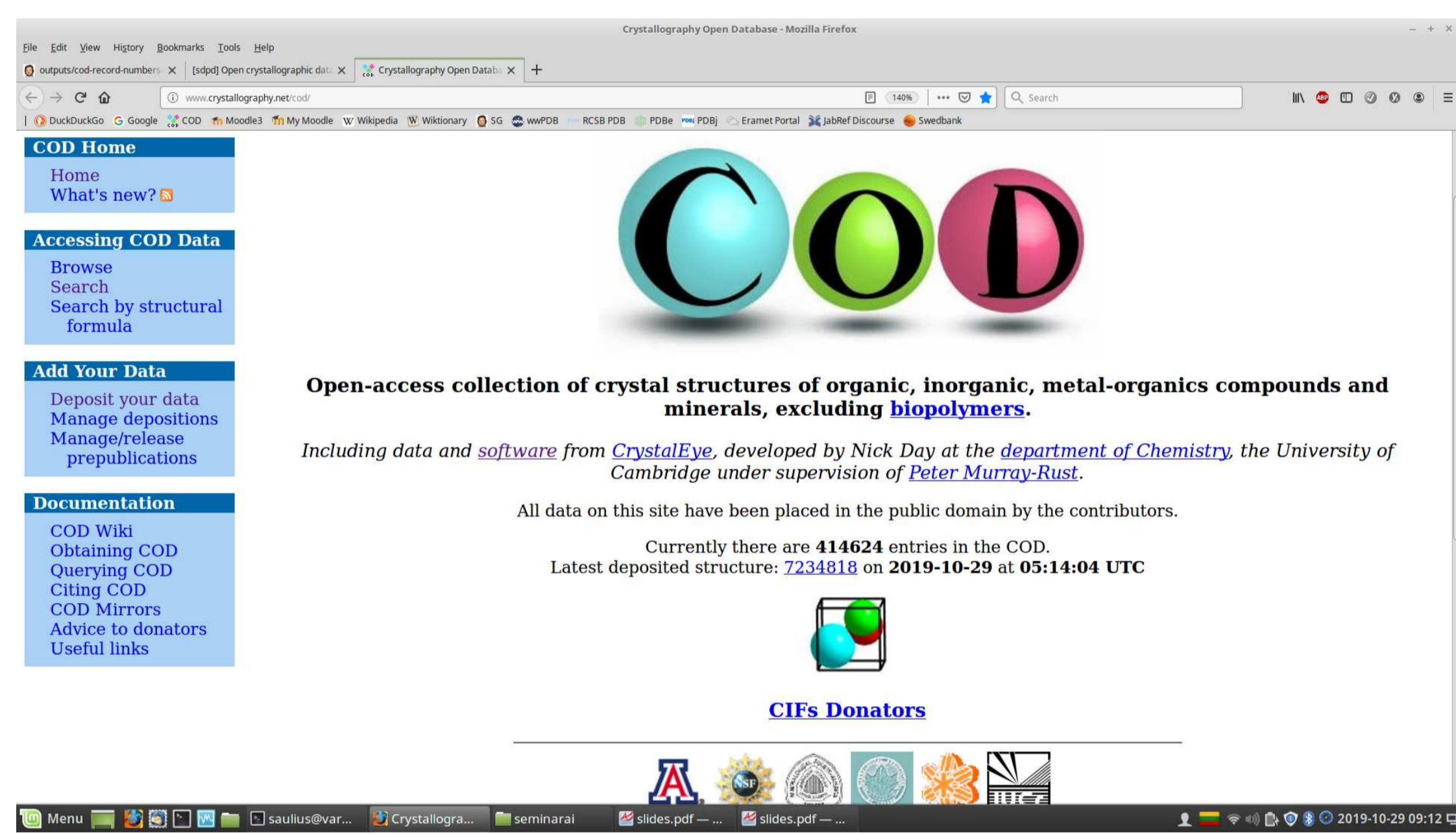


Abstract

Crystallography Open Database (the COD [2]) is the largest FAIR crystal structure collection available. Over 16 years, data collection and curation was performed using automatic and semiautomatic tools based on Unix architecture, using standard and custom built command-line programs, Unix pipes, shells and GNU Make, as well as editing text files. The collection of COD programs used for CIF [3] processing and COD data curation is released as F/LOSS in the `cod-tools` package [5]. To make all changes traceable and computations reproducible, all changes are recorded in a Subversion [1] repository. All curation history since 2007-11-30 is available at svn://crystallography.net/cod.

Introduction

Crystallography Open Database (the COD) is the largest FAIR crystal structure collection available.
<http://www.crystallography.net/cod>



Data curation methods

Data curation records are visible as Subversion revision logs, both in the COD repository and on the Web:

https://www.crystallography.net/cod/1100110.html

▼ Version history

Revision	Date	Message	Files
184082 (current)	2016-06-30	cod/ (saulius@kolibris) Marking entries 1100110 and 9007968 as duplicates of 2203515.	1100110.cif
184080	2016-06-30	cif/1/10/ (saulius@kolibris) Marking COD data sources and updating COD depositor comments in 01/1100110.cif.	1100110.cif
184079	2016-06-30	cif/1/10/ (saulius@kolibris) Merging the 01/1100110.cif CIF data with the data from the IUCr supplementary file describing the same structure.	1100110.cif
184078	2016-06-30	svn cat 1100110.cif \ cif_merge - <(curl -sSL http://scripts.iucr.org/cgi-bin/sendcif?wn6225sup1 \ cif_filter --renumber --start-data 1100110) \ cif_filter --exclude-empty-tags --add-cif-header \$(codid2file 2105696) \ sponge 1100110.cif	1100110.cif

Interactive Unix command line is often used to assess COD structures . . . :

```
curl -sSL http://crystallography.net/cod/7224530.cif | \
cif_molecule -i --p1 | buffer jmol 2> /dev/null &
```

```
curl -sSL http://crystallography.net/cod/2227697.cif | \
cif_molecule | cif2molecule | obabel -i SDF -o SMI | \
cdkdepict | ~/src/xmlsplit/xmlxargs konqueror 2> /dev/null &
```

. . . and curate them *en masse*:

```
r134283 | antanas | 2015-03-24 01:00:56 +0200 (Tue, 24 Mar 2015) | 9
lines
```

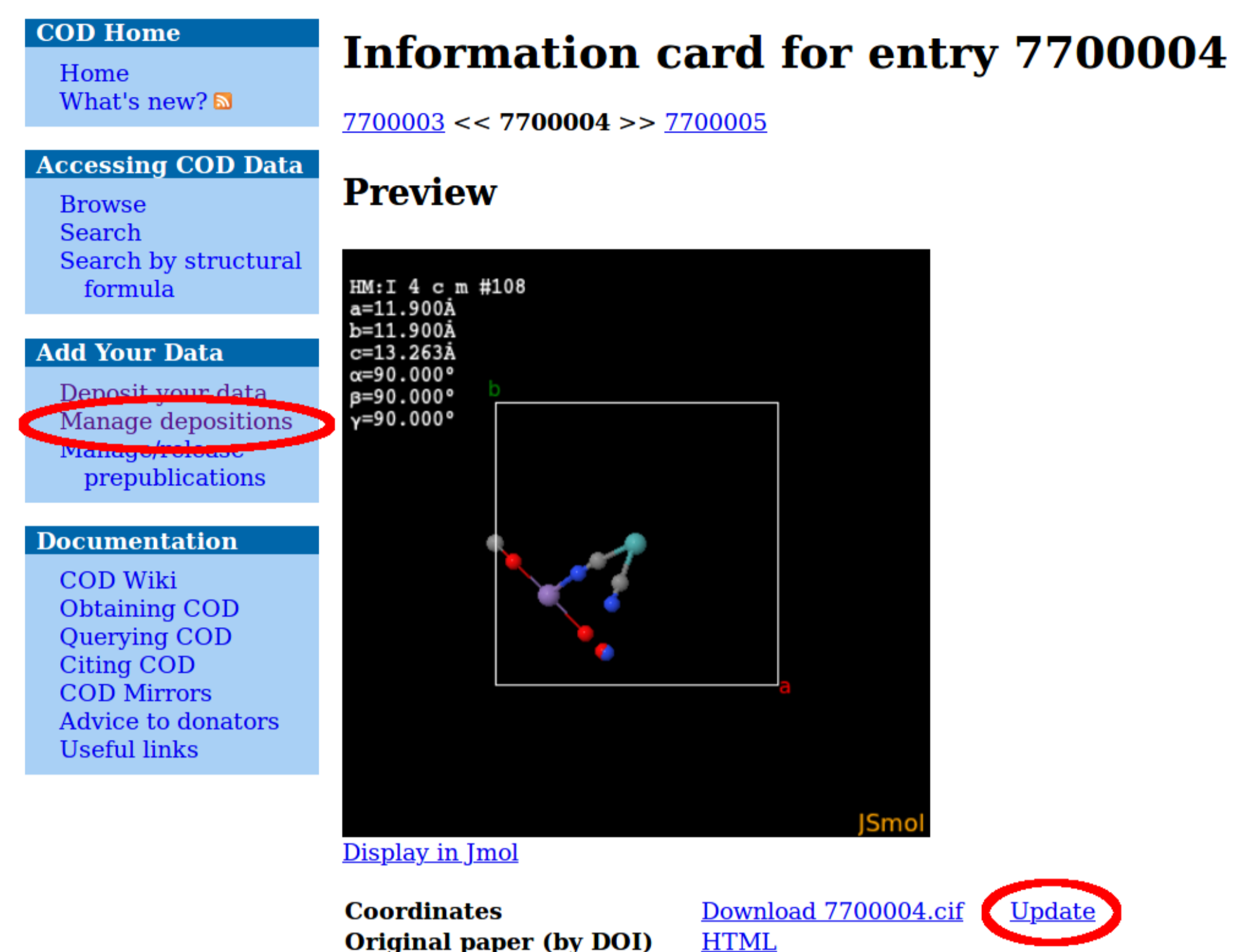
```
cif/ (antanas@echidna.ibt.lt)
Replaced '_reflns_observed_expression', '_atom_type_scatter_imag' and
'_atom_site_calc_flags' tags with their correct tag counterparts
('_reflns_threshold_expression', '_atom_type_scatter_dispersion_imag',
'_atom_site_calc_flag'). The following command was used on the CIFs
containing the tags:
```

```
find ~/struct/cod/cif -name '*.cif' \
| xargs -n1 -I{} sh -c \
'cif_correct_tags \
--r cod-tools/trunk/perl-scripts/inputs/replacement_tags.lst \
{} \
| cif_filter -h {} | sponge {}'
```

A total of 124 files were changed.

Data curation methods (cont.)

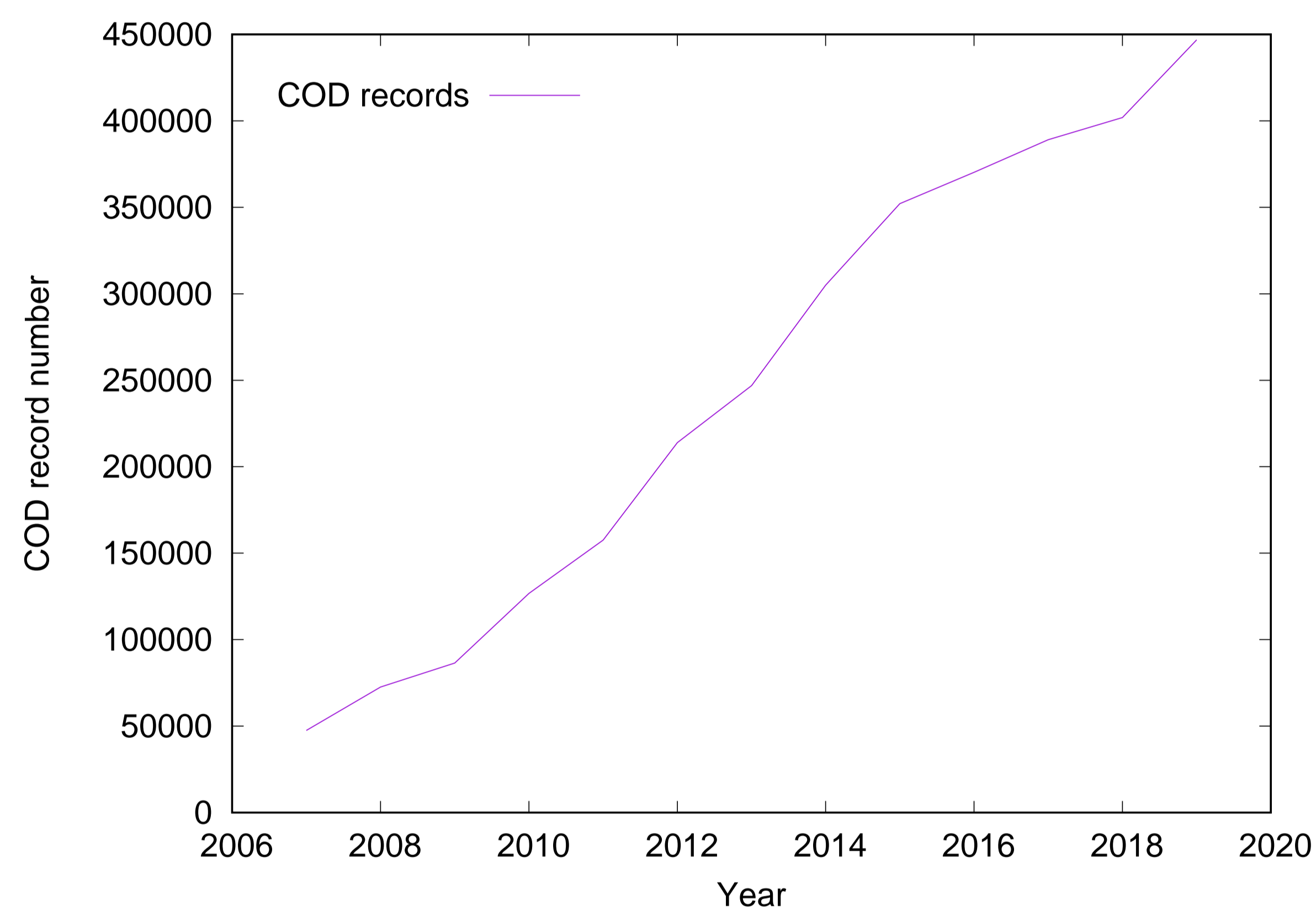
Records can be curated (updated) using Web interface, one record at a time:



Jmol [4] molecule viewer is used to assess the crystal structures by the curators.

Results

COD is on-line for 16 years, increased 8-fold over the last 10 years; currently it contains over 440 000 records (2019):



Conclusions

- ▶ Using F/LOSS, Unix-architecture based tools allow one to build, curate and maintain an open scientific data collection;
- ▶ Unix architecture tools allow both manual maintenance of data as well as integration into automated workflows;
- ▶ version control systems, traditionally used for software development, are instrumental for scientific data curation as well.
- ▶ Web-based sites provide easy access for users who would not run Unix command line for the task involved; easy and safe integration of Unix tools into interactive Web pages would be very welcome.

Bibliography

- [1] Collins-Sussman et al. *Version Control with Subversion*. 2011, <http://svnbook.red-bean.com/>.
- [2] Gražulis et al. Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research*, 40:D420–D427, 2012, <http://nar.oxfordjournals.org/content/40/D1/D420.abstract>.
- [3] Hall et al. The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A*, 47:655–685, 1991, <http://dx.doi.org/10.1107/S010876739101067X>.
- [4] Hanson. Jmol – a paradigm shift in crystallographic visualization. *Journal of Applied Crystallography*, 43:1250–1260, 2010, <http://dx.doi.org/10.1107/S0021889810030256>.
- [5] Merkys et al. COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *Journal of Applied Crystallography*, 49(1), Feb 2016, <http://dx.doi.org/10.1107/S1600576715022396>.

License

This poster is distributed under Creative Commons Attribution 4.0 International license 