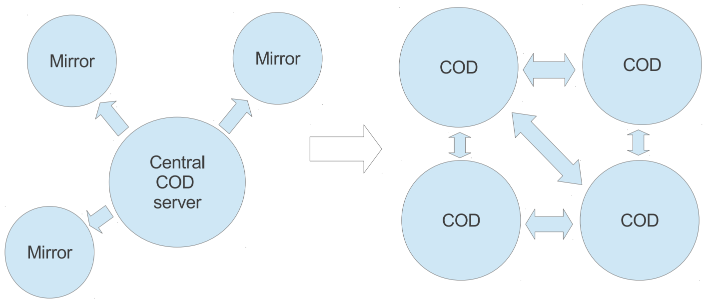
COD completeness challenge

nr	journal	publisher
45157	Inorganic Chemistry	American Chemical Society
42069	Acta Crystallographica Sect. E	International Union of Crystallography
28775	Dalton transactions (Cambridge ...	Royal Society of Chemistry
26752	Organometallics	American Chemical Society
25493	Journal of the American Chemic ...	American Chemical Society
19824	Acta Crystallographica Sect. C	International Union of Crystallography
19028	Chemical Communications	Royal Society of Chemistry
17858	CrystEngComm	Royal Society of Chemistry
13225	Crystal Growth & Design	American Chemical Society
11083	The Journal of Organic Chemist ...	American Chemical Society
9358	Acta Crystallographica Sect. B	International Union of Crystallography
7910	Organic Letters	American Chemical Society
7516	Dalton Transactions	Royal Society of Chemistry
5751	New Journal of Chemistry	Royal Society of Chemistry
5283	Organic & Biomolecular Chemist ...	Royal Society of Chemistry

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

COD durability assurance

- Best price/performance ratio
- Capability to build a distributed, equal-peer database



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

Acknowledgments

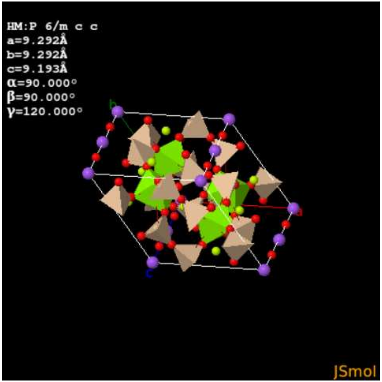
VU Institute of Biotechnology	**QM community**	**COD Advisory board**
	Björkman	Daniel Chateigner
Virginijus Siksnys	Torbjörn	Robert T. Downs
(*head of the dept.*)	Stefaan Cottenier	Werner Kaminsky
	Nicola Marzari	Armel Le Bail
Andrius Merkys	Giovanni Pizzi	Luca Lutterotti
Antanas Vaitkus	Lubomir Smrcok	Peter Moeck
	Linas Vilčiauskas	Peter Murray-Rust
	Chris Wolverton	Miguel Quirós

Thanks to commercial COD users and supporters – Bruker, PANalytical, Rigaku; thanks to IUCr for support and consultations.

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

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

References

Boullay, P., Lutterotti, L., and Chateigner, D. (2012).
Quantitative analysis of electron diffraction ring patterns using the MAUD program.

Boullay, P., Lutterotti, L., Chateigner, D., and Sicard, L. (2014).
Fast microstructure and phase analyses of nanopowders using combined analysis of transmission electron microscopy scattering patterns.
Acta Crystallographica Section A, 70:448–456.

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

Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N., and Kozinsky, B. (2016).
AiiDA: automated interactive infrastructure and database for computational science.
Computational Materials Science, 111:218–230.

Sadowski, P. and Baldi, P. (2013).
Small-molecule 3d structure prediction using open crystallography data.
Journal of Chemical Information and Modeling, 53:3127–3130.

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