Software for maintaining and expanding the (Predicted) Crystallography Open Database

Saulius Gražulis[a], Justas Butkus[a], Robert Downs[b], Miguel Quirós Olozábal[c], Armel Le Bail[d]

[a] Institute of Biotechnology, Graiciuno 8, LT-02241 Vilnius, Lithuania
[b] Department of Geosciences, University of Arizona, Tucson, Arizona 85721-0077, USA
[c] Departamento de Química Inorgánica, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain
[d] Université du Maine, Laboratoire des Oxydes et Fluorures, CNRS UMR 6010, Avenue O. Messiaen, 72085 Le Mans Cedex 9, France

Institute of Biotechnology

EU Centre of Excellence

Introduction

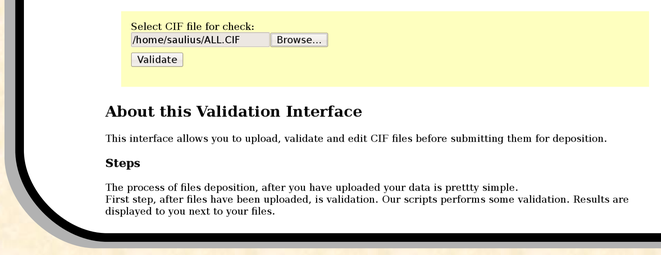
Data are arriving in ever increasing rates. Thus, automated software tools are needed to cope with the growing amount of data, to provide consistent, uniform and accurate information.

The following goals are set by COD team:

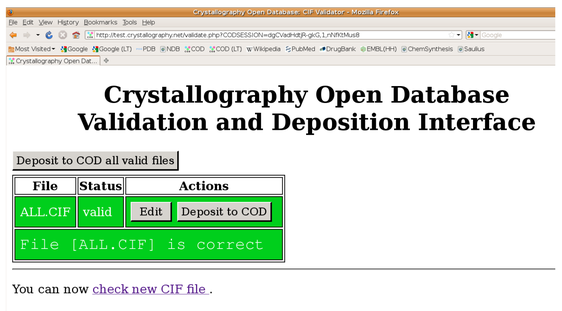
- build automated structure deposition tools;
- build a collaboration platform for structure validation and curation;
- ensure data quality – uniformity, integrity, and trustworthiness;
- make scientific data freely accessible to anyone.

Deposit your data to COD for publication!

http://www.crystallography.net/



Automatic data deposition



CIFParser

A new CIF parser was developed, to meet the following demands:

- identify, locate and clearly show errors in the CIF source (existing parsers were found to be not adequate for this task);
- callable from Perl for ultimate flexibility;
- capable to correct most common CIF errors in a predictable and documented way (via extension a documented and switchable extension of the IUCr CIF grammar);
- easy to extend and check against official CIF grammar for conformance – written in yapp notation, close to a Backus-Naur form.
- free software – distributable under an OSI – compatible license

cif-tools

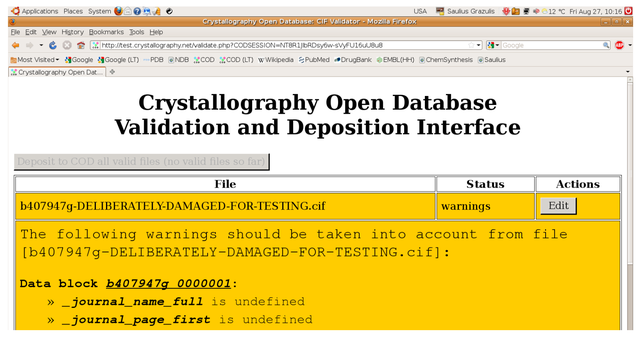
available at **svn://www.crystallography.net/cif-tools**

A set of tools based on CIFParser was developed and deployed:
- All basic programs a *x style command programs, developed for Linux;
- agile development – fast deployment, automatic unit testing ensures high quality and predictability
- new tools can be added as needed by anybody;
- integrable into GUIs and Web applications.

Example programs:

cif_filter – extract uncopywritable data from CIFs, fix errors, compute missing data values;
cif_cod_check – check whether a CIF satisfies COD requirements

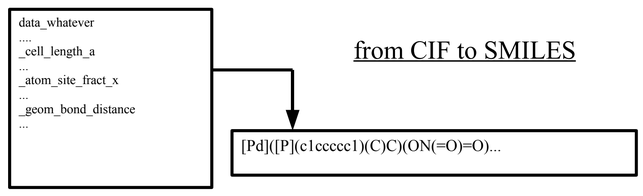
Automatic error detection/correction



Automated CIF download

Remember: atomic coordinates are not copyrightable.
The scripts introduce delays so that no noticeable charge will be put in the servers.
COD identifies itself as the downloader.
The bibliographic information is downloaded from Openurl or Pubmed and appended to the CIF's.

Search by SMILES



Use of available free tools: Openbabel (http://www.openbabel.org).
Very compact ASCII format
Human easily readable, understandable and editable.
Stores ONLY chemical connectivity, discard any other information.

SMILES-building helper scripts

Classification of structures: Is it easy to define what is the "molecule" ?
Choosing between openbabel-generated connectivity and author-supplied bonds.
Automatic fixing of some common mistakes.
Removing of chirality markers in non-chiral groups.
Human inspection of the results and fixing if necessary.

Expanding the PCOD

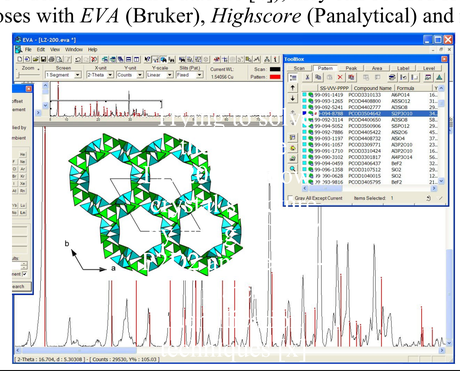
2010 update : 898.707 SiO_2 entries were added from *ZEFSAII* zeolite predictions and the contributions from *GRINSP* increased to 163.520 (silicates, phosphates, sulfates of Al, Ti, V, Ga, Nb, Zr, or zeolites, fluorides, etc). The PCOD is the first database to attain and offer **more than one million of CIF entries**.

Software : a new *GRINSP* version is now available [3] for parallel computing (for instance using fully the 8 processors of an INTEL core i7).

Other data from other prediction computer programs (*CASTEP*, *CERIUS2*, *CRYSTAL*, *G42*, *GULP*, *USPEX*…) are expected, just send them, please.

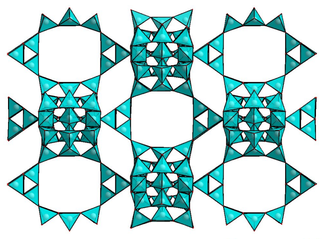
PCOD Powder P2D2

All powder patterns (> 1 million) were calculated and gathered in the P2D2 (Predicted Powder Diffraction Database [4]), they can be used for search-match purposes with *EVA* (Bruker), *Highscore* (Panalytical) and more soon.

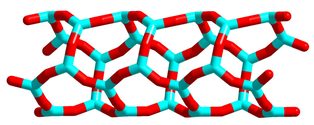


VIRTUAL MODELS in PCOD

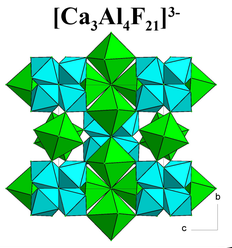
Zeolites



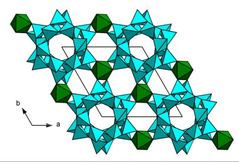
B_2O_3 nanotubes



$[Ca_3Al_4F_{21}]^{3-}$



Titanosilicates



Acknowledgments

The Vilnius COD development group is financed by the Research Council of Lithuania, contract No. MIP-124/2010

We thank Adriana Daškevič and Andrius Merkys for help with software development and data curation.

Authors thank to all CIF donators, listed on our Web page, and to numerous anonymous volunteers who help to collect data and keep COD running.

The COD Advisory Board thanks comercial supporters for donated hardware and financial support.

CONCLUSIONS

COD server is technically in position to store and serve all structures that are currently solved. COD deposition procedure ensures syntactic correctness and presence of the most essential CIF data.

COD and PCOD are constantly expanded to meet more needs of crystallographic community

We rely on the help of crystallographic community to add more data and ensure the data correctness!

Use and Add more structures to:
http://www.crystallography.net/

THANK YOU !

REFERENCES

[1] http://www.crystallography.net/
[2] S. Grazulis *et al.*, (2009). *J. Appl. Cryst.* 42, 726-729.
[3] A. Le Bail (2010). *Phys. Chem. Chem. Phys.* 12, 8521-8530.
[4] A. Le Bail (2008). *Powder Diffr.* Suppl. 23, S5-S12.

COD/PCOD : cod@crystallography.net