D420–D427 *Nucleic Acids Research, 2012, Vol. 40, Database issue* *Published online 8 November 2011*
doi:10.1093/nar/gkr900

Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration

Saulius Gražulis[1,2,*], Adriana Daškevič[1], Andrius Merkys[1], Daniel Chateigner[3,4,5], Luca Lutterotti[6], Miguel Quirós[7], Nadezhda R. Serebryanaya[8,9], Peter Moeck[10], Robert T. Downs[11] and Armel Le Bail[12]

[1]Department of Protein – DNA Interactions, Vilnius University Institute of Biotechnology, Graiciuno 8, LT-02241 Vilnius, [2]Department of Mathematical Computer Science, Vilnius University Faculty of Mathematics and Informatics, Naugarduko 24, LT-03225 Vilnius, Lithuania, [3]Université de Caen-Basse Normandie, UMR 6508 CRISMAT, F-14032 Caen, [4]ENSICAEN, [5]CNRS, UMR 6508 CRISMAT, F-14050 Caen, France, [6]Department of Materials Engineering, University of Trento, via Mesiano, 77 - 38050 Trento, Italy, [7]Departamento de Química Inorgánica, Facultad de Ciencias, Universidad de Granada, 18071 Granada, Spain, [8]Technological Institute of Superhard and New Carbon Materials, Ministry of Education and Science of the Russian Federation, ul. Tsentralnaya 7a, [9]Institute of Spectroscopy, Russian Academy of Sciences, Fizicheskaya ul. 5, Troitsk, Moscow oblast 142190, Russia, [10]Department of Physics, Portland State University, PO Box 751, Portland, OR 97207-0751, [11]Department of Geosciences, University of Arizona, Tucson, AZ 85721-0077, USA and [12]Université du Maine, Laboratoire des Oxydes et Fluorures, CNRS UMR 6010, Avenue O. Messiaen, 72085 Le Mans Cedex 9, France

Received August 14, 2011; Revised September 25, 2011; Accepted October 5, 2011

ABSTRACT

Using an open-access distribution model, the Crystallography Open Database (COD, http://www.crystallography.net) collects all known 'small molecule / small to medium sized unit cell' crystal structures and makes them available freely on the Internet. As of today, the COD has aggregated ~150 000 structures, offering basic search capabilities and the possibility to download the whole database, or parts thereof using a variety of standard open communication protocols. A newly developed website provides capabilities for all registered users to deposit published and so far unpublished structures as personal communications or pre-publication depositions. Such a setup enables extension of the COD database by many users simultaneously. This increases the possibilities for growth of the COD database, and is the first step towards establishing a world wide Internet-based collaborative platform dedicated to the collection and curation of structural knowledge.

INTRODUCTION

Modern research experiments provide us with large amounts of digital data. Computers permit efficient storage and processing of this data, allowing new scientific inferences to be drawn from them. Effective computer processing, however, requires that the data are collected and organized in a homogeneous and coherent manner, yielding with necessity some sort of a database.

The need of uniform data representation (1) and of comprehensive databases in the field of crystallography was recognized a long time ago. Databases for organic compounds (2) as well as in other areas of crystallographic research, i.e. structural biology (3), metals and alloys (4), inorganic compounds (5), minerals (6) and powder diffraction data for material identification purposes (7) were created.

Traditionally, in a pre-Internet era the databases were distributed using a subscription-based approach, similar to the one used by the scientific journals of those days. Users of a database were required to pay a license fee for the usage of the databases. In modern times, however, when the Internet is becoming more and more ubiquitous, the limitations of such a subscription licensing model can be overcome. Not surprisingly, novel online services

*To whom correspondence should be addressed. Tel: +370 684 49802; Fax: +370 5 261 2116; Email: grazulis@ibt.lt

aggregating open data are emerging, like Crystal Eye (http://wwmm.ch.cam.ac.uk/crystaleye/) or PubChem (8).

Over the Internet, data can be transmitted instantly between computers, processed simultaneously at many places and flexibly combined from many sources. Such technical possibilities open new ways for the development of science, where more eyes and more brains thinking on a subject as a rule have more success than a single person in isolation, and, when working together, create more new knowledge to everyones advantage; growth of scientific knowledge is cumulative and faster (9).

It is out of such considerations that the Crystallography Open Database (COD) project originated in 2003. The goal was (and remains) to collect all known 'small molecule/small to medium sized unit cell' crystallographic structures in one high-quality open-access database. Towards this goal, 150 000 structures have already been collected and are available for search and download at http://www.crystallography.net. The collection of new entries is done continuously, resulting in a growth rate of ~40 000 entries/year. There is the technical capacity to double this rate. Automated data checks and, where necessary, manual inspections by COD maintainers ensure that all our entries are syntactically correct, complete with the most essential data and, whenever possible, free of semantic errors.

MATERIALS AND METHODS

Data storage and retrieval

Database contents. The COD crystallographic database collects all crystal structures of 'small to medium sized unit cell' crystals—structures of organic, inorganic, metal-organic compounds and minerals. The structures are to be determined using state of the art experimental techniques: single crystal or powder diffraction of X-ray or gamma photons, neutrons, electrons or other particles; or calculations using the proven theoretical methods such as the density functional theory.

The distinctive feature of the database contents is that the structures should be determined without prior assumptions of molecular geometries, except perhaps for a limited number of parameters of hydrogen atoms or of disordered, poorly defined 'minor regions' of the structure (such as solvent molecules, peripheral groups) where parameters that are well defined from a large number of previous studies may be used. For structures determined using single crystal diffraction, for example, this means using full matrix least squares refinement, at least for non-hydrogen atoms, without assumptions of interatomic distances or angles. For theoretical calculations, fairly general assumptions about the base wave function set should be made. Thus, structures of biological macromolecules as a rule should not be deposited into the COD database, since their refinement usually uses prior knowledge of bond parameters and their distributions, and this knowledge is derived from the independent structures of 'small molecules' with similar chemistry.

This decision on the content of the COD has been taken for two reasons. The epistemological reason was to provide in the COD sets of data that were obtained using similar initial assumptions and are thus comparable with each other. The practical reason was that structures of large biological macromolecules are already stored very efficiently in the open-access databases PDB (10) and NDB (11). The COD inclusion criteria are satisfied by all newly published structures in peer-reviewed chemical and crystallographic journals.

The exclusion of data based on their belonging to a certain class of materials, e.g. to minerals or organic compounds, was considered to be inappropriate for modern applications. Indeed, currently computers and storage facilities are powerful enough to accommodate all data in the fields of organic and inorganic chemistry, metal-organic chemistry and mineralogy. The COD policy is to provide tools for selecting an appropriate subset of data. Such selection is much faster and arguably easier than collecting data from several independent database sources.

Structure identification. Each structure deposited into the COD receives a unique seven-digit number, called COD number. A COD number identifies a particular instance of a structure determination. As a rule, COD does not accept duplicate structures. If, however, two structures of the same compound are published in two different peer-reviewed journals, both can be deposited to COD and receive distinct COD numbers. Similarly, if a structure of some compound is solved with significantly higher resolution or using more advanced methods, this new instance will be deposited into the COD under the new COD number. The COD thus intends to capture and represent the current state of crystallographic knowledge. 'Old structures' may be filtered out using crystallographic quality indicators, special COD markers or publication dates.

Currently, the COD deposition software uses a very simple algorithm to detect duplicates by comparing chemical brutto formulae and crystal unit cells. For an incoming structure, structures with matching chemical formulae are selected. Unit cells are compared as provided by the authors, taking into account measurement errors. Chemical formulae are checked for exact match, and both formulae provided by the authors and computed from the crystal structure contents (atomic coordinates and Z values or crystal density and molecular weight) are compared with each other (resulting in four comparisons if author-provided formulas are present in both files). If experimental conditions (temperature, pressure or sample histories) are provided, these are also checked for equality (numeric values are compared as numbers, text values—for an exact match). If all compared values match, an incoming structure is declared to be a potential duplicate of the exising one.

This simple algorithm is not yet suitable for searching of similar compounds in different unit cell settings. The main purpose of the duplicate detection is to exclude 'technical' duplicates, i.e. to prevent exactly the same structure file from being deposited again, and it proved to be efficient.

File tree layout. The data in the COD are stored in the Crystallographic Interchange File/Framework (CIF)

format (1), one structure per file. A file name is chosen to be the COD number of the corresponding structure. Each file contains all data necessary to describe the structure, interpret experimental data and find the corresponding publication(s). If necessary, bibliographic and experimental condition information is disseminated into all relevant CIFs. If present, bibliographic data values are taken from the `data_global` section and, if the corresponding tags are not present in the coordinate data sections, they are inserted there. Values in the coordinate sections, however, take precedences over the values given in the `data_global` section. Only bibliography data is treated this way (_publ_author_name, _publ_section_title and _journal_... CIF tags). In addition, whenever external bibliography is supplied in the process, it has precedence over the `data_global` or coordinate data block values. After this, each COD file can be used independently from the others.

The COD files are stored into a split 'prefix directory tree'. On the top level, directories numbered 1 to 9 are created, each holding files from n000000.cif to n999999.cif, where n is the digit corresponding to the directory name. In the early versions of the COD, files were stored immediately in these subdirectories. As the number of records in the COD grew, however, managing tens of thousands of files in one directory became inconvenient. Therefore, the current COD file layout introduced two intermediate layers of directories, where the second and the third digit was used for the second-level directory name, and the fourth and the fifth digit were used to name the third-tier directories. For example, COD file `1234567.cif` would be stored into the directory location `1/23/45/1234567.cif`. Such scheme ensures that no directory contains more than a hundred entries (files or subdirectories), but at the same time the tree can accommodate all structures to be stored in the COD. To list all COD data files, a recursive subdirectory scan should be performed (see Supplementary Data 'COD-scan.sh' for a example).

Use of the version control software. Ideally, data deposited into the COD should be written to the database just once and not changed afterwards. In real life, such an ideal is not achievable. Occasionally, typing errors or inaccuracies are spotted in structures that need to be corrected to maintain the quality of the database. New fields describing a structure are sometimes added and they need to be inserted into old records as well. Some missing data might be acquired later, after the original deposition of a structure. For example, measurement temperature might be missing from the older CIFs, but provided in the paper text. In such cases, a field needs to be inserted into a CIF file, which should retain the original COD number.

Since such changes may affect inferences drawn from the data of the COD database, and thus, for the reproducibility of scientific computations, previous versions of the database need to be maintained. In addition, some changes may inadvertently introduce new mistakes that need to be identified and corrected. A record concerning database changes must, therefore, be maintained in a scientific database.

For the amount of data and the nature of the files that the COD deals with, a nearly ideal tool for managing change histories is a version control system, one of the kind routinely used by software development teams. The COD project uses the Subversion (12) version control system.

Thus, each structure file in the COD may be available in a number of revisions, each revision identified in the Subversion repository by a revision number. Knowing the COD number of a structure and a desired revision number, one can unambiguously restore a sequence of bits stored in a particular COD file at a given point in history. By default, the most recent revision is returned by the server, but any previous revision of a file or of the whole tree can be extracted if requested.

Apart from the structure files themselves, the COD Subversion repository contains the versioned COD Website software and the SQL data table dumps. Thus, the repository is necessary and sufficient to construct a working COD Web site and to restore the whole history of the database.

SQL database layout. The COD crystallographic database provides tools for selecting an appropriate subset of data by many structural parameters. The selection is powered by the SQL database, which is accessible by several protocols.

Each structure in the COD crystallographic database is described as an entry in the SQL database. Such entries are generated automatically from the COD CIFs and consist of bibliography and parameters that describe the size and contents of an unit cell, space group, the diffraction experiment and the quality of the data, such as the R factor and goodness-of-fit parameter. Structures with disorder are flagged with the 'has disorder' flag in the COD SQL 'data' table; all CIF data describing disorder are preserved in the COD CIF.

Parameters of the unit cells are stored together with their respective measurement precisions. Both Hermann-Mauguin and Hall symmetry space group symbols are included in the entries. To make sure that the same spelling of the symmetry symbols is used, and that they are always present, upon deposition Hermann-Mauguin and Hall symbols are derived from symmetry operators using table generated with the help of the CCTBX library (13). Recently, the SQL database layout was extended to include parameters describing the radiation type and wavelength. Searches using these criteria can yield the COD numbers of structures satisfying the search criteria.

There are several ways to access the structural parameter SQL database. The Web interface allows the selection of data by unit cell parameters and contents. Numerical values can be searched between defined bounds. The MySQL protocol allows to directly connect to the COD SQL database with permission to read and search the whole dataset using the standard SQL query language. Finally, one can download the SQL database dumps and build it on a local site.

Data collection and deposition

Data sources. The COD predominantly collects data published in the peer-reviewed scientific press. Since many contemporary publications are accompanied with data in electronic form, most often in the CIF format, the data deposition from such sources can be automated to a significant degree, and manual intervention is only needed when syntactic errors are detected or when some information (most often, bibliographic data) is missing. Data from older publications, when they present significant interest, need to be keyed in manually.

It has been estimated (http://www.icsti.org/IMG/pdf/ICSTI-IUCr-finalreport.pdf) that a large amount of crystallographic data remains unpublished. The COD advisory board invites all researchers, therefore, to publish such results as personal communications to the COD so that they are not lost to mankind. As the COD might in such case be the only safeguard of the scientific data quality, the checks performed for fully automated deposition of such structures are stricter than for published structures. In particular, all IUCr data quality tests must be satisfied. In the future, we plan to extend our Web deposition software so that peer review will be possible to decide whether a structure is suitable for deposition into the COD. Thus, the COD should become not only the repository of the published crystallographic information, but also provide an extra layer of review and data correction capabilities.

Data deposition Web site. A website for deposition of data to the COD was developed recently, allowing the scientific community to participate directly in extending the COD data collection. The deposition web page can be accessed via the COD main page, following the link 'Deposit your data'. All processes concerning the insertion of new data were automated. The detailed data processing pipeline was described previously (14). Currently, we have merged this pipeline with the web interface so that it can be used by all interested researchers cooperating on further COD developments. Acknowledging a wide concern about the preservation of the original research data and in line with recent IUCr publication standards (http://www.iucr.org/home/leading-article/2011/2011-06-02#letter), the COD now accepts structure-factor (Fobs) files in addition to atomic coordinates.

Crystal structures are accepted either as published or in some pre-publication format or as personal communications. Each structural information file, supplied for deposition, is checked for syntactic errors and validity against IUCr data validation criteria (http://journals.iucr.org/services/cif/checking/autolist.html). Files that contain errors or do not match these criteria can be edited interactively and validated again. Deposited structural information of the published data and personal communication depositions are put into the public domain and become immediately accessible. Structures that are deposited prior to publication are not released until the corresponding publication is issued or until the hold period expires. In the latter case, the authors are contacted regarding the further processing of their data.

Manual data curation. Although a great amount of checks and corrections are performed automatically, not all mistakes can be corrected or even detected using automated systems. The COD permits combination of automated checks and manual corrections, performed when necessary by a crystallographer or computer scientist. Recently, a validation against the IUCr core CIF dictionary and a check using the COD deposition tools of all COD CIFs was performed on all of our data. The resulting comments and error messages were stored in a COD database table.

From this table, a list of the most prominent dictionary violations were picked, and the errors were either fixed manually or with the help of specifically developed Perl (15) scripts. In this manner, ~46 000 CIFs were corrected (of which 694 CIFs were fixed manually) so that ~31 000 files do not have validation issues any more. Such validation process will be done on a regular basis so that future releases of the COD should have significantly fewer semantic errors. So far the corrections are done by the COD project participants, who have edit access to the COD Subversion repository. In the future, however, we will implement the possibility to correct COD CIFs using the COD Web interface, thus enabling all registered COD depositors to improve the records. We expect that this world-wide collaborative effort will result in a very high-quality curated database, adding significant value to the current automatic checks.

The general data curation policy of COD is, of course, 'do not invent data'. We view a COD record as a COD depositor's statement that the specified authors have made the reported claims about their determined crystal structure in their particular publication. COD maintainers strive to ensure that the COD statement is correct, i.e. that COD files indeed convey what the authors intended to say. To this end, the COD data curation process may change data representation when the data are clear to a qualified crystallographer but do not match formal requirements of the CIF framework. As an example, temperatures specified in degrees Celsius with a unit of measurements given in a text string are transformed to Kelvins without the measurement unit, as mandated by the IUCr Core CIF dictionary (e.g. http://www.crystallography.net/2006444.cif), or the case of enumerator values are changed to match the dictionary (e.g. http://www.crystallography.net/2000571.cif). Such obvious changes are performed on the sole discretion of the COD curators and sometimes even automatically. In the cases of ambiguities, we consult original publications or contact the depositing authors. All changes are documented in such CIFs and in their Subversion log records (e.g. http://www.crystallography.net/2300355.cif). We have to keep correspondence to the minimum, however, since we have little manpower so far dedicated to such tasks. Apart from such attempts to restore the author's original intent, the COD does not introduce any other changes; of course, COD does not re-refine structures, and the reported coordinates are changed only by the authors or after a confirmation from the authors.

RESULTS

Growth of the COD database content

Right after the COD creation, the first large set of data came from the American Mineralogist Database (16) which is the main source for our structures of minerals. In 2007, the IUCr allowed all crystallographic databases (including the COD) the free download of IUCr published CIFs. (The inclusion of these CIFs allowed the COD to reach the 50 000 entries milestone that year.) The IUCr CIFs have been included regularly and almost automatically since then. Data from several other journals were provided by authors or volunteers who donated data from their private structure collections. In the last few years, CIFs are being downloaded from several journal websites using gentle web crawler scripts. The scripts introduce delays to avoid any significant loads on these websites and identify the COD as the downloader so that the publishers may easily contact the database maintainers if they wish. This allowed the COD to quickly reach 100 000 entries. At the time of the writing of this article, the COD has grown to ~150 000 entries.

Search and data retrieval capabilities

Search using the web interface. The COD website allows for the searching of the database using queries such as unit cell parameters, chemical composition and bibliographic data (authors, journal names, paper titles). For chemists, a substructure search using SMARTS strings and SMILES structure representations has been implemented. Currently, the free software package Openbabel (http://www.openbabel.org) is used both for the CIF → SMILES transformation and for the actual searches. But, even if this is an excellent tool, the transformation task is intrinsically difficult due to several reasons (e.g. disorder, 'molecules' not equal to 'asymmetric units', missing H atoms, ambiguous bond orders or aromaticity, non-standard valences and so on). SMILES need, therefore, to be checked and in many cases corrected. This task has been done for a fraction of the COD database (≈26 000 at present) for which a preliminary substructure search engine is available at the COD website.

Results of all searches are displayed in a simple tabular form on the result page (Figure 1). In addition to descriptions of the found structures, a link to download all structures in a ZIP file is provided. Also, a list of COD numbers for the found structures can be downloaded in a plain text format. For each individual structure, links to download CIFs with coordinates and structure factors are provided when these items are available. If a publisher provided the COD with the relevant linking information, a link to the original paper that published the found structure is also present. Each structure can be visualized and described in an extended table using the information card format (Figure 2), after following the 'COD ID' link.

Ways to download the COD contents. In addition to the online access via the WWW and the search possibilities outlined above, users have an option to download the whole COD content for offline processing. The COD is available for download and update via the Subversion (12), Rsync and as a single archive file via the HTTP protocol.

The Subversion protocol gives anybody an anonymous read-only access to the public COD database repository.

Figure 1. Typical search results obtained over the Web interface.

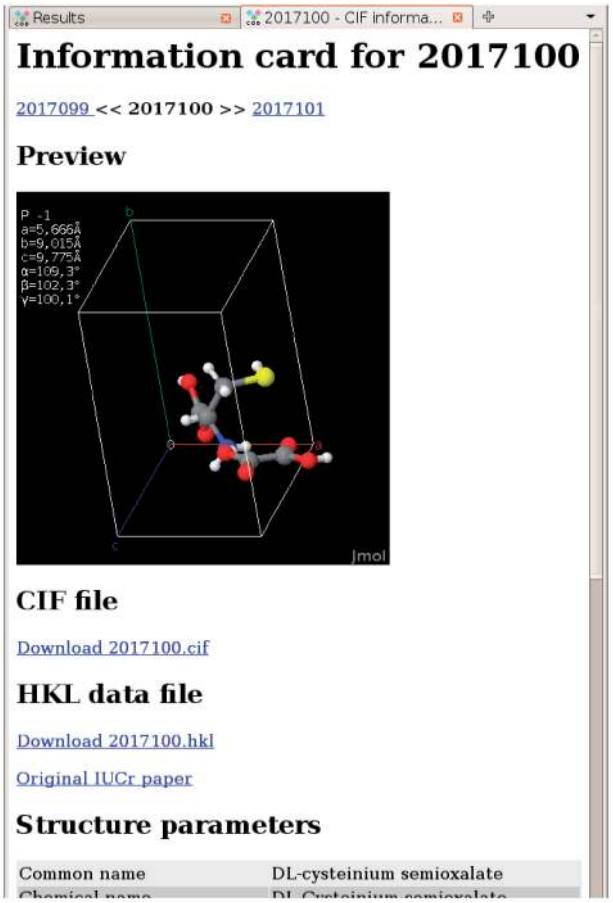

Figure 2. Extended description of a single structure.

This repository contains all information that the COD possesses—CIFs, Fobs data, MySQL table dumps, SMILES strings for the stored structures and the corresponding Website software. One can also obtain all log records provided by COD maintainers from the repository. Subversion is intended to be the primary distribution channel for those who wish to have all COD materials at hand. In fact, mirror sites are created by checking out the COD repository, starting the Web server from the working copy, and later updating this working copy at regular intervals. Subversion gives access to the complete history of changes of the COD files and permits to synchronize the checked out data in a very efficient way by transferring only compressed differences between the old and new COD files. At the moment, Subversion repository can be accessed at the URI svn://www.crystallography.net/cod. Towards the end of this year, access via http://www.crystallography.net/cod will be launched.

We realize that the Subversion software may be unsuitable for some of the COD users. Thus, we provide essential COD data via different protocols. Coordinates and structure factors can be accessed using Rsync tool at URIs rsync://www.crystallography.net/cif and rsync://www.crystallography.net/hkl, respectively. CIFs can be downloaded and later updated from these read-only sources. Examples of commands to download the COD file collection and to search the COD file tree for the required files is provided in the Supplementary files.

The whole COD data set can be downloaded as one packed file in the .zip, tar.gz or .tag.bz2 formats using a http protocol with the help of a web browser or a command line tool such as `wget` or `curl`. To save server disk space and bandwidth, the contents of these files are restricted to atomic coordinate CIFs and MySQL table dumps. Fobs data, SMI files and software are not distributed in the archive files due to their size. Moreover, these files are updated only at major releases of the COD, which happen less often than COD repository updates. One must also take into account each update of the COD as using pre-packed archives requires downloading all structures, old and new. Thus, one should resort to the .zip file download only when the synchronization options, Subversion or Rsync, are unavailable.

Improvement of quality of the COD records

Syntactical correctness. Syntactic correctness is a requirement for a CIF being included in COD. So, it is checked and fixed if necessary for all files before they are accepted in the database. Initially, all files are checked with an error-correcting CIF parser developed for the COD project. Some syntax errors, such as missing closing quotes on the same line, missing 'data_' headers, missing comment symbols at the beginning of the CIF are fixed automatically. Files that enter via the automatic deposition system are presented to the depositor and the depositor is asked to edit the file if syntax errors are found. Files downloaded from journal sites are usually correct if corresponding to recently published papers, especially if journals ask authors to check files at the checkcif IUCr facility. Still, occasionally syntax errors are found and fixed.

After the deposition, the COD collection is checked routinely at regular intervals with the help of the freely available program vcif2 (17) to find any errors that might slip through due to a bug in the COD parser.

Semantic checks and corrections. Regarding semantic correctness, the errors are much more difficult to detect and fix. Nevertheless, the basic crystal data (chemical formula, unit cell parameters, space group related tags, coordinates) are specially checked. The Hall and Hermann-Mauguin space group symbols are regenerated from the symmetry lines, replacing the original entries if necessary. The possible presence of non-filled templates created by the software ('ENTER AUTHOR NAME HERE') is also detected. The correspondence of cell lengths and angles to the spacegroup symmetry constraints are checked as well.

Reliability and availability

For a resource to be usable as a source of scientific data, it must be available at all times (when it may become necessary to reproduce or validate inferences made based on this resource). The COD offers both physical and organizational security of the collected data.

Physically, COD servers reside on mirrored disks which are backed up nightly at four geographically different locations (in Vilnius, Granada, Caen and Portland/Oregon). In addition, regular backup archive copies of the whole repository are made on DVD-ROMs and stored offline. This level of backup storage is deemed enough to protect the investment of scientists' labour and supporting agencies grants. The open nature of the COD, however, permits all interested parties to store their own copies, should this be considered necessary. Unlike closed databases where data preservation depends solely on the owner of the database, open databases can be backed up flexibly, balancing backup costs against the value of data for the stakeholders.

The COD team strives to provide convenient and unambiguous access to data. To this end, the domain name 'www.crystallography.net' has been registered by the COD Advisory board. The data items will be indefinitely maintained as available over designated URIs. Thus, an URI containing a COD number in a form http://www.crystallography.net/<COD-number>.cif (e.g. http://www.crystallography.net/1000000.cif), is permanently mapped to the corresponding CIF, no matter what file layout or internal representation the COD is using. The URIs of COD repositories and rsync modules mentioned in the previous sections are also intended to be durable. So far we have maintained the described URIs since 2003. Thus, researchers can rely on the web services provided by the COD server, and on the possibility to obtain local copies or restore previous data in a standard way if needed.

DISCUSSION

Application areas of the COD database

With the current state of the COD database, several important applications are immediately possible. The coverage of compounds that are already included in the COD permits the usage of the COD for material identifications using powder diffractograms and theoretical computed powder patterns from the COD entries. Search-match software databases of such a kind are available for several diffractometers (http://www.crystallography.net/archives/2011/PANalytical/, http://www.crystallography.net/archives/2011/Rigaku/, http://www.bruker-axs.com/crystallography_open_database.html).

Even without 100% coverage of all known compounds (and elements), COD can be used for gathering statistics about chemical and structural properties of organic compounds, assuming that the COD has a representative subset of the compounds of interest. Since the COD may be accessed and downloaded free of charge by universities and students, the COD can also be used for teaching, providing a real life and high volume example of a scientific database for formulating challenging tasks in data mining and processing.

A lucrative use of the COD arises in areas where large amounts of computing power is necessary and grid or volunteer computing is viable. The open nature of the COD permits the downloading of this database or its parts to any number of computers, thus taking advantage of huge amounts of parallel resources available in grid/volunteer computing environments.

CONCLUSION

It is hoped that users not finding in the COD some data he or she knows to exist (especially those in the own private or institutional archives) remember how the COD grew and continues to grow. Hopefully, she or he becomes another volunteer, depositing new entries in the CIF format to the COD. Only that way the COD collection will come to close completion (estimated to be $\approx$800 000 entries currently) providing full services to those requiring exhaustive coverage. The definite advantages of the COD collection of small molecule / small to medium unit cell structures are first its direct availability from any personal computer connected to the Internet and second the full spectrum of structure types covered and available (organics, inorganics, organometallics and minerals) in one place, which more and more researchers in multi-disciplinary laboratories require. The current COD growing speed ($\approx$40 000 entries/year) is not sufficient, it should be $\approx$200 000 entries/year during the four next years in order to attain the $\approx$1 000 000 small structures that will be available at that time. This is because >50 000 crystal structures are determined per year nowadays. Given the efficient system described in this article, developed and installed at Vilnius, the task looks now less impossible with more users help, with your help.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Files.

FUNDING

This research is funded by a grant (No. MIP-124/2010) from the Research Council of Lithuania. Funding for open access charge: the Research Council of Lithuania grant (MIP-124/2010).

Conflict of interest statement. None declared.

REFERENCES

1. Hall,S.R., Allen,F.H. and Brown,I.D. (1991) The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr. Sec. A*, **47**, 655–685.
2. Allen,F.H., Bellard,S., Brice,M.D., Cartwright,B.A., Doubleday,A., Higgs,H., Hummelink,T., Hummelink-Peters,B.G., Kennard,O., Motherwell,W.D.S. *et al.* (1979) The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Crystallogr. Sec. B*, **35**, 2331–2339.
3. Bernstein,F.C., Koetzle,T.F., Williams,G.J., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M. (1977) The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.*, **112**, 535–542.
4. White,P.S., Rodgers,J.R. and Le Page,Y. (2002) Crystmet: a database of the structures and powder patterns of metals and intermetallics. *Acta Crystallogr. Sec. B*, **58(Pt 3 Pt 1)**, 343–348.
5. Kaduk,J.A. (2002) Use of the inorganic crystal structure database as a problem solving tool. *Acta Crystallogr. Sec. B*, **58(Pt 3 Pt 1)**, 370–379.
6. Rajan,H., Uchida,H., Bryan,D., Swaminathan,R., Downs,R. and Hall-Wallace,M. (2006) Building the american mineralogist crystal structure database: a recipe for construction of a small internet database. In: Sinha,A.K. (ed.), *Geoinformatics: Data to Knowledge*, Vol. 397. Geological Society of America, Boulder, CO, United States, pp. 73–80.
7. Kabekkodu,S., Faber,J. and Fawcett,T. (2002) New powder diffraction file (pdf-4) in relational database format: advantages and data-mining capabilities. *Acta Crystallogr. Sec. B*, **58**, 333–337.
8. Wang,Y., Bolton,E., Dracheva,S., Karapetyan,K., Shoemaker,B.A., Suzek,T.O., Wang,J., Xiao,J., Zhang,J. and Bryant,S.H. (2010) An overview of the pubchem bioassay resource. *Nucleic Acids Res.*, **38**, D255–D266.
9. Zucker,L.G., Darby,M.R., Furner,J., Liu,R.C. and Ma,H. (2006) Minerva unbound: knowledge stocks, knowledge flows and new knowledge production. *Working Paper 12669*. National Bureau of Economic Research, Cambridge, Massachusetts, United States.
10. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. Sec. D*, **58(6 Part 1)**, 899–907.
11. Berman,H.M., Olson,W.K., Beveridge,D.L., Westbrook,J., Gelbin,A., Demeny,T., Hsieh,S.-H., Srinivasan,A.R. and Schneider,B. (1992) The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.
12. Collins-Sussman,B., Fitzpatrick,B.W. and Pilato,C.M. (2008) *Version Control with Subversion*. O'Reilly Media, Sebastopol, CA, United States.
13. Grosse-Kunstleve,R.W., Sauter,N.K., Moriarty,N.W. and Adams,P.D. (2002) The *Computational Crystallography*

Toolbox: crystallographic algorithms in a reusable software framework. *J. Appl. Crystallogr.*, **35**, 126–136.
14. Gražulis,S., Chateigner,D., Downs,R.T., Yokochi,A.F.T., Quirós,M., Lutterotti,L., Manakova,E., Butkus,J., Moeck,P. and Le Bail,A. (2009) Crystallography Open Database – an open-access collection of crystal structures. *J. Appl. Crystallogr.*, **42**, 726–729.
15. Perl.org (2011) http://www.perl.org/ (20 October 2011, date last accessed).
16. Downs,R. and Hall-Wallace,M. (2003) The american mineralogist crystal structure database. *Am. Mineralogist*, **88**, 247–250.
17. Todorov,G. and Bernstein,H.J. (2008) *VCIF2*: extended CIF validation software. *J. Appl. Crystallogr.*, **41**, 808–810.