



Institute of Biotechnology

Crystallography Open Database: plans, wishes, perspectives

Saulius Gražulis
Oviedo 2011

Acknowledgements

- COD development group
 - J. Butkus
 - A. Daškevič
 - M. Margelevičius
 - A. Merkys
- BNSTS
 - V. Šikšnys
- X-ray group
 - D. Golovenko
 - E. Manakova
 - R. Sukackaitė
 - G. Tamulaitienė
- COD AB
 - D. Chateigner
 - R. T. Downs
 - S. Gražulis
 - A. Le Bail
 - L. Lutterotti
 - P. Moeck
 - M. Quirós Olozábal

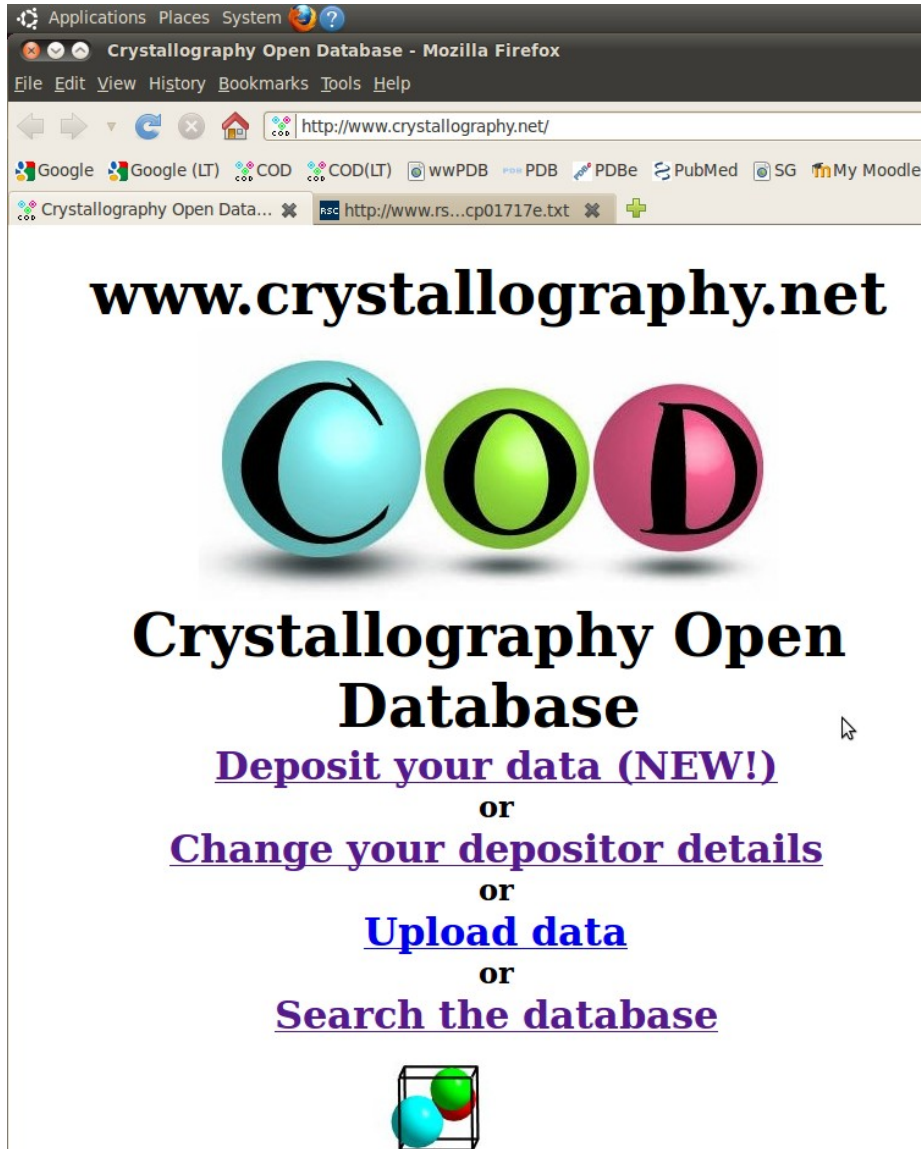
We acknowledge financial support of the:

- Lithuanian Research Council (contract MIP-124/2010)
- Commercial users

Thanks to our numerous

- Volunteers

The COD project



The screenshot shows a Mozilla Firefox browser window with the address bar displaying <http://www.crystallography.net/>. The page content includes the URL **www.crystallography.net**, the COD logo (three spheres labeled C, O, D), and the text **Crystallography Open Database**. Below this, there are four menu options: [Deposit your data \(NEW!\)](#), [Change your depositor details](#), [Upload data](#), and [Search the database](#). At the bottom, there is a small icon of a crystal structure.

An open access database!

- Nearly **149 000** records (August 2011)
- growth 40 000/year
- all records syntactically correct
- records reviewed and updated
- full history of changes

COD goals

- Collect all solved “small unit cell” structures (approx. 10^6 ?)
- Implement a collaboration and data sharing platform for crystallographers
- Become reliable source and repository of structural knowledge
 - What is knowledge?
 - Information that can verifiably guide our actions towards a desired result.

COD mission

- **Provide open access to crystallographic data for all interested parties**
- **Be free like in “free speech”**
- **Facilitate a free exchange of knowledge**

COD search interface

The screenshot shows a browser window with the address bar containing "Crystallography Open Database". The page title is "Search". Below the title, there is a note: "(Output limited to 300 entries maximum, see the [hints and tips](#))". There are two search input fields: "Search by COD ID" and "Search by SMARTS", each with a "Search" button. Below these is a table with search filters. The table has two columns: the left column contains filter names, and the right column contains input boxes for each filter. At the bottom of the table are "Reset" and "Send" buttons.

Crystallography Open Database

Search

(Output limited to 300 entries maximum, see the [hints and tips](#))

Search by COD ID Search

Search by SMARTS Search

Note: substructure search by SMARTS is currently available in a subset of COD containing 25 000 structures.

text (1 or 2 words)	<input type="text"/>
1 to 8 elements	<input type="text"/>
NOT these elements	<input type="text"/>
volume min and max	<input type="text"/>
strict number of elements	<input type="checkbox"/>
<input type="button" value="Reset"/>	<input type="button" value="Send"/>

Done

Collaboration - automatic structure deposition

**Crystallography Open Database
Validation and Deposition Interface**

Please be patient - deposition may take several minutes to complete.

Log in Upload a file Validate data **Deposit structures** Finish

File name	Notes
om9010406_si_002.cif	Storing for deposition: cif-deposit.pl: Structures 4063346 4063347 4063348 were successfully deposited into COD

Continue

Creating a user account

Applications Places System USA Wed May 18, 10:38 AM andrius

Crystallography Open Database: CIF Validator - Mozilla Firefox

File Edit View History Bookmarks Tools Help

http://andrius/~andrius/cod/check_user.php?CODESESSION=llrq3ssar158is7gpub2qrqrg2tm9g

Crystallography Open Databas...

User **Andrius Merkys** does not exist - let's create an account for your (a star symbol * indicates required fields):

Depositor details

* Your account name
Andrius Merkys

* Retype your COD password (must match the previous one)
●●●●●●●●

* Your real name in Latin transliteration (please use only ASCII characters)

Your real name in your national script (please use characters supported by UTF8)

Areas of your scientific interests:

X-ray crystallography Powder diffraction Material identification
 Software development Scientific databases Structural biology
 Other (please specify): _____

Your highest academic degree earned so far:
(none) _____

Institution that has awarded the degree:
Name of the institution

Address of the institution:

ITQs, emerve

Type the two words:

no CAPTCHA™
stop spam.
read books.

Create an account

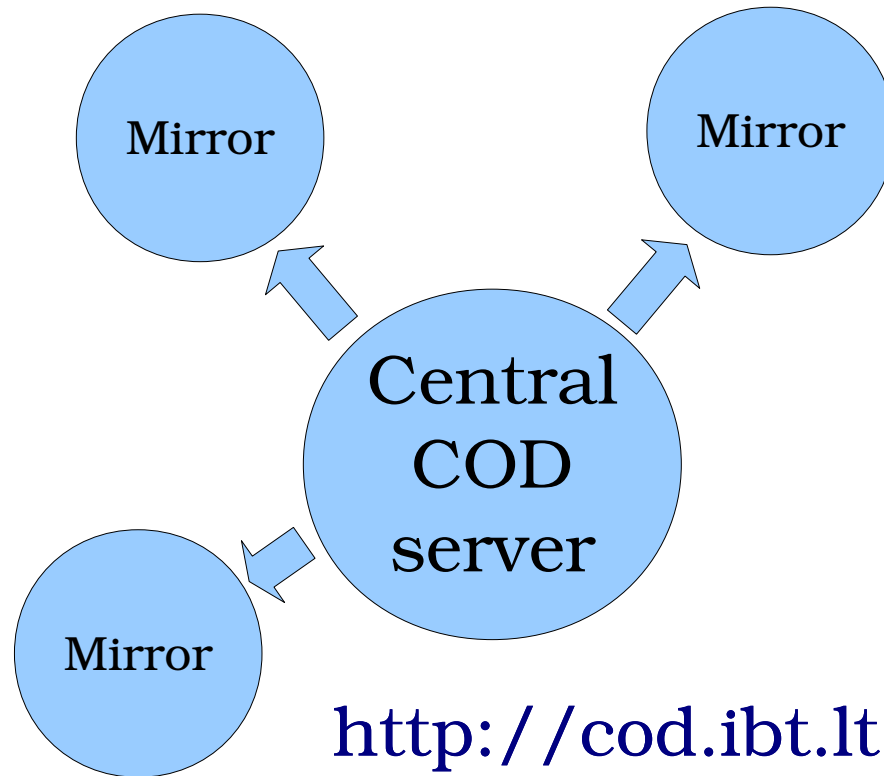
Done

andrius@andrius: ~/c... (IUCr) CIF Definition da... cod-bakalaurinis.odt - ... praktikos-ataskalta.od... cod-bakalaurinis.odp - ... Crystallography Open D...

Current COD architecture

<http://qiserver.ugr.es/cod/>

<http://cod.ensicaen.fr/>



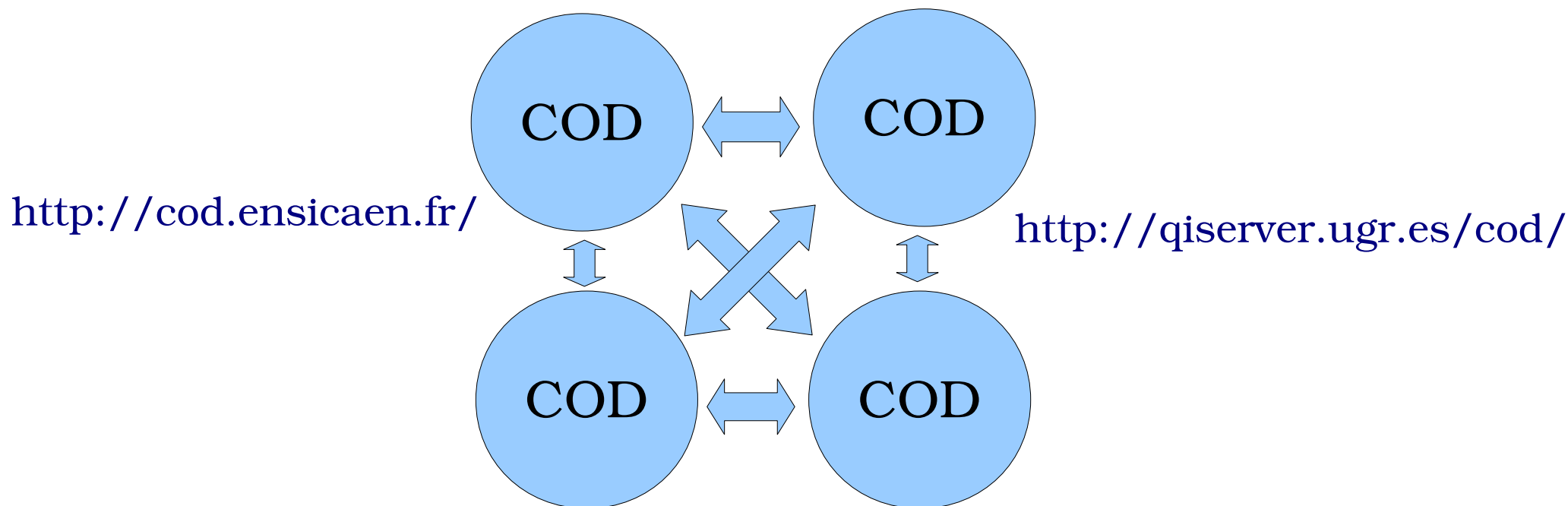
<http://cod.ibt.lt/>

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

<http://nanocrystallography.org/>

Future COD architecture vision

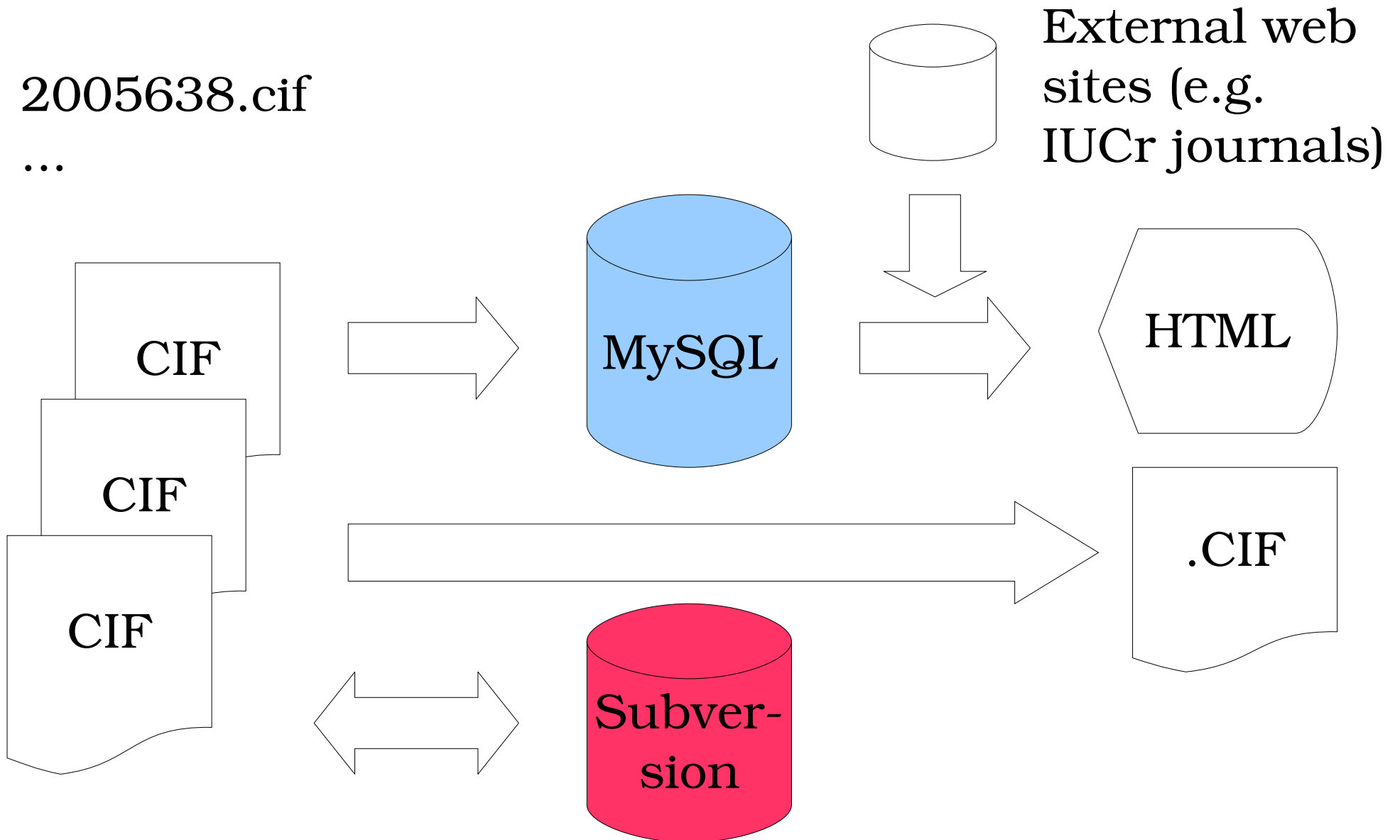
<http://www.crystallography.net/>



<http://nanocrystallography.org/>

```
_cod_server_run_by  
;  
# PUT YOUR NAME HERE  
;
```

COD organisation



Obtaining COD

- Via the Subversion protocol:
 - `svn://www.crystallography.net/cod`
- Via the Rsync protocol:
 - `rsync://www.crystallography.net/cif`
- Via the WWW:
 - <http://www.crystallography.net/>
 - <http://www.crystallography.net/cif/>

COD desiderata

- Syntactically correct data
- Semantically correct data
- Reliable data
- Full coverage of “small unit cell” crystals
- Uniformity (epistemological, methodological)
- Durability

COD data check

- Check syntax, report syntax errors nicely
- Check for data essential for COD (cryst. data, coordinates, chemical info, bibliography)
- *Validate against CIF dictionaries*
- *Run consistency checks (e.g. PLATON, Fobs:Fcalc CC, R values, packing, etc.)*

“Deposit anyway” button

- How to maintain high data reliability standards at low cost?
 - Use stringent automated checks (some false-negatives possible)
 - Have a “deposit anyway” button to send a structure with explanation to a reviewer's mailing list
 - collect +3 votes to deposit, -3 votes to reject.

COD validation - the amount of work

```
mysql> select count(*) from validation;
```

```
+-----+
```

```
| count(*) |
```

```
+-----+
```

```
| 248642 |
```

```
+-----+
```

```
1 row in set (0.09 sec)
```

New COD depositions

depositor_id	nr	min(date)	max(date)
10	14612	2010-11-06	2011-08-19
1	6474	2010-10-16	2011-08-19
9	484	2010-11-24	2011-07-15
55	122	2011-06-27	2011-07-08
43	15	2011-05-11	2011-05-13
41	6	2011-05-06	2011-06-20
64	5	2011-07-21	2011-07-21
14	4	2010-11-21	2010-11-21
16	4	2010-12-11	2010-12-13
39	2	2011-04-28	2011-04-28
4	2	2010-10-18	2010-10-18
44	1	2011-05-17	2011-05-17

12 rows in set (0.15 sec)

Total 21731 structures.

COD data sources

- Donations from laboratories:
 - Mineralogical Society of America
 - Mineralogical Association of Canada
 - Laboratoire de Cristallographie et Physicochimie du Solide
 - Laboratoire de Cristallographie et Sciences des Matériaux CRISMAT
 - Laboratoire des Oxydes et Fluorures, Institut de physique de la Matière Condensée
- Donations by journals:
 - IUCr journals
- Automated Internet crawling
- Collections by volunteers from the peer-reviewed journal supplementary data.

Problems with published data

- Impression: ~40% of published CIF files contain syntax errors; ~1% in the IUCr journals in recent years contain semantic problems...
- “These [i.e. syntactically incorrect and missing data] problems, which affect about 40% of incoming CIFs” (Frank H. Allen, The Cambridge Structural Database: ..., Acta Cryst. B, 2002, **58**, 380 – 388)

Community based effort?

- Wiki (Wikipedia) – like review, error correction, annotation, linking with other resources
- Invited editors and reviewers?
- Depositors as editors and reviewers?

COD applications

- Source of ligands for macromolecular crystallographers
- Collecting statistics (representative data subset?)
- Search-match software
- Teaching
- Software testing and validation

COD prospects

- Docking
- Crystallographic publication validation and review
- Rational drug design
- QSAR
- Materials research
- Semantic web
- Etc., etc., etc...

COD funding

- Government grant funding
- Funding by project participants
- Funding by (industrial) COD users
- *Donations by the community*
- *Contract research*
- *Services*
- Ultimate price/performance ratio!

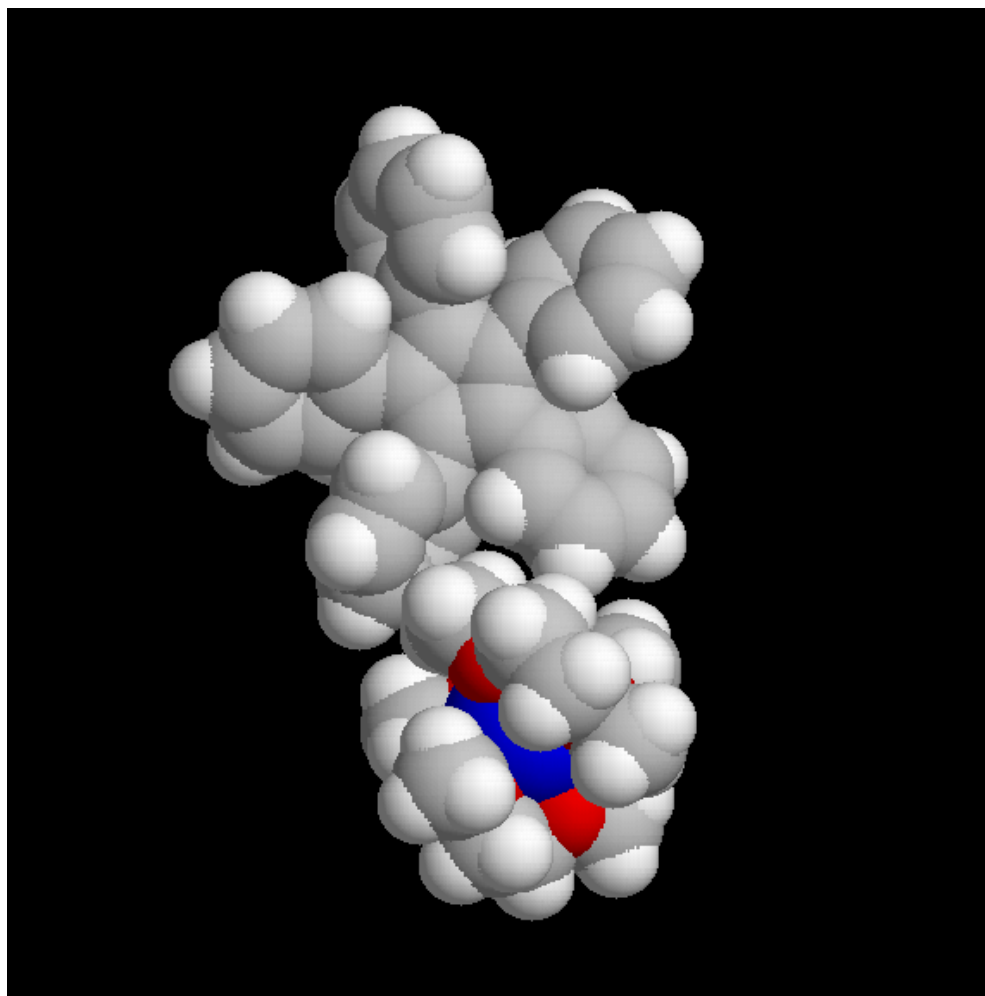
Conclusions (COD)

- A platform for international collaboration is being created
- Added value: syntactically correct, uniform, searchable, open data in one place
- 50 registered depositors
- ~20 000 structures deposited via the Web interface (14% of all COD structures)

Data to knowledge

- Database is not yet knowledge...
- ... but it is nowadays an important prerequisite!
- Semantic webs?
- Automatic inference?

Thank you!



Copyright? What copyright?

- Copyright covers works of authorship (novels, verse, sci. papers, computer programs)
- Copyright covers **only** the expression of the ideas
- Copyright **does not** cover:
 - Ideas
 - (scientific) facts
 - Simple forms (i.e. ones that do not contain individual's “trace of the hand”)

COD copyright policy

- Include data:
 - `_atom_site_fract_x 0.333`
- Exclude potentially copyrighted text:
 - `_publication_text`
;
Introduction

We have solved ...
;