Crystallography Open Database: An Open Access Tool for Searching and Retrieving Crystallographic Data.

(http://www.crystallography.net)

Speaker: Miguel Quirós Olozábal
Departamento de Química Inorgánica, Universidad de Granada, Granada, Spain
(on behalf of all people that make COD possible)

What is COD?

- It is a large collection of CIF files (around 150000 at present) of "small molecule" structures. The ideal goal is to gather all available data of this kind.
- CIFs are world-wide freely accessible through a web interface at http://www.crystallography.net
- Data are organized in an SQL database so they can be searched by unit cell, chemical composition or arbitrary text (authors, title, bibliographic reference, ...). A basic substructure search engine has been just put on service.
- Data may be added to COD by crystallographers through the automatic deposition service. Structures can be made instantly worldwide accessible or put on hold until publication.

COD Development history

- February 2003. Challenge launched by Michael Berndt: "What if crystallographers work together to establish a public domain database with all relevant crystallographic data?"
- March 2003. COD is created by a team lead by Armel Le Bail with the server placed at Le Mans (Université du Maine). The first large set of data comes from an AMCSD donation (Robert T. Downs).
- December 2003. Creation of PCOD (Predicted Crystallography Open Database, Armel Le Bail).
- May 2005. Petition for Open Data in Crystallography (supported by more than 2000 signatures, including a Nobel laureate, when closed in 2008).
- September 2007. CIFs of IUCr journals made freely available to all databases.
- December 2007. COD main server and decision centre moves from Le Mans to Vilnius (Vilnius University Institute of Biotechnology). Main development group lead by Saulius Gražulis.

COD Development history (cont.)

- March 2008. COD reaches 50000 entries.
- August 2009. Publication of an article describing COD at J. Appl. Crystallogr. (**2009**, *42*, 4726-4729).
- ~ 2009. Start of systematic downloading from journal websites. First manually and now semiautomatically.
- June 2010. Financial support for the Vilnius development group from the Lithuanian Research Council.
- August 2010. Automatic deposition service operative at COD website.
- September 2010. COD reaches 100000 entries.
- May 2011. Substructure search for a subset of the database operative at COD website.

What is there inside a COD CIF-file?

Always:

_chemical_formula_sum
....
_cell_length_a
....
_symmetry_equiv_pos_as_xyz
....
_atom_site_fract_x
....

Never:

_publ_section_abstract
....
_publ_section_comment
....

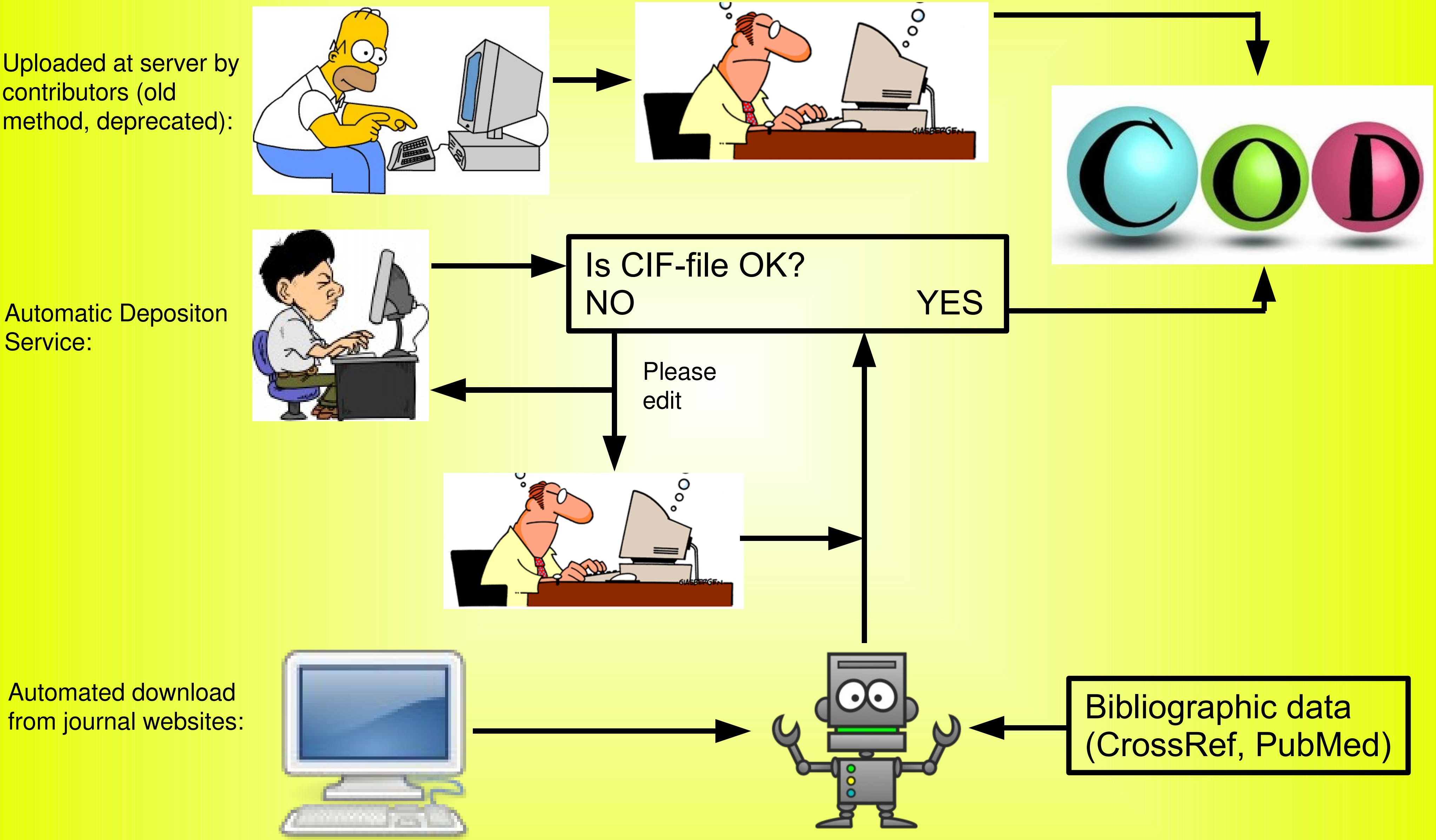
©

If available:

_cell_measurement_temperature
....
_diffrn_radiation_wavelength
....
_atom_site_aniso_U_11
....
_publ_author_name
....
_journal_name_full
....
_refine_ls_number_parameters
....
_refine_ls_wR_factor_ref
....
_geom_bond_distance
....

(in a separate file:)
_refln_F_squared_meas
....

Gathering data



Uploaded at server by contributors (old method, deprecated):

Automatic Depositon Service:

Is CIF-file OK?

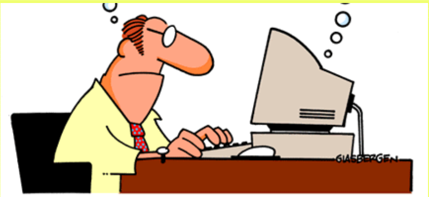
NO

YES

Please edit

Automated download from journal websites:

Bibliographic data (CrossRef, PubMed)



Cleaning CIFs

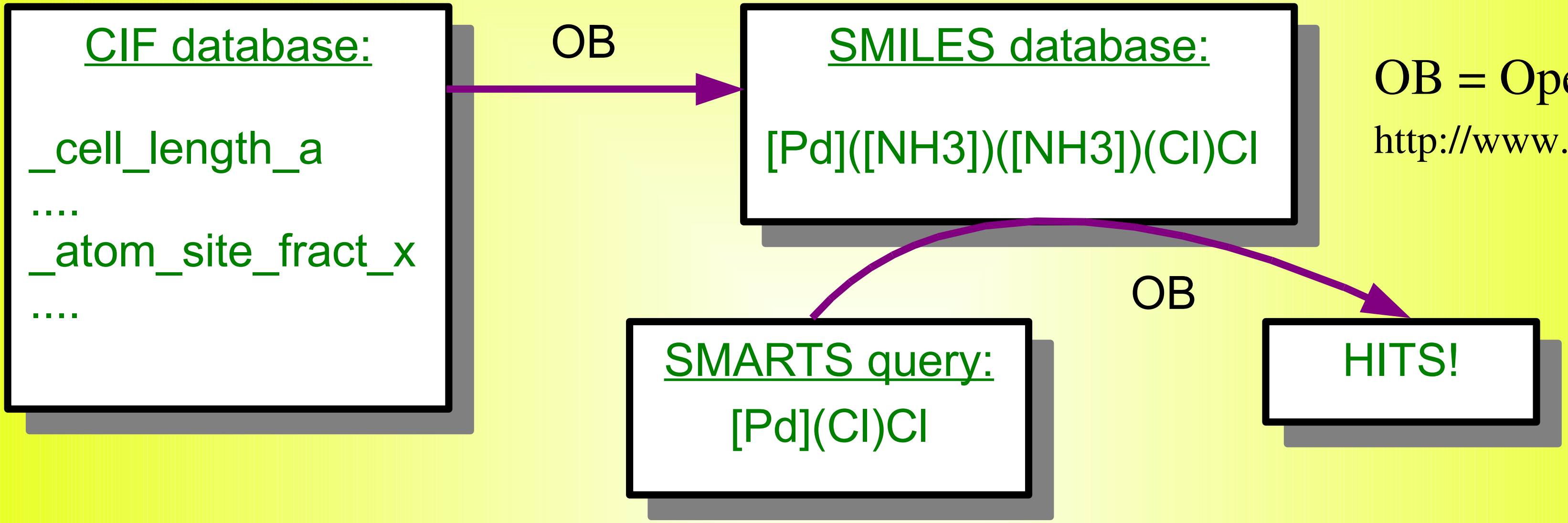
```
data_manyerrors
A very interesting structure

_publ_author_name
   'Miguel Quirós'
_publ_author_address
;
ENTER ADDRESS HERE
;
_symmetry_space_group_name_H-M
P_2(1)/c
_refine_special_details
' broken quoted string into
  two lines'
_geom_special_details
;
forgot closing semicolon
loop_
 _atom_site_fract_x
 _atom_site_fract_y
 _atom_site_fract_z
0.02O3(2)   -0.1723(3)  0.4387(**)
...
```

- CIFs uploaded by contributors or downloaded from journal websites sometimes contain errors.
- CIFs are checked by vcif or other tools. Fixing is automatically made for some common errors (cif-filter) and manually in other cases. The Automatic Deposition Service checks again syntax and presence of essential data.
- Extraction of data for insertion in the MySQL Database (CIF2COD), now automatically done by the Automatic Deposition Service.

Substructure search

- The most usual way for searching chemical compounds in organic and metal-organic chemistry. It is essential for COD being widely used.



OB = OpenBabel:
http://www.openbabel.org

CIF ---> SMILES. A task with technical challenges.

- "Molecule" not equal to "asymmetric unit".
- Atoms with "non-organic" valences.
- Calculate bonds or use CIF bonds?
- Correct bond orders? Aromatic or not?
- How to break polymers?

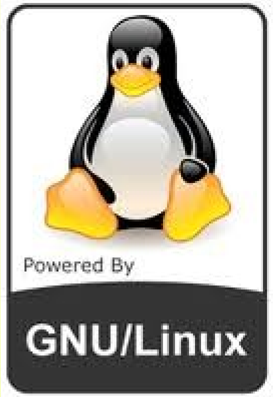
- Customize OpenBabel for COD??

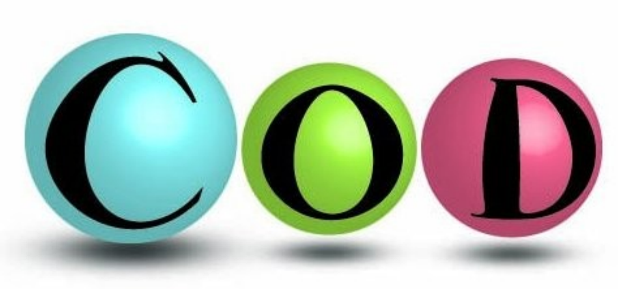
The (long) TODO list

- Downloading from journals websites.
  - Keeping current journals up to date.
  - Develop scripts for journals of other main publishers (Elsevier, Wiley, Taylor, ...).
  - How to get data of the pre-CIF era?
- New functionalities for the automatic deposition system.
  - Structure factors.
  - Improve previously deposited data.
- Setting up more mirrors.
- Improve the search engine.
  - Extract chemical connectivity from more CIFs.
  - Develop an "user-friendly" interface for substructure building and search.
- Any other thing you can dream of (it's open! you can do it!)

CALL FOR VOLUNTEERS!

FREE/LIBRE OPEN KNOWLEDGE





People that build COD

- The COD Advisory Board:
  - Armel Le Bail, Saulius Gražulis, Daniel Chateigner, Robert T. Downs, Peter Moeck, Luca Lutterotti, Hareesh Rajan, Alexandre F.T. Yokochi, Yoshitaka Matsushita, Xiaolong Chen, Marco Ciriotti, Miguel Quirós.
- The Development Group at Vilnius:
  - Saulius Gražulis, Justas Butkus, Andrius Merkys, Adriana Daškevič, Mindaugas Magelevičius, Mindaugas Laganeckas.
  - (Research Council of Lithuania is acknowledged, contract No. MIP-124/2010)
- The Contributors
  - Anyone that at any moment has deposited or edited one CIF at COD.

Please deposit your CIF's!!!