Crystallography Open Database: history, development, perspectives

Saulius Gražulis[a], Andrius Merkys[a], Antanas Vaitkus[a], Daniel Chateigner[b], Luca Lutterotti[c], Peter Moeck[d], Peter Murray-Rust[e], Miguel Quiros[f], Robert T. Downs[g], Werner Kaminsky[h], Armel Le Bail[i]

[a] *Vilnius University Institute of Biotechnology, Saulėtekio al. 7, LT-10257 Vilnius, Lithuania*

[b] *Normandie Université, CRISMAT-CNRS, ENSICAEN, IUT-Caen, Université de Caen Normandie*

[c] *Department of Industrial Engineering, University of Trento*

[d] *Department of Physics, Portland State University, P.O. Box 751, Portland, OR 97207-0751*

[e] *The University of Cambridge, Department of Chemistry*

[f] *Departamento de Química Inorgánica, Facultad de Ciencias, Universidad de Granada*

[g] *Department of Geosciences, University of Arizona*

[h] *Department of Chemistry, University of Washington at Seattle*

[i] *Université du Maine, Institut des Molécules et des Matériaux du Mans, CNRS UMR 6283, 72085 Le Mans*

Table of Contents

Abstract

Data are important assets of modern science, enabling reliable conclusions from present and historic measurements, and facilitating new discoveries, especially when cross-disciplinary data sets are compared. This article briefly discusses how data are managed and disseminated in the field of crystallography, and describes the construction, use and application of the Crystallography Open Database, a large resource publishing chemical crystallography data as an open-access database.

Introduction

Science is crucially based on observational data. As an example of an ancient data-driven discovery, the observation of equinox precession by Hipparchus in around 130 BCE comes to mind (Wikipedia 2016) – Hipparchus compared the longitudes of Spica and Regulus and other bright stars with the measurements from his predecessors, Timocharis and Aristillus, who lived about 100 years earlier, and concluded from the differences that the equinox points drift with time. Needless to say, this discovery could only be made because old observations of Timoarchis school were meticulously recorded, accurate enough and preserved for future generations. Today, the amount of data that scientists collect each year has grown by roughly 10 orders of magnitude, with fields such as astronomy or particle physics currently accumulating from several TB (Annis et al. 1999) to as much as 15 PB (petabytes) of data per year (Hewett 2006; PPARC 2006).

In the field of crystallography, the need of long term data preservation was recognised very early in the field. Currently, the International Union of Crystallography (IUCr) and the crystallographic community take great care with respect to data archiving and data reuse. The IUCr has rigorously described mathematical definitions necessary for crystal structure and experiment description in the International Tables for Crystallography (Hahn 2006), and created the CIF standard for crystallographic data exchange (Hall et al. 1991; Fitzgerald et al. 2006), which is constantly maintained to address new challenges in data management (Bernstein et al. 2016). Crystal diffraction data is being accumulated systematically in a number of databases since as early as 1941 (Faber and Fawcett 2002), archived in various crystallographic databases, the largest ones being the COD (Gražulis et al. 2012), the CSD (Groom et al. 2016), the ICSD (Belsky et al. 2002), the Pauling File (Villars et al. 2004), the PDB (Berman et al. 2012), the PDF from ICDD (Faber and Fawcett 2002), and the CRYTSMET (White et al. 2002). Several other, more specialised databases exist that specialise on specific aspects of crystallographic data; the structures they mention are usually included in one or several above mentioned databases. References to these specialised databases will be given below.

Before 2003, of the above mentioned crystal structure archives, only PDB offered full open-access to the crystallographic data it contained; all other databases followed a subscription-based model, offering little or no data on the Web for the general public or non-subscribers, and requiring purchase of a license for systematic data searches and, occasionally, restricting publication of derived data (Baldi 2011; Sadowski and Baldi 2013). The advent of the Web, ubiquitous computing and advantages of open, linked data on the web prompted a group of crystallographers to initiate the Crystallography Open Databases (COD), offering crystal structures for chemical crystallography on similar grounds as PDB provides them for macromolecular crystallography. As of current, the COD and the PDB remain the two largest databases offering the Open Access model to crystallographic data, and together covering the largest domain of crystal structures in an open way. While other databases contain larger collections of crystals structures, and claim higher level of data curation than COD (Bruno and Groom 2014), they still require acquisition of licenses for systematic data search.

In this chapter, we will review the COD contents, data collection and data curation policies. We will then describe various ways how COD data can be accessed and used. Finally, we will give examples of COD applications in fields of crystallography, chemistry, material identification and teaching.

Open databases for science

Over years, various researchers find that open access to articles consistently increases citations of these publications (Harnad and Brody 2004; Eysenbach 2006; Zucker et al. 2006; Harnad et al. 2008; Eger et al. 2013). Similar trends are observed for data in the field of bioinformatics (Piwowar and Vision 2013), and one would expect crystallography to follow similar trends. Thus there is a pure pragmatic reason for researchers to deposit data openly, so that they are findable, reusable and citable. For the users of data, the absence of pay-walls, access and use restrictions provides the convenience of one-click access to data. Finally, there are ethical considerations – most published research was funded by public money, and the society members who's taxes were used to produce scientific results has reasonable expectations that these results will be available to them without demand of extra payment and without restrictions. Understandably, then, many funders require that researchers whom they have supported publish their results under open access licences for both publications and data.

To answer the above-mentioned concerns, many open databases have been established by researchers. Below, we describe topic-specific databases, in addition to more general databases outlined above.

Table 1. Material property and structure databases available on-line.

Nr.	Database	Approx. no. of records	License	Current URL	Est.	Reference
1.	MPOD	300	Public domain	http://mpod.cimav.edu.mx	2010	(Pepponi et al. 2012)
2.	RRUFF	47 000	Open access	http://rruff.info/	2015	(Lafuente et al. 2015)
3.	AMCSD	20 000	Open access	http://rruff.geo.arizona.edu/AMS/amcsd.php	2003	(Downs and Hall-Wallace 2003; Rajan et al. 2006)
4.	IZA Zeolite database	176[1]	Open access	http://www.iza-structure.org/databases/	1996	(Baerlocher et al. 2007)

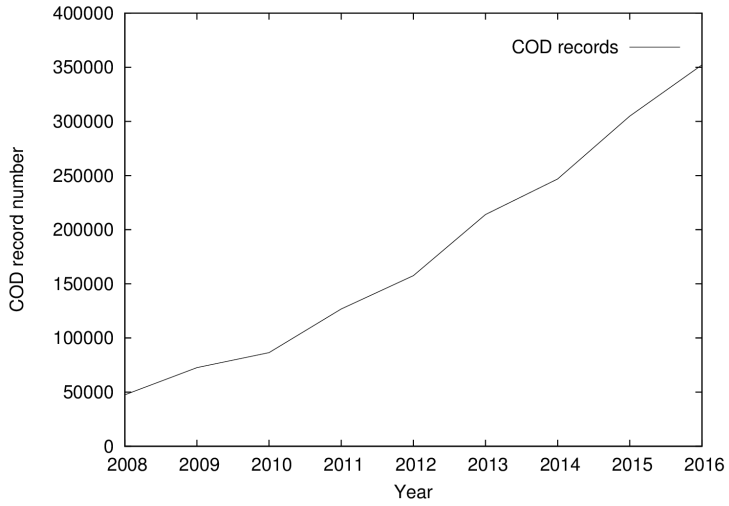
1 The number of unique zeolite Framework Types that had been approved and assigned a 3-letter code by the Structure Commission of the IZA

Nr.	Database	Approx. no. of records	License	Current URL	Est.	Reference
5.	Bilbao server			http://www.cryst.ehu.es	1997	(Aroyo et al. 2011)
6.	Bilbao Incommensurate Structures Database	140	Open access	http://webbdcrista1.ehu.es/incstrdb/	2010	(Aroyo et al. 2006)
7.	Bilbao Magnetic structure database	428	Open access	http://webbdcrista1.ehu.es/magndata/	2015	(Perez-Mato et al. 2015)
8.	NDB	8600	Open access	http://ndbserver.rutgers.edu/	1992	(Berman et al. 1992; Coimbatore Narayanan et al. 2014)
9.	COD	367 000	Public domain	http://www.crystallography.net/cod	2003	(Gražulis et al. 2009; Gražulis et al. 2012)
10.	PCOD	1 000 000	Public domain	http://www.crystallography.net/pcod	2003	(Le Bail 2005)
11.	TCOD	2 000	Public domain	http://www.crystallography.net/tcod	2013	(Chateigner et al. 2015)
12.	CSD	800 000	Subscription based	http://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/	1965	(Groom et al. 2016)
13.	ICSD	200 000	Subscription based	https://icsd.fiz-karlsruhe.de/	1987	(Belsky et al. 2002)
14.	PDF	380 000	Subscription based	http://www.icdd.com/products/pdf4.htm	1941	(Faber and Fawcett 2002)
15.	CRYSTMET	170 000	Subscription based	http://www.TothCanada.com [2], https://cds.dl.ac.uk/cgi-bin/news/disp?crystmet	1996	(White et al. 2002)
16.	Pauling file	290 000	Subscription based	http://paulingfile.com	1995	(Villars et al. 1998; Villars et al. 2004)
17.	PDB	124 000	Open access	http://www.rcsb.org/pdb	1971	(Bank 1971; Berman et al. 2012)
18.	BMCD	43 000	Open access	http://xpdb.nist.gov:8060/BMCD4	1995	(Gilliland et al. 2002)

2 The page at the http://www.TothCanada.com advertised in (White et al. 2002) seems no longer operational, bu the access for subscribers is advertised at https://cds.dl.ac.uk/cgi-bin/news/disp?crystmet.

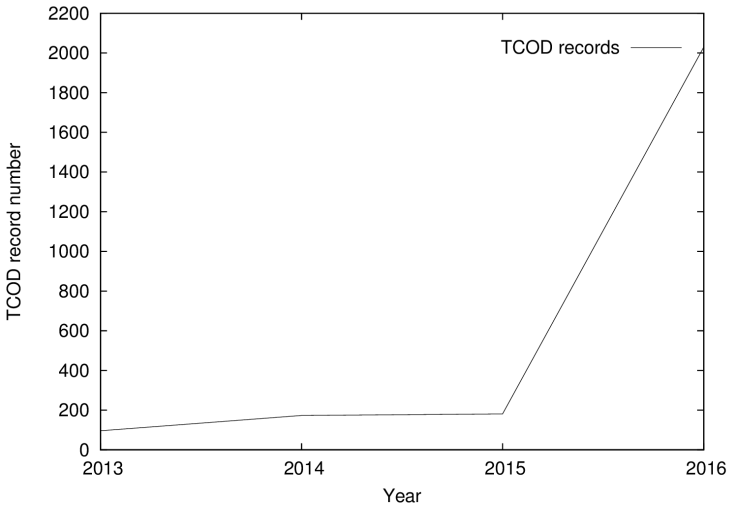
A constantly updated list of scientific databases in the field of biosciences can be found in the Nucleic Acids Research (Galperin and Cochrane 2010), and crystallographic databases are listed by the IUCr (http://www.iucr.org/resources/data).

Figure 1. COD record number growth.



The Crystallography Open Database incorporates a continuously increasing number of determined crystal structures, reaching > 367000 entries at the time of writing this article (fig. 1). The equivalent of COD for structures obtained from first-principle calculation and/or optimisation is TCOD, started in 2013, with consequently a more modest number of entries around 2000 (fig. 2). However such entries require long calculation times and one can expect larger increases in the years to come.

Figure 2. TCOD record number growth.



The Crystallography Open Database (COD) was founded in February 2003. The COD is a grass-root initiative – its establishment was proposed in a letter published at the SDPD (Structure Determination by Powder Diffractometry) mailing list by Michael Berndt:

« What if crystallographers work together to establish a public domain database with all relevant crystallographic data? This would not only overcome the current situation with 'fragmented' databases, it would also prevent for becoming dependent from monopolists. What would be needed ?

1. A small team of engaged scientists with some experience in database and software design to coordinate the project.
2. The authors (i.e. the scientific community = you) who provide the project with database entries (note, that if you haven't sold your experimental results exclusively, you are free to distribute the data to such a database, even if they have already been part of a publication - and a lot of good data have never been published).
3. Free software a) for maintaining the database, b) for data evaluation and calculation of derived data (e.g. calculated powder pattern from crystal structures for search-match purposes), c) for browsing and retrieval. We are not in the same situation as decades before when the well-known databases (ICSD, CSD, PDF) started. Today we have the Internet, fast computers, and a big pool of free available software. The question is: Do we have enough scientists who are willing to cooperate? »

Several laboratories contributed a lot to the COD at its very beginning. Bob Downs offered his collection of mineralogical data, including the whole AMCSD (Rajan et al. 2006) dataset (all the crystal structures previously published in the American Mineralogist which were made freely accessible from the websites of the Mineralogical Society of America). The MySQL/PHP scripts written by Hareesh Rajan. Meantime, Daniel Chateigner joined, and less than 3 weeks after the letter from Michael Berndt, the COD project was announced at various Internet media (Newsgroups, various Mailing Lists and What's New pages) by the following letter :

> « Dear Crystallographers, a project of Crystallography Open Database (COD), accommodating crystal structure atomic coordinates prior to their publication, is under development. It is intended to give faster access to the latest structure determinations, openly. Its development and success depends completely on your contributions, either by data download or/and by giving help in software improvements. Visit the COD project Web pages (www.crystallography.net) for more details and a crystallography database(s) quiz. Thanks for your future help, the COD is yours, it is the right time to do something for an open database controlled by crystallographers, now or never!
>
> The advisory board (wishing to enlarge) : Michael Berndt, Daniel Chateigner, Robert T. Downs,
>
> Lachlan M.D. Cranswick, Armel Le Bail, Luca Lutterotti, Hareesh Rajan »

This letter produced a lot of positive as well as negative comments. Some researchers who responded positively joined the COD team and the number of entries in the COD increased, attaining more than 5000 entries by the end of March 2003 (3725 CIFs from the AMCSD, 450 CIFs from the LdOF, 850 CIFs from the CRISMAT). The CIF2COD computer program (FORTRAN) was built up on the basis of CIF2SX with permission from Louis Farrugia. CIF2COD reads several CIFs (from n.cif to n+m.cif), performs several quality tests, and produces a .txt file containing m+1 lines with the MySQL database (cod) unique table (data) fields (including a, b, c, alpha, beta, gamma, volume, number of elements, space group, chemical formula, reference and additional text). The first minimal COD search page was coded in PHP language. Donations continued in April 2003 (1200 CIFs from IPMC) and the IUCr was contacted, asking for permission to download systematically the CIFs freely available at the IUCr Web site. The decision had to wait for the next IUCr Executive Committee meeting in August 2003. After 4 months, the number of entries in the COD attained 12000, essentially by donations, from individuals or laboratories and the AMCSD.

Then came bad news. Sadly, Michael Berndt died on June 30th, 2003 after a long, serious illness at the age of 39. Despite the loss, the COD team continued to implement his plan and to work on the database. Five years after its foundation, the COD passed a major milestone in 2008, by archiving the

50000-th entry. To attain completion some day, the COD should add much more than 40000 new entries per year, and also to digitise older data that were published in paper form. The required growth rate of the COD was attained in 2011 (fig. 1), when automated procedures for crystallographic data collection were implemented. Nevertheless, we still have much to do, and the COD wishes to convince more crystallographers in order to accelerate its completion. During these past ten years, the COD Advisory Board undergoes some variations, departures, new admissions, so that the list of co-authors of the present chapter, intended to stimulate more cooperation, is representative of the current situation, listing the main actors of the COD development until now.

Building COD

COD collects all published small molecule crystal structures. To facilitate this process, a Crystallographic Interchange File/Framework (CIF) is employed. Currently, COD uses the CIF 1.1 (Hall et al. 1991) version of the framework. The framework files, CIFs, are used to input data into the COD, as an intermediate versioned archive for storage, and for providing data to the users.

The main founding principle of the COD database is Open access – all data are readily available on the Internet. COD data records are identified by stable URIs and accessible via the REST interface. The COD main page on the Web (http://www.crystallography.net/) states that "All data on this site have been placed in the public domain by the contributors", which we assume binding for COD AB, data maintainers and contributors. All deposited data, unless embargoed by depositors for a fixed amount of time as a "pre-publication deposition", is immediately after the deposition available on the Internet and accessible via the automatically generated stable identifiers. Such arrangement enables immediate and permanent linking of COD structures into the World Wide Web fabric.

Each data item that is committed to the COD repository is first of all checked for the syntactic correctness of the incoming CIF file. Since not all submitted files can be guaranteed to conform to the formal CIF definition (IUCr 2016), an error-correcting CIF parser (Merkys et al. 2016) is employed. This ensures that all COD CIFs can be automatically parsed and supports unassisted COD data processing.

Scope and contents

COD aims at collecting all experimentally determined small-molecule crystal structures into an open-access resource. “Small-molecule” category encompasses all inorganic, metal-organic and organic compounds with an exception of macromolecules – organic polymers. The latter are being collected into dedicated well-known open-access databases such as Protein Data Bank (PDB) (Berman et al. 2000) and Nucleic Acid Database (NDB) (Berman et al. 1992; Coimbatore Narayanan et al. 2014).

As an experimental database, COD collects structures determined by any experimental method. However, there are sister databases PCOD and TCOD, which aim to collect predicted and theoretically determined structures, respectively (see Sister databases (PCOD, TCOD) for more thorough description).

COD structures may be refined using just X-ray data and the first physical principles (using full-matrix least squares), but they may also be refined using restrains (especially when determined using powder diffraction methods), or, more recently, hybrid methods (from experimental powder data using Rietveld and Le Bail methods + from first principles using DFT).

Data sources

The COD finds most of its structures (over 90%) from peer-reviewed scientific publications. The rest is deposited by authors either as Personal communications or as pre-publication depositions. Data published in papers are subjected to checks for conformance with CIF syntax, CIF dictionary definitions and the completeness of bibliographic and other provenance information. Personal communications and pre-publication depositions are in addition checked for conformance to the IUCr data criteria[3]. The COD permits both manual deposition by crystallographers using a Web interface (http://www.crystallography.net/cod/deposit), or an automated deposition using various Web inspecting engines. Automated web searches are conducted on the journals that publish openly accessible crystallographic supplementary data. Data are also automatically extracted from open-access publications. Data from other crawlers, such as CrystaEye (Day et al. 2012), and other open databases (e.g. AMCSD (Rajan et al. 2006)) are incorporated into the COD on a regular basis, either using automated or semi-manual procedures. Such strategy permits broad coverage of published structures with little resources required; it leverages power of Internet automation, at the same time permitting humans to intervene at critical points when necessary.

It must be noted with regret that some journals still do not provide the supporting data for their papers openly. Data are either located behind the pay-walls, or available only in subscription based databases with explicit restrictions on their reuse. Unfortunately, this makes a technically simple task of collecting all currently published crystal structures in an open databases virtually impossible, not for technical but for purely organisational reasons. The barriers are not even related to intellectual property, since published data and facts of Nature are uncopyrightable. We thus urge everyone who sees virtue in open scientific data exchange and has benefited from open-access database to approach every publisher and to ask them provide underlying publication data for deposition to open-access databases, or to deposit her or his crystallographic data directly to the COD.

Data maintenance

Scientific databases are an indispensable resource in the modern day research and as such they must adhere to the criteria of all properly designed experiments – reproducibility and traceability. Obtained results are of little value if repetition of the same procedures under the same conditions yields a different outcome. The same holds true if the experiments are purely computational in origin such as simulations or compilation of statistical data. In addition to that, any conclusions drawn from claims of untraceable origin become unverifiable and run the risk of polluting every sequential experiment they are used in. As a result, the employment of the WORM (Write Once Read Many) principle which ensures that once data is written it is never completely forgotten becomes a necessity for scientific databases.

Collecting and preserving scientific data is an important endeavour, but maintaining it is a task of no less importance. Reasons behind the need to modify the data are numerous – from a simple human error to new insights about the data or even the introduction of a novel way of describing certain phenomenon. The means for updating scientific articles via the issuing of addenda and errata are well established, however, the same mechanism is usually not applied to the supplementary material. A more common approach is to silently replace the outdated version with a new one leaving the returning reader with a very unexpected sense of *jamais vu*. The situation is only worsened by the fact that supplementary material is rarely well-reviewed before publishing, resulting in an even greater need for a proper data maintenance strategy.

3ftp://ftp.iucr.org/pub/dvntests and http://journals.iucr.org/services/cif/checking/autolist.html

Data discrepancies addressed by the COD maintainers can be grouped into three main classes: syntax errors, semantic errors and errors relating to the crystal structure. Each of these classes requires different detection and correction strategies as well as affects the data usability in varying degrees (see table 2).

Error class	Ease of detection	Ease of correction	Effect on data usability
Syntax	Detected automatically by the parser	Mostly automatic	Unreadable file
Semantic	Detected mostly automatically by specialized software, requires occasional manual analysis	Automatic and manual	Incorrect supporting information
Crystal Structure	Detected by specialized software and manual analysis	Mostly manual	Incorrect crystal structure

Table 2: Error classes routinely addressed by the COD maintainers

The initial step of data management in the COD is the detection and correction of syntactic errors. This kind of discrepancies are especially important since they render the files unreadable and limit the possibility of any further data maintenance. Crystallographic structures in the COD are stored as CIF files, a format that has been adopted by the crystallographic community. However, even with the widespread use of the CIF format, none of the parsers available at the time were capable enough to satisfy the specific needs arising from the curation of large data sets. As a result, maintainers of the COD have developed an open-source error-fixing CIF parser, which is able to correct some of the most prominent syntax errors (Merkys et al. 2016). Initial file parsing upon deposition as well as the routinely database-wide checks guarantee that at any given moment all files in the COD can be read correctly according to the CIF format rules.

Syntactical correctness ensures that the files are readable, but does not guarantee the validity of the data stored inside the files – this is the task for semantic validation. Due to a great variety of semantic errors and the fact that they usually only affect a portion of the data in the file, the COD has adapted a very flexible policy regarding discrepancies of this kind. Upon the initial deposition semantic errors are recognised, automatically corrected, reported to the depositor and in case an automatic correction is not possible, recorded in an internal database for further analysis. Once a significant amount of similar errors accumulate, heuristics-based programs are developed to automatically fix the errors in question. Since it is unreasonable to expect perfect detection of all possible semantic error cases in advance, the file validation strategy also addresses the handling of new kinds of semantic discrepancies that were previously missed during the initial deposition. In this case, heuristics-based programs are developed for the detection of these new errors and the whole database is revalidated based on the new criterion. In the end, both the new error-correction programs and the new error-detection programs are eventually integrated into the deposition step. The described work flow ensures that the overall semantic validity of the COD dataset will only increase in the future.

The set of computer programs developed by the COD maintainers for the detection and correction of syntactic errors are collectively called cod-tools. These tools are capable of recognising all of the problems listed in the IUCr validation criteria (http://journals.iucr.org/services/cif/checking/autolist.html), such as misspellings of data item names or their enumerated values, as well as some other common issues identified by scanning the COD. Examples of such discrepancies include data items designated to specify temperature containing values in other units than Kelvins or data items used to describe density of the crystal containing values in

kg/m^3 instead of g/cm^3. Instances of errors like these might not seem significant when handling individual files, but they do complicate the work flow and skew the results of database-wide analyses. Luckily enough, some of the errors can be automatically corrected by using heuristics (for example, unit designators after the temperature values); others, however, require manual curation.

One type of manually curated errors is the incorrect number of implicit hydrogen atoms. This number, provided using the '_atom_site_attached_hydrogens' data item, specifies the amount of hydrogen atoms attached to the atom site excluding the hydrogen atoms for which coordinates are given explicitly. Such discrepancies are easily spotted even by a novice chemist, but they are much harder to detect automatically. Incorrectly marked hydrogen atoms result in erroneous calculated atom charges, mismatch between the declared and the calculated formulae and skewed distributions of geometric parameters.

Errors in the coordinates, cell constants and symmetry are especially difficult to locate and correct. Nevertheless, the structures in the COD are routinely scanned for “bumps” (suspiciously small interatomic distances) and voids. Examination of “bumps” usually reveal modelling errors, unmarked disordered sites or redundant atoms; several non-P1 structures, that had all symmetric atoms listed, have been spotted and corrected while scanning the COD. Voids, on the other hand, are a sign of missing atoms or their groups, wrong cell constants or incorrectly low symmetry. Currently, new means of detecting other geometric anomalies in deposited structures based on statistical distributions of geometric parameters are being developed. Such checks will make the identification of unfinished refinement, missing atoms, typographical errors in coordinates and cell constants possible.

Not all structures, however, can be successfully corrected. To inform the user as well as enable the recognition of such entries in automated analyses, a warning or an error flag is added to the CIF file manually. Currently there are around twenty such entries in the COD.

Another type of structures in the COD unfit for normal use are the retracted ones. Retraction rate, as reported by RetractionWatch, is around 500-600 retractions per year (http://retractionwatch.com/help-us-heres-some-of-what-were-working-on/) and the field of crystallography is not immune to incorrect conclusions and scientific fraud. Since, at least to the knowledge of the COD maintainers, there is no open database listing all retracted publications, the process of retraction in the COD is completely manual. Each entry coming from retracted publications is blanked and excluded from the search so as not to bias automated analyses. However, since history of all structures is preserved in the COD, retracted structures can be accessed if necessary.

Alongside retractions, there are a few more types of entries, that are not desired in the COD, but often are identified as such only after the deposition. One of them is duplicates: in order to not overcrowd the COD with repeated entries and thus bias statistical results, deposited structures are compared to the rest of structures in the database during an attempt to locate duplicates. Currently, two structures are assumed to be duplicates if they originate from the same publication, have the same lattice cell constants and contents, are measured at the same temperature and pressure and are not enantiomers or suboptimal of one another. However, this method is not perfect. Moreover, a handful of duplicates has slipped in during the era before the automatic deposition system. Therefore, new methods to locate duplicates are devised and employed in the COD, almost always requiring supervision of a data curator. As entries are not removed from the COD, duplicates are marked with a special flag, indicating the original entry.

In 2013 results of theoretical calculations being deposited to the COD were spotted. This caused the policy of accepting only experimentally detected structures to be reiterated and a sister database, the TCOD, was opened to house all kinds of theoretically defined structures. Since then more than 400 theoretical structures were identified and marked as such in the COD. Difficulty to identify theoretical

structures from data given in CIFs hinders automatic detection of such depositions. However, properties like high numeric precisions of cell constants and coordinates, missing standard uncertainties and experimental details may be used to guide this otherwise manual task. As with any other structure not fitting the scope or criteria of the COD, theoretical structures are also marked as such instead of being removed.

Version control

Scientific data, when used, must be properly cited and available for verification of the conclusion drawn from them. The availability must be ensured both during the research, for the benefit of the scientist conducting it, but also at later stages, for peer review and for replications of conclusions reached. Curated databases, however, change over time, and databases like COD that follow immediate release policy can change at any time and at high rate, comparable with the rate at which data are queried for computations. To make sure that computations done with COD are repeatable, and inference drawn from them are reproducible, it is crucial that any previous state of the database can be restored. We implement this requirement by using version control on COD data.

Currently, a Subversion (Collins-Sussman et al. 2004; Collins-Sussman et al. 2011) server is used to register version of COD data in CIFs. Subversion is a powerful, off-the-shelf open source software system that enables track of changes in a tree of files, assigns each state of the file tree an unchangeable sequential revision number, and allows restoring any previous revision from the repository. Although originally designed as a tool for software development, Subversion offers precisely those functions that are needed for scientific database of the medium size, such as COD. In particular, a text nature of the CIF format makes them particularly well suited for tracking with revision control systems.

Since introduction of Subversion, all COD data curation history is available, and any state of the database can be restored. Since Subversion records also movements of files in the file tree and rename operations, a full data provenance of each COD CIF is provided in the version control system since its insertion into the repository. When a COD ID of a structure **and** a revision number of the structure is known, a unique string of bits (a digital object) describing that structure at a given revision can be retrieved.

The COD MySQL data tables are automatically produced from each current COD revision. These tables themselves are not currently versioned, i.e. currently MySQL tables contain only data from the most recent revision of the COD (although a nightly dump of the COD MySQL database is inserted into the COD Subversion repository). Such implementation was deemed satisfactory, since the primary COD data are CIFs, and MySQL tables for any revision can in principle be reconstructed from the CIFs of that particular revision.

As the database grows, however, and more queries are executed on the MySQL database, and not on the CIF tree, the need arises to quickly perform historic SQL queries, without reconstructing MySQL tables for each revision. This need is explicitly recommended in the RDA Recommendations for Data Citation (Rauber et al. 2015; Rauber et al. 2016). In future, therefore, COD will implement a possibility to query every revision of COD database on-line (historic states of MySQL tables will be restored from COD CIFs and marked with corresponding time-stamps and revision numbers), and to cite COD queries in a durable and reproducible way, enabling to rerun each historic query, both on the original data and on newer database revisions.

Data curation policies

Since COD record contents can change during data curation, a question arises what rules does the COD curation policy follow and what a researcher can rely upon. The current COD data policy is at follows. A COD record is essentially a *claim* made by a *data depositor* that the specified *authors* have published

certain findings about the structure described in the COD entry. To this extent, the COD data curation team makes reasonable efforts to make each COD entry to represent the publication authors' *intent*. To that end, data in COD entries can be enhanced during the data curation; addition data from the original publication may be added. Data values in CIFs may be corrected if a correct value is clearly specified by the authors in the original publication, and it is clear that the authors meant that value to be published (usually, such corrections also make good physical sense, making it obvious that the curated structure describes better the physical reality). In cases where the intent of the author is not so clear, or where essential data items such as coordinates of atoms or atomic symbols have to be changed, authors are first contacted to approve the changes. In all cases it must be clear that the original finding of the authors meant exactly the curated value, and is not a new interpretation of the experiment.

Data curation **never** involves a new structure solution from the same data, re-refinement, guessing values from common chemical knowledge or similar investigative steps. Such processes are possible, but in that case, a new COD ID must be assigned to the new structure solution, and is will be treated by the COD as a new publication.

The data curation process has data uniformity and accuracy of claims as its main aim. All COD structures must use the same conventions to describe analogous situations. In most cases, the IUCr CIF standard provides adequate means for uniform description, and we curate the data records to adhere to these standards. For example, atomic coordinates must be provided either as fractions of cell vectors along the crystal axes, or as Cartesian coordinates in an orthogonal frame (in which case orthogonalisation matrices relating the used Cartesian frame and the crystal axes must be given). Another instance is crystalline material melting point which must be given in Kelvins. If an original publication contains these data items recorded in different ways (different coordinate systems, different units), COD data curators convert them to the common mandated format, leaving original values in specific COD data items for reference. Sometimes, however, there is no standard way to express certain circumstances; for example, sometimes authors are not sure what is the chemical nature of atom occupying certain site in a crystal unit cell, and they mark such sites using different codes (such as "I1", "M2" or so on). COD introduces a uniform notation, "X" for completely unknown atom at a site, and "M" for an unknown metal. In that case the original authors' designators might be changed; the curated version (atom site "X"), however, expresses the authors' message "unknown atom" better than the original "I" designator, since the latter can be confused with iodine in COD context.

Quarterly releases

The COD follows a continuous release policy – each commit to a COD database is immediately available on the Web and in the Subversion repository. Each such commit introduces a new COD revision. COD contents is mostly updated day, and several revisions can be generated each day. It is therefore important that COD users keep track of which revision they are using for their calculations and data searches. Since such tracking might introduce extra burden, we are providing, after a popular request, quarterly releases of COD data snapshots. Four times a year a current COD revision is exported, both CIFs and MySQL table dumps, and packed in several most popular data formats. The revision and time-stamp of the most recent release is available at http://www.crystallography.net/cod/archives/LAST_RELEASE.txt. Each current release is available for download in the COD archive area:

- current release:
  - http://www.crystallography.net/cod/archives/cod-cifs-mysql.tgz
  - http://www.crystallography.net/cod/archives/cod-cifs-mysql.txz
  - http://www.crystallography.net/cod/archives/cod-cifs-mysql.zip

(the contents of all three files is identical, so only one is needed to obtain a release)

- historic releases: can be found in each year's 'data' directory, following the URIs of the type http://www.crystallography.net/cod/archives/<year>/data/; for example, all four releases of 2015 are in http://www.crystallography.net/cod/archives/2015/data/.

While the use of COD releases is conceptually simple and does not require the use of version control software and revision tracking, it must be noted that the releases get outdated quickly. Also, downloading a new release repeatedly downloads all previous data anew, wasting bandwidth and time. Thus, frequent COD user's should consider incremental means of updating their COD collection, such as Subversion ('svn') or Rsync.

Sister databases (PCOD, TCOD)

The growing need for COD-like databases for other than experimental structures has sparked the creation of two sister databases: PCOD for predicted structures and TCOD for theoretically constructed structures. PCOD, Predicted Crystallography Open Database (http://www.crystallography.net/pcod/), was launched in December 2003 with the goal of collecting computationally predicted structures. It was expected that the number of such entries could easily exceed the number of experimentally determined ones. In January 2004, the PCOD offered 200 entries. In February 2007, the number of entries were boosted to more than 60 000 by the deposition of crystal structure predictions using GRINSP software (Le Bail 2005). As COD passed a major milestone by archiving the 50 000th entry in 2008, PCOD climbed over the 100 000 structure limit in the same year. A year later PCOD reached one million entries, most of them being generated by ZEFSA II (Falcioni and Deem 1999). As a fork of COD, PCOD has inherited most of features, such as stable unique data identifiers, data versioning and Web and MySQL interfaces for searching. Automatic deposition service is to be implemented in PCOD.

TCOD, Theoretical Crystallography Open Database (http://www.crystallography.net/tcod/), was launched in May 2013 thus addressing the need for an open repository of theoretically computed chemical structures. As methods of computational chemistry enjoy unprecedented growth and computer power increases, a large number of atomistic simulations can be carried out, producing theoretical material structures and calculating their properties using DFT, post-HF, QM/MM and other methods. By the end of that year TCOD offered around 200 entries. To ensure high quality of deposited data, development of ontologies in a format of CIF dictionaries was initiated. In addition to that, COD-like pipeline to check each deposited structure against a set of community-specified criteria for convergence, computation quality and reproducibility was developed and installed in the TCOD. As of the time of writing, TCOD contains more than 2 000 entries.

Use of COD

Data search and retrieval

Open-access Web resources pave the way for unprecedented applications that interconnect and reuse data hosted by many different organisations without the need of coordination between them. Key elements for such cooperation are the interfaces for data accession. Commonly used architectural style for both human- and machine-usable Web interfaces is REpresentational State Transfer (REST), according to which RESTful interfaces are built (Fielding 2000), that use common HTTP requests to stable URLs for data retrieval.

Data identification (unique identifiers)

Each entry in the COD consists of a CIF data block, listing the atomic positions of the crystal of interest, and an optional diffraction data data block (Fobs, powder diffractograms). If an experiment results in more than one CIF data block (N data blocks), they are split across N COD entries.

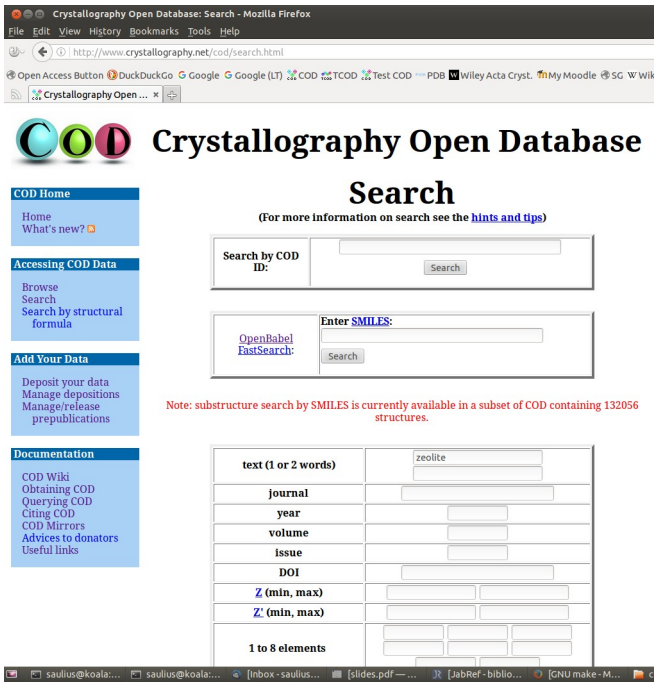
To provide permanent descriptors, unique identifiers – integers from range 1000000–9999999 – are assigned for each deposited entry upon the deposition into the COD. The COD identifiers are promised to be permanent – both retracted and duplicate entries, that are detected after their deposition, are marked as such instead of removal.

COD identifiers are straightforwardly transformed into stable URIs by prefixing them with http://www.crystallography.net/cod/ and postfixing with file type (.html for general review of an entry, .cif for CIF file with atomic positions and .hkl for diffraction data file). For example, files of entry 1504719 can be accessed via http://www.crystallography.net/cod/1504719.html and http://www.crystallography.net/cod/1504719.cif (diffraction data is not present for this entry).

Web search interface

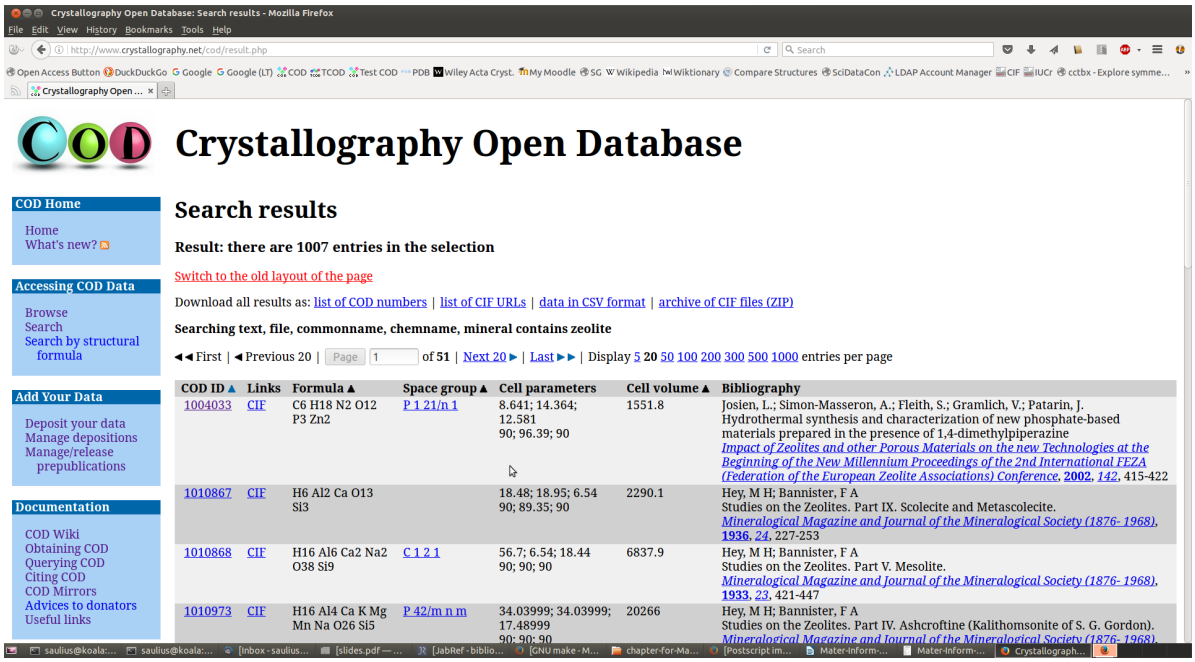
Data can be search on the Web using simple web forms that use the COD MySQL database as a fas-search index (Fig. 3):

Figure 3. COD search Web interface form.



The COD server returns found results as a paged HTML table (Fig. 4). From this page, results can be also downloaded in bulk as an archive. COD currently supports ZIP archives for downloaded data. The result table can be downloaded as a CSV (comma-separated value) file, and the list of selected structures can be obtained as a text file, either one COD number or one COD URI per line.

Figure 4. COD search result page, obtained as of 2016-11-05 from the query shown in Fig. 3.



RESTful interfaces

The same search interface can also accessed programmatically using the COD RESTful API. The base URL for search is http://www.crystallography.net/cod/result, while search terms have to be defined as HTTP GET or POST parameters. An example of such query using the 'curl' command line tools is given in fig. 5.

Figure 5. Example of the COD programmatic search interface.

```
sh% curl -sSL 'http://www.crystallography.net/cod/result?
text=ibuprofen&year=2014&format=urls'

http://www.crystallography.net/cod/4510385.cif
http://www.crystallography.net/cod/4510386.cif
http://www.crystallography.net/cod/4510387.cif
http://www.crystallography.net/cod/4510388.cif
```

A list of supported search terms is given in a list below:

- text: textual search; for example, text=caffeine;
- id: search by COD identifier; for example, id=3000000;
- el1, el2, …, el8: search for elements in composition; for example, el1=Ba&el2=O4;
- nel1, nel2, …, nel8: exclude entries with given elements; for example, nel1=Os;
- vmin, vmax: minimum and maximum volume of the cell, in Å^3; for example, vmin=10&vmax=20;
- minZ, maxZ: minimum and maximum Z value;
- minZprime, maxZprime: minimum and maximum value of Z';
- spacegroup: search by spacegroup;
- journal, year, volume, issue, doi: search by terms in bibliography.

By default, the result of the structure request is returned in CIF format. Other output formats can be requested.

Output formats

Combination of search parameters results in logical conjunction (OR operation). Output format can also be controlled using HTTP GET or POST parameter 'format', with one of the following values: 'html', 'csv', 'zip', 'json', 'count'. In addition to them, 'lst' value can be used to get the list of COD identifiers, 'urls' to get the list of COD URLs and 'count' to get the number of entries matching the search query. The default format currently used for the 'result' query is 'html', returning a paginated HTML table. Since the request of the search result with no search terms selects all COD database, this URI can be also used for browsing the COD database by COD ID. Other browsing pages (currently by journal or by publication date; the full list is available at http://www.crystallography.net/cod/browse.html) are actually also implemented using the 'result' requests.

Accessing COD records

As presented in Data identification (unique identifiers) (see p. 14), each entry in the COD is identified by unique seven-digit number. COD presents the following URLs for accession of entry-related data:

- Coordinates: http://www.crystallography.net/cod/XXXXXXX.cif
- Diffraction data: http://www.crystallography.net/cod/XXXXXXX.hkl
- Metadata in RDF: http://www.crystallography.net/cod/XXXXXXX.rdf

Here, the XXXXXXX placeholder should be replaced by a single COD number. An example of a query made using these identifiers from the Unix-style command line is show in fig. 6.

Figure 6. Retrieving a specific COD structure using the stable COD URI identifier.

```
sh% curl -sSL http://www.crystallography.net/cod/2001546.cif | head -n 30
#------------------------------------------------------------------------------
#$Date: 2016-02-19 16:29:56 +0200 (Fri, 19 Feb 2016) $
#$Revision: 176759 $
#$URL: svn://www.crystallography.net/cod/cif/2/00/15/2001546.cif $
#------------------------------------------------------------------------------
#
# This file is available in the Crystallography Open Database (COD),
# http://www.crystallography.net/. The original data for this entry
# were provided by IUCr Journals, http://journals.iucr.org/.
#
# The file may be used within the scientific community so long as
# proper attribution is given to the journal article from which the
# data were obtained.
#
data_2001546
loop_
_publ_author_name
'Freer, A. A.'
'Bunyan, J. M.'
'Shankland, N.'
'Sheen, D. B.'
_publ_section_title
;
 Structure of (<i>S</i>)-(+)-ibuprofen
;
_journal_issue                   7
_journal_name_full               'Acta Crystallographica Section C'
_journal_page_first              1378
_journal_page_last               1380
_journal_paper_doi               10.1107/S0108270193000629
```

Depositions to the database in the form of CIFs are also available using the RESTful interface. Currently, registration of a depositor account at the COD is required beforehand. The URL of the RESTful deposition interface is http://www.crystallography.net/cod/cgi-bin/cif-deposit.pl. All parameters along with a CIF file must be provided via HTTP POST:

- username: depositor's username;
- password: depositor's password;
- user_email: depositor's e-mail address;
- cif: contents of to-be-deposited CIF file;
- hkl: contents of to-be-deposited diffraction data file (optional);
- deposition_type: type of deposition, either “published”, “prepublication” or “personal”.

MySQL interface

The Web based interfaces are readily available, can be accessed using standard software such as Web browser or URL downloader, and do not require any sophisticated programming, their capabilities are naturally limited since we can not expose a full data query language such as SQL at the moment. To alleviate this limitation, COD exposes a read-only version of the COD MySQL database for queries. When accessed as a user 'cod_reader', this database grants SELECT privilege to that user to enable full use of the SQL query language. A special host dedicated for such public queries us

'sql.crystallography.net'. An example of such query using the Linux 'mysql' command line client is illustrated in fig. 7.

Figure 7. Querying the COD MySQL database.

```
sh% mysql -u cod_reader -h sql.crystallography.net cod -e \
    'select file as codid, formula, year, a, b, c, vol from data \
         where vol between 100.0 and 100.1 and formula != "?" \
         order by year'
+---------+--------------------------+------+--------+--------+---------+---------+
| codid   | formula                  | year | a      | b      | c       | vol     |
+---------+--------------------------+------+--------+--------+---------+---------+
| 1011228 | - Mn O3 -                | 1920 |   5.84 |   5.84 |    5.84 |     100 |
| 1528581 | - F6 Li2 Zr -            | 1960 |   4.98 |   4.98 |    4.66 | 100.086 |
| 9009168 | - Cl Ho O -              | 1963 |  3.893 |  3.893 |   6.602 | 100.056 |
| 1538762 | - Pu Zr -                | 1964 |  4.642 |  4.642 |   4.642 | 100.027 |
| 1537490 | - Cl N Ti -              | 1964 |  3.937 |  3.258 |   7.803 | 100.087 |
| 1532423 | - Li0.29 Si0.88 Zr1.83 - | 2002 |  3.701 |  3.669 |   7.581 | 100.013 |
| 1510152 | - Au Ga O2 -             | 2002 | 3.0427 | 3.0427 | 12.4836 |  100.09 |
+---------+--------------------------+------+--------+--------+---------+---------+
```

The structure of the 'data' view can be queried using standard SQL commands (fig. 11). A human-readable and machine-verifiable description of the semantics for each 'data' column is currently provided as an XML file (http://www.crystallography.net/cod/sql/CODDictionary.xml).

When querying data using SQL, user has access to the raw SQL tables, and is therefore responsible for filtering the data to get the desired results. In particular, the COD 'data' table may contain structures that are flagged as retracted ('status = “retracted”' in SQL 'where' statements) or containing errors. These structures are most probably not desired, unless we investigate sociology of structural science, not the structures themselves. In addition, the COD 'data' table contains a small number of marked duplicates, and some structures that were computed by theoretical methods and thus do not represent experimental results (such structures are systematically collected in TCOD). These records are most probably also to be excluded from searches when investigations of crystal structures are carried out. This can be done by the SQL query provided in fig. 8. This query method is recommended for most material structure searches in COD and in its sister databases. The queries performed via the REST interface already perform such filtering, as indicated by the result count in both examples of the fig. 8.

Figure 8. Filtering out structures from the COD MySQL queries.

```
sh% mysql -u cod_reader -h sql.crystallography.net cod -e \
       'select count(*) from data \
        where \
            (status is null or (status != "retracted" and status != "errors")) \
            and duplicateof is NULL \
            and (method is NULL or method != "theoretical")' -NB

365477

sh% curl -sSL 'http://www.crystallography.net/cod/result?format=count'

365477
```

Currently, the COD MySQL tables do not contain atomic coordinate data. A common strategy to get coordinates from SQL queries is to get the list of COD IDs and then convert them either to COD CIF URIs or to local file names than can be retrieved. An example of both strategies (assuming that local COD CIF tree is checked out in the directory ~/struct/cod/cif) is presented in figs. 9 and 10. Fetching coordinates from a copy on a local file system is of course much faster but requires preparation and maintenance of the up-to-date COD copy. We describe in section "Installing a local copy of the COD" on the page 20 how to build such COD copy.

Figure 9. A COD CIF data retrieval after a MySQL query using COD URIs. The requested structures are experimental structures of silicon solved after the year 2000. The '-NB' option provides a plain tab-separated value list (TSV) which is suitable for Unix pipe processing. Please note the 'sleep 1' command inserted after each download which delays the queries and saves the public COD servers from the overload.

```
sh% mysql -u cod_reader -h sql.crystallography.net cod -e \
       'select file from data \
        where \
            (status is null or (status != "retracted" and status != "errors")) \
            and duplicateof is NULL \
            and (method is NULL or method != "theoretical") \
            and formula = "- Si -" \
            and year > 2000' -NB \
    | awk '{print "http://www.crystallography.net/cod/"$1".cif"}' \
    | xargs -i sh -c 'curl -sSL {}; sleep 1' \
    > Si.cif
```

Figure 10. Preparing coordinates for an SQL query using a locally installed COD copy.

```
sh% mysql -u cod_reader -h sql.crystallography.net cod -e \
       'select file from data \
        where \
            (status is null or (status != "retracted" and status != "errors")) \
            and duplicateof is NULL \
            and (method is NULL or method != "theoretical") \
            and formula = "- Si -" \
            and year > 2000' -NB \
    | awk '{print "'${HOME}/struct/cod/cif/'" \
                  substr($0,1,1)"/"substr($0,2,2)"/"substr($0,4,2)"/" \
                  $1".cif"}' \
    | xargs cat \
    > Si.cif
```

Figure 11. Finding column definitions of the COD 'data' view.

```
sh% mysql -u cod_reader -h sql.crystallography.net cod -e 'describe data "R%"'
+-----------+----------------+------+-----+---------+-------+
| Field     | Type           | Null | Key | Default | Extra |
+-----------+----------------+------+-----+---------+-------+
| radiation | varchar(32)    | YES  |     | NULL    |       |
| radType   | varchar(80)    | YES  |     | NULL    |       |
| radSymbol | varchar(20)    | YES  |     | NULL    |       |
| Rall      | float unsigned | YES  |     | NULL    |       |
| Robs      | float unsigned | YES  |     | NULL    |       |
| Rref      | float unsigned | YES  |     | NULL    |       |
| RFsqd     | float unsigned | YES  |     | NULL    |       |
| RI        | float unsigned | YES  |     | NULL    |       |
+-----------+----------------+------+-----+---------+-------+
```

Alternative implementations of COD search on the Web

Since COD is openly accessible on the Web, and all data are free for download, anyone can implement an alternative Web based search engine for COD, and indeed such sites have been implemented already. The oldest is probably the www.nanocrystallography.net web page that uses a subset of COD for teaching purposes. Another chemist-oriented search tool existing at the moment are the MolView on-line molecular viewer by Herman Bergwerf (http://molview.org/) and the DataWarrior standalone Java program by Thomas Sander (http://www.openmolecules.org/), to mention just two mature open-source projects. Other similar endeavours exist on the Web as well.

In addition, the Web base abstractors of chemical information such as PubChem (Bolton et al. 2008) and ChemSpider (Pence and Williams 2010) now provide links to some of the COD structures, and we expect number of such links to grow in the future. In this way, various types of information resources

can be seamlessly integrated on the Web, providing instant access to multiple facets of object description.

When implementing an alternative COD interface, all implementers are encouraged to use the latest revision of the COD, either by regularly updating their local copies using one of the methods described in this chapter, or by querying the on line COD servers. If a subset of COD data is deliberately selected, this should be indicated so that the users of the resource are not confused. If such preclusions are met, additional independent services will provide more possibilities for end users of scientific data, and thus allow them to leverage full potential of open databases, something that is completely impossible with closed archives of data.

Installing a local copy of the COD

Since COD is an open-access database, each user can and may install a local copy of the COD database, a practice which is in fact encouraged.

The first method to obtain a full copy of the COD is to use a Subversion client and to check out a working copy of COD files. The COD Subversion repository is world-readable and can be accessed using Subversion protocol at svn://www.crystallography.net/cod/, with CIF collection only available as a subtree at svn://www.crystallography.net/cod/cif. A command to check out the COD working copy on a Linux operating system is provided in fig. 12; for other platforms, alternative SVN clients can be used (for example TortoiseSVN, https://tortoisesvn.net/, for Windows).

Figure 12. Obtaining (checking-out) a working copy of the COD data using the command line 'svn' Subversion client.

```
sh% svn checkout svn://www.crystallography.net/cod
sh% cd cod; svn update
```

Alternatively, another client that can be currently used to fetch data from the COD Subversion repository is GIT, with the GIT SVN plug-in (readily available in Linux software repositories for most popular Linux distributions). The corresponding cloning commands are provided in fig. 13.

Figure 13. Cloning COD data directory with GIT and GIT SVN.

```
sh% git svn clone svn://www.crystallography.net/cod
sh% cd cod; git svn fetch; git svn rebase
```

Access via Subversion stands out of other methods to obtain COD data by an advantage of easier retrieval of recent changes. Once cloned, a local copy (called “working copy” in Subversion parlance), can be updated, say, per-regular basis, fetching only the changes – modifications, additions and deletions. In addition to that, the 'svn log' (or 'git log' if GIT client is used) commands provide the full history of data additions and changes, with all metadata (dates, committers, changed files) and with human readable log messages. Thus maintaining a Subversion working copy is arguably the best method to have the most up-to-date local mirror of COD data.

If the full history of the COD changes is not needed and the use of Subversion clients is undesired, an incremental update of the local COD copy can be performed using the 'rsync' tool (Davison 2015). The COD file collection is presented to 'rsync' users as the rsync modules 'hkl', 'cif' or 'cod-cif' (for the COD data), 'pcod-cif' (for PCOD data) and 'tcod-cif' (for TCOD data). The commands to synchronise a local tree with the COD database are provided in fig. 14.

Figure 14. Using the 'rsync' program to download and update the COD file collection.

```
sh% rsync -av --delete rsync://www.crystallography.net/cif/ cod-cif/
sh% rsync -av --delete rsync://www.crystallography.net/hkl/ cod-hkl/
sh% rsync -av --delete rsync://www.crystallography.net/cod-cif/ cod-cif/
sh% rsync -av --delete rsync://www.crystallography.net/pcod-cif/ pcod-cif/
```

The provided 'rsync' commands ensure that the local COD file tree becomes exactly the same as the one on the COD server, including deletion (option '--delete') of the removed files. User may want to use additional options, such as '--backup' and '--backup-dir' to preserve copies of the removed files if such references are needed.

The 'rsync' method provides a lean and fast way to synchronise two directories. However, COD file change history is not available when using this method. Moreover, while 'svn' updates are atomic, i.e. they always transfer a complete latest revision even if new commits are taking place simultaneously, the 'rsync' protocol has now knowledge about the Subversion repository transactions and can not ensure that a complete revision is transferred. If an update of COD happens during the 'rsync' process, some transferred files may end up from the newer revision, while the others will be from the older one. To guard against this, running two or more 'rsync' commands in a row is recommended, so that the last command does not fetch any new updates.

File system based queries

When all COD files are available on a local disk, another kind of COD queries becomes possible, namely a query the COD CIFs directly using the standard Unix file processing utilities. While such queries are as a rule slower than the database queries (although with fast disks and large RAM caches they can be speeded up a lot), they are more flexible and do not require building local SQL database or connecting on-line to an existing one.

Being ASCII-encoded text files, CIFs can be searched using the Unix 'grep' and other tools. A query in fig. 15 will find all CIFs that contain line 'diamond' in them, regardless of case. The first command will print out all lines that have this word, and the second command will list names of all files that contain this word (note the 'diamond' in this case can be a name of a mineral, a name of a program or something else).

Figure 15. Search COD CIFs using 'grep'. Options of this command are supported by a GNU 'grep' utility on the Ubuntu-12.04 operating system or higher.

```
sh% grep --exclude '*.svn-base' -H -iR --color diamond cod/cif/
sh% grep --exclude '*.svn-base' -H -iR -l diamond cod/cif/
```

Another powerful method to query and possibly process COD files is the use of 'find' and 'xargs' Unix tools, or employment of the 'make' tool to organise computations. The use of these methods is beyond the scope of our present chapter, but it should be noted that all of them permit running arbitrary programs, written in any programming language, on any subset of COD CIFs.

When using home-written programs for CIF processing, one must take into account that CIF is a structured, free-text format described by a formal syntax (IUCr 2016), and thus requires a correct parser to extract data properly (simple tools like 'awk' or Perl's 'split()' function are not sufficient). Fortunately, numerous parser libraries for proper CIF parsing exist: COD employs a correcting parser from the 'cod-tools' package (Merkys et al. 2016) which has C, Perl and Python bindings; other parsers have been proposed by various authors (Hester 2006; Todorov and Bernstein 2008; Gildea et al. 2011).

For quick composition of different processing tools, however, one can employ simple command line utilities to extract values. The 'cod-tools' package (Merkys et al. 2016) contains such utility, 'cifvalue', that is written entirely in C and permits fast extraction of requested CIF values and their printout in a space-separated-value form which is then easily processed by 'awk', 'perl', 'R', most spreadsheet programs and a multitude of other tools. An example in fig. 16 shows how to use 'cifvalue', in conjunction with the afore-mentioned 'find' and 'xargs' programs, to extract molecular weight, unit cell volume and melting point data from the COD collection.

Figure 16. Use of 'find', 'xargs' and 'cifvalues' from the 'cod-tools' package to extract requested data from CIFs.

```
sh % find cod/cif -name .svn -prune -o -name '*.cif' -print \
     | xargs cifvalues \
         --tags _chemical_melting_point,_chemical_formula_weight,_cell_volume \
     > volumes.dat
```

Programmatic use of COD CIFs

The proper usage of any resource requires mutual understanding between the resource provider and the resource consumer. Since the COD is a completely open database, there are no legal restrictions on the use of data, however, one should be aware of certain COD policies to ensure the optimal utilization of the COD and the validity of the desired results. The COD promises to retain stable structure identifiers, document any changes introduced by the COD maintainers and providing the means of recognising structures unfit for conventional use that were identified as such. Reciprocally, the user of the COD is expected to make use of these premises and apply critical thinking when examining the results; the data set is not yet perfect nor complete, but voluntary collaboration is the driving force behind projects rooted in openness. As a result, reporting of any observed errors and the deposition of new structures to the COD is highly endorsed. Finally, whether one is planing on using the COD for viewing individual structures, processing the whole data set using intricate programs or getting more involved into the project the knowledge of basic COD conventions is tantamount.

Since definitions of structure classes, such as organic compounds and minerals, are often under debate, there is no programmatic classification of structures in the COD. Nevertheless, the user can narrow the search by selecting structures by chemical composition or symmetry, and remove the false positives according to one's needs.

CIFs describing natural minerals can be detected by checking the presence of _chemical_name_mineral CIF data item. However, the addition of this data item is relied upon the depositor, thus the COD can not guarantee that all mineral structures in the database are marked as such.

As described in Data maintenance, CIFs of entries with issues are marked with special data items to be recognised as such both by human users and programs. The main data item to look at is _cod_error_flag, which indicates entries with warnings (enumeration value “warnings”) and errors (enumeration value “errors”). Furthermore, the same data item with value of “retracted” indicates structures, retracted by the authors.

At the time of writing there are around 1100 entries without coordinates in the COD (excluding retracted structures). Most of these entries were created as references of otherwise inaccessible published crystal structures, such as from pre-CIF or paywalled publications. Although the practice of creating such entries is not common and number of them is small, all of them can be filtered out according to the following principle: such entries have _atom_site_ CIF loop with a mock atom site, whose all parameters (label and coordinates) are equal to the value “unknown”, denoted as a lone question mark (“?”, ASCII character 63 decimal)).

Automatically identifying chemical types of the atoms in the CIF file is a bit more complex task than it may seem at first glance. Even though the core CIF dictionary describes a way of specifying the chemical species of the observed atoms, it is often ignored or misused. The recommended practice is to use the “_atom_site_type_symbol” data item that is designated just for this purpose. Alternatively, the chemical type symbols can be prepended to the “_atom_site_label” data item values; for example, following this naming scheme, the “C11”, “Au” and “Pb*” labels would be used to specify carbon, gold and lead (Pb) atoms accordingly. The latter approach seems to be preferred in practice, however, it introduces a lot of ambiguity. First of all, it is not clear whether the user mean to use the labels for this purpose or if he simply forgot to include the “_atom_site_type_symbol” data item. In addition to that, some ambiguity also arises when trying to extract the chemical symbol from the label. Usually, it is sufficient enough to take the first one or two letters from the atom label as its chemical symbol (“/^([A-Za-z]{1,2})/” in regular expression form), however, this approach fails when labels are constructed following some additional arbitrary rules. For example, “HO” and “HOH”, often used to indicate hydroxide and water molecules accordingly, would be recognised as holmium (Ho); other labels often used for water molecules (“Wat”, “W” and “Ow”) demonstrate the flaws of this simplistic approach even further. The maintainers of the COD have adopted a practice of manually putting chemical types to “_atom_site_type_symbol” data items values, if previously empty, thus removing any ambiguity. This, however, is not yet done automatically, as it often requires manual double-checking.

Current widely used approach of splitting same-site atoms into separate _atom_site_ loop entries results in often misinterpretation of sites, that are mixtures of two or more different chemical types. For example, grunerite structure in COD entry 9000000 contains four iron-magnesium sites, that can only be identified as such by comparing their coordinates. We have adopted a practice of marking atoms in such sites as alternative using CIF's _atom_site_disorder_ data items in order to present downstream applications with semantically connected _atom_site_ entries. However, instead of transforming all COD CIFs, we use this practice on-the-fly, as implemented in command line tool cif_mark_disorder from cod-tools package (Merkys et al. 2016).

It is a well-known fact in crystallography that low resolution experiments extract very little to no information on the positions of hydrogen atoms in the structures. There is a wide spectrum of methods for hydrogen position treatment from restraints to geometric prediction. Of course, sometimes hydrogen atoms are completely excluded from crystal structures, especially if their positions are of little interest in the research. It is important, though, to detect such cases for computational analyses in order to avoid misinterpretations. For a known number of hydrogen atoms, attached to a known site, CIF standard defines data item _atom_site_attached_hydrogens. However, there is no recommended notation for a known number of hydrogen atoms, whose sites of attachment are unknown. We have made a decision to “attach” them to a “fake” atom with unknown coordinates (all equal to the special CIF value "unknown", denoted as a lone question mark (“?”, ASCII character 63 decimal)).

Data deposition

An automatic deposition interface was opened in 2010, allowing the scientific community to directly participate in the expansion of the COD data collection. The whole process of insertion of new data, which was detailed beforehand (Gražulis et al. 2009), was automated and embedded into a set of Web pages (accessible at http://www.crystallography.net/cod/deposit) to guide all interested researchers through the deposition of their data in CIF format. Acknowledging a concern about the preservation of the original research data, the COD accepts diffraction data files (in CIF format) as well as atomic coordinates, in line with the publication standards by the IUCr (http://www.iucr.org/home/leading-article/2011/2011-06-02#letter),

The COD accepts three types of depositions:

- Data that was published before the deposition and has a full bibliographic record. Such depositions are accepted from anyone registered at the COD Web site and are immediately put into public domain;
- Pre-publication structures are accepted from the authors of future publications. Contrary to the published material, such structures are not released until the corresponding publication is issued or the hold period expires, although details such as lattice constants, symmetry, summary chemical formula, substance name and the list of authors are made public under persistent COD identifiers that are retained after the release. Coordinates and diffraction data are thus retained confidential within the COD and we assume that such depositions maintain the originality of the submitted work and publications of such structures are eligible as original research. Depositors are granted possibility to extend the hold period up to 18 months after that they are contacted via e-mail and asked either to indicate the publication, make the records public as personal communications (in case the publication does not happen) or, as a last resort, to withdraw it from the COD;
- Structures are also accepted as Personal communications to the COD. Such structures are assumed to be published at the COD by their authors personally and are immediately put into the public domain.

Prior to the automatic deposition interface, all data was collected, corrected and placed in the COD by it's maintainers. Since 2010 all depositions have been directed to the novel interface, thus sparing many man-hours of effort.

Applications

The more obvious application occurs once a crystallographer has determined the cell parameters of a supposedly new phase. Then these cell parameters and the corresponding cell volume can be used in a simple search in COD so as to avoid to waste time if the crystal structure is already published. Full confidence in the result of such a search will wait for the COD attaining completion.

Material identification

Crystal structure databases have for long been used to identify phases in polycrystalline materials. Subsets of databases designed for specific user application (e.g. inorganics, organics, metals …) have been developed and sold separately. Databases containing only diffraction peak positions have also been constructed from structure databases. In both cases (from crystal structure or peak lists) the usual search-match commercial softwares work only on the comparison between peak positions from the database and the ones of the samples to be identified. Consequently, only these structures stored in the actual database can be identified, e.g. organics, forgetting the other phases (inorganics, metal-organics …), except if the user can afford for all databases and corresponding softwares.

Another approach resolving the just previous drawbacks of classical databases is clearly provided using COD. Since COD records all structures independently of their “classification” as inorganics … the search-match results extend to wider ranges of materials (obviously selection on elements, bonds or whatsoever and even phase class can be introduced if necessary). This warrants a more ab initio phase identification whatever the material of concern. But also, the COD open character allows any user to benefit of this aspect using its own software. Such application has recently been developed, called Full-Pattern Search-Match (FPSM), which allows COD-based identification, quantification and microstructural characterization, in an automated way through the internet.

COD and sisters is free for download and use to everybody, even companies. This wonderful value-addition from the academic to the industrial and technological worlds has rapidly been noticed by companies constructing x-ray diffractometers. Crystal Impact was the first company to incorporate COD in the 2000's in its Search-Match software, rapidly followed by Panalytical (Highscore+ software), Bruker (Eva) and Rigaku (PDXL). More recently the 3D Systems company used it for 3D printing crystallographic models, and Kagaku Benran incorporated COD for its Handbook.

Applications for the mining industry

The usefulness of COD for mineral identification proved very useful for practical applications in mining field; in the SOLSA[4] (Sonic Drilling coupled with Automated Mineralogy and chemistry On-Line- On-Mine-Real- Time)[5] project that started this year, the COD is used as essential data provider for identifying minerals for characterisation of the drill cores. The COD is also planned as a vehicle of the subsequent data dissemination, storing results of crystallographic investigations drill cores. All properties of the COD are essential here – open access regime permits efficient distribution and fast access to data; well-established crystallographic CIF framework provides a sound foundation for describing measurement results, and the RESTful interface enables easy integration. It is anticipated that the results of the SOLSA project will be openly available to the community after the project is completed.

Extracting chemical information

Many of the potential users of COD are chemists and so, they will be more interested in the chemical features of any crystallised compound than in the purely crystallographic facts. And for organic and metal-organic chemists, the chemical features of the compound are mostly defined by the statement of how atoms are directly bounded or not to each other: this is the so-called "chemical connectivity" or "molecular structure". Hence, a chemist is more likely to be interested in particular associations of atoms (functional groups, coordination environments) than in unit cell parameters or space groups.

But the molecular structure is not usually explicitly established in the CIF files uploaded to COD and it needs to be deduced from atom coordinates and/or the bond list (if present). And this chemical connectivity should be written in a format suitable to chemically define the compound and to perform searches. Among many available possibilities, we have chosen the SMILES format for this purpose (there are two specifications for this format, the original one elaborated by the Daylight Chemical Information Systems, (Funatsu et al. 1988), http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html and an open specification established afterwards, http://opensmiles.org, both are essentially identical). This format represent a chemical species by a single chain of ASCII characters and has the advantages of storing only the molecular structure and nothing else, which makes it very compact, and of being both human and machine writeable/readable, which is convenient for both automatic or manual edition. With some practice, it is possible to directly "see" the molecular structure (in simple cases) or at least important features of it (in more complicated ones) just reading the SMILES and there are several informatics tools able to depict the molecular structure for a given SMILES (for example, indigo-depict: http://lifescience.opensource.epam.com/indigo/).

The SMILES format presents, however, also important drawbacks: it has been designed with the valence-bond theory in mind (but the very concept of "chemical connectivity" somehow implies the valence-bond theory) and hence, it has problems representing species that are not well-explained by this theory, like delocalized bonds (other than aromatic rings) or polycentric bonds (metalocenes,

4 http://www.solsa-mining.eu/
5 https://ec.europa.eu/easme/en/printpdf/7079

boranes, …). Another drawback is that it can only represent discrete species and not polymeric ones, for which only a fragment may be represented.

Deriving the molecular connectivity as a SMILES chain from the corresponding CIF file is however far from being a trivial task. We are using Open Babel toolbox (O'Boyle et al. 2011) (http://openbabel.org) which, in principle, has the ability for performing the CIF to SMILES conversion, but the result is not optimum in many cases. To begin with, Open Babel reads the atoms as they are in the input file, does not perform any symmetry generation or consider the occupancy factors, hence does not handle properly chemical species placed in symmetry elements of the lattice neither it considers the possible disorder. To circumvent these problems, algorithms and corresponding software have been developed by COD maintainers (Gražulis et al. 2015).

But even if we have a set of atoms chemically representative of our compound, there are still important problems to face regarding the choice of the best representation for any particular chemical species since Open Babel, in many cases, does not yield a SMILES displaying the schematic image that most chemists will have about it, image that, after all, is just a conventional one. Most problems arise from the fact that Open Babel has been designed from the point of view of an organic chemist and in the realm of valence bond theory, trying to settle every atom acting with its usual valence. The number of bonds that an atom can form is also limited (probably for program efficiency reasons) making it necessary very often to supplement the bonds found by Open Babel with those provided by the authors in the _geom_bond_distance_ loop.

For the previous reasons, the obtained crude SMILES usually represent accurately organic compounds (easily recognisable by the absence of square brackets) that may be accepted without further treatment, but it does not happen the same with metal-organic compounds, for which is very frequent to find missing bonds, spurious or lacking H-atoms, wrong bond representations, etc.: the list of compound families showing these kinds of problems is quite large. At present, the curation of such SMILES is done mostly by human intervention with the aid of a number of helper scripts that identify and, in some cases, automatically solve the problems associated with some of the more frequently found families of compounds. It is noteworthy that human intervention in this task has not been eliminated even by the proprietary and not released software and not described-in-detail algorithms used by commercial databases (Bruno et al. 2011).

Due to these reasons, the number of entries with SMILES that has been considered as acceptable is, at present, just about one third of the total number of COD entries. The procedure needs to be improved in order to accelerate the conversion and diminish the need for human intervention.

The establishment of the chemical identity of COD entries is quite useful to cross-link COD with other chemical databases. In this sense, the available SMILES have been already used to set around 35000 links between COD and the open chemical database ChemSpider ((Pence and Williams 2010); http://www.chemspider.com) and it is expected that the same can be used for other important open databases like PubChem.

The built SMILES are also used to perform substructure searches, in which the user of the database tries to find all compounds containing a given molecular fragment, this is surely the main kind of search that an organic or metal-organic chemist is interested in, since such molecular fragments is the main way of defining families of compounds. COD website implements such searches by allowing the user to introduce the fragment also in SMILES format and then use Open Babel fast search utility to get the hits. For the benefit of users that are not familiar with the SMILES format, the query may also be built in COD website using graphical interfaces written either in JavaScript (Bienfait and Ertl 2013) or Java (http://www.molinspiration.com/jme/) languages. The whole SMILES collection is also downloadable as a single file (http://www.crystallography.net/cod/smi/allcod.smi) so that the user can

perform the search locally with any software of his/her choice. An interesting possibility is to use Open Babel package without the involvement of a fast search index: this procedure is much slower than the above mentioned fast search (it takes several minutes, which makes it difficult to implement in the web interface), but it yields more accurate results and the query can be written in the SMARTS language (http://www.moldb.net/opensmarts; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html), which allows for more versatile and sophisticated searches than SMILES.

Property search (Seeking properties?)

Modern methods of computational chemistry are able to greatly reduce the efforts in such fields as material science. In silico experiments can quite accurately predict various properties of the materials without the need of time- and cost-intensive synthesis and experimentation. For example, knowledge of crystal contents and densities is sufficient enough to carry out the search of possible hydrogen storage materials, as demonstrated by Breternitz and Gregory in their research using the COD (Breternitz and Gregory 2015). Mounet et al. has embarked on the screening for crystal structures with periodic layered compounds in order to identify novel graphene-like compounds in both COD and ICSD (Mounet et al. 2016).

Geometry statistics

In order to simplify and encourage similar research in the COD, we are developing a database for the geometry of COD structures. Our main goals are to collect bond lengths, valence and dihedral angle sizes, and provide their descriptions in the form of statistical models. To achieve that, we have devised a novel descriptor for chemical environment, that is, a “name”, allowing to group geometric parameters, measured from similar compounds (Long et al. 2016a,b). We have chosen a “fuzzy” descriptor as a balance between too strict matching, that would yield huge numbers of classes with small number of observations, and short-sightedness. However, there are cases when geometry parameters from chemically different environments fall in the same class thus yielding multimodal or skewed distributions. In order to accommodate such irregularities we have chosen mixture models of Gaussian and Cauchy distributions.

Thus, we have developed fully automatic software, capable of extracting aforementioned geometry parameters from crystal structure descriptions without the need of human supervision. With this software we have extracted geometric parameters from more than 300 000 small-molecule entries from the COD to date. To ease browsing of the collected geometry dataset and the describing models, we have launched a Web interface. Currently, browsing is implemented using aforementioned atom descriptors.

One of the possible uses of our geometry database is the detection of common geometric features. The semi-automated search for artefacts and outliers in the crystal structures is another possible use. Furthermore, derived statistical distributions from our database could be used to generate forcefields in modelling, as well as constraints or restraints for the refinement of crystal structures. This particular approach is being used by Murshudov et al. to compile a dictionary of constraints for macromolecular structure refinement using REFMAC5 refinement program (Vagin et al. 2004; Long et al. 2014).

High-throughput computations

Successful usage of the results from high-throughput in silico research is somewhat hindered by the problem of reproducibility. A key to this problem is to preserve provenance for all steps, leading from the inputs to the results (Mesirov 2010). To aid the field of atomistic simulations, Pizzi et al. have developed AiiDA framework (Pizzi et al. 2016), based on Open Provenance Model (Moreau et al.

2007; Moreau et al. 2008). AiiDA is able to automate the execution of computations, automatically store inputs and results in a tailored database, while keeping track of data provenance and helping to share the results.

In order to ease the importing and exporting data to and from AiiDA, it was interfaced with the COD and TCOD. Current pipeline allows seamless importing of experimental data from the COD to AiiDA for further atomistic simulations at the same time preserving all metadata, required for unambiguous identification of inputs, and exporting of the results, bundled together with all metadata, required for reproducibility to the TCOD.

Applications in college education and complementing outreach activities

Crystallographic open access databases have been built up form 2004 onwards for educational purposes at Portland State University (Sondergeld et al. 2016). The focus of these activities has always been interactive visualizations of crystal structures with educational relevance. The well known Java-based Jmol plug-in (now replaced by the more secure JavaScript version known as JSmol) into web browsers by Bob Hanson and his team at St. Olaf Collage in Minnesota/USA has been adapted for this purpose (Moeck et al. 2005).

In recent years, we augmented our educational activities with 3D printed crystallographic models (Moeck et al. 2014). The key to these activities was a Windows executable program by Werner Kaminsky (Stone-Sundberg et al. 2015) that converts *.cif files directly into *.stl or *.wrl files, as required for the 3D printing process. Note that there are also Windows executable programs by Werner Kaminsky that create 3D print files for crystal morphology models (Stone-Sundberg et al. 2015) and longitudinal representation surfaces for anisotropic crystal-physics properties (Stone-Sundberg et al. 2015).

While the CIF dictionaries contain provisions to encode crystal morphologies in *.cif files directly so that it can be read into Werner’s program (Kaminsky 2007), the developers of the Material Properties Open Database (Pepponi et al. 2012) needed to write their own modified CIF extension dictionary. 3D print files can also be created directly at the website of the Material Properties Open Database (Fuentes-Cobas et al. 2014). Selected 3D print files and CIF encoded crystal morphologies are available for download at the above mentioned educational project of Portland State University (Sondergeld et al. 2016).

Perspectives

Historic structures

As of August 2016, most of the structures in COD are published in “CIF era” (1990s and later), with the contribution of older structures equal to 8% (27 000 entries). However, it is assumed that the amount of published pre-CIF structures is much larger, and much effort has to be made to digitalise and deposit them as CIFs. Therefore, we have produced a few dozens of such entries manually, but laborious nature of such task prevents the conversion from attaining speed. Nevertheless, the collection of historic structures can be sped up by harnessing crowdsourcing for detection of coordinate tables in scanned publications, optical character recognition and evaluation of geometry as a means for error detection.

Theoretical data in (T)COD

For some time CIF format has enjoyed being de facto standard for reporting and archiving of the results of experimental crystal structure solutions. It was adopted and used by the most of crystallographic journals as well as structural databases. New CIF dictionaries are being developed to define ontologies in such fields as macromolecular crystallography (Fitzgerald et al. 2006), powder diffraction (Toby et al. 2003), electron density studies (Mallinson and Brown 2006). However, much effort is still needed to consolidate the knowledge in the field of theoretical material science, which is expanding rapidly currently. Nevertheless, there are a few disjoint attempts, namely, European Theoretical Spectroscopy Facility (ETSF) (Caliste et al. 2008), (Gonze et al. 2008) and NoMaD (Mohamed 2016). Addressing this issue, the TCOD has been launched, adopting the practice of using the CIF format, approach-specific dictionaries (for example, cif_dft dictionary for DFT) and defining data validation criteria for automated checks. In addition, TCOD puts emphasis on the provenance of the results and reproducibility by devising a special dictionary for related metadata – cif_tcod (Gražulis 2016). TCOD, accompanied with a huge collection of experimental structures in the COD (Gražulis et al. 2012), opens an immediate potential for the cross-validation of experimental and theoretical data.

Concluding remarks

The 13 years of COD development demonstrate that it is possible to build a fully open-access, high quality database in a well-defined area of scientific inquiry, namely in the field of crystallography. During its history the COD was on-line most of the time, expect for short technical glitches. Its volume grows constantly over time, and it enjoys increasing number of citations as well. Although not yet covering every published structure, the COD is suitable for many applications and impossible to substitute when openness is an essential requirement. We see a large potential of open data in the new, connected world, with many self-evident but also unanticipated uses of scientific results for the benefit of everyone, and will continue to develop and support the COD into the future.

Acknowledgements

We acknowledge financial supports from the Research Council of Lithuania (grant numbers MIP-124/2010 and MIP-025/2013), the European Community (SOLSA, 2016-2020, grant agreement n°689868) and the Conseil Régional de Normandie (COMBIX project, 2013-2014, Chair of Excellence of LL).

References

Annis J., Bakken J., Holmgren D., Petravick D. and Rechenmacher R. 1999 The Sloan Digital Sky Survey data acquisition system, and early results. 2-5.

Aroyo, M. I., Perez-Mato, J. M., Capillas, C., Kroumova, E., Ivantchev, S., Madariaga, G., Kirov, A. and Wondratschek, H. 2006 Bilbao Crystallographic Server: I. Databases and crystallographic computing programs. *Zeitschrift für Kristallographie - Crystalline Materials*, 221(1): 15-27. DOI: 10.1524/zkri.2006.221.1.15

Aroyo, M. I., Perez-Mato, J. M., Orobengoa, D., Tasci, E., de la Flor, G. and Kirov, A. 2011 Crystallography online: Bilbao Crystallographic Server. *Bulgarian Chemical Communications*, 43(2): 183-197.

Authors Of Wikipedia 2016 *Hipparchus*. Available at https://en.wikipedia.org/wiki/Hipparchus [Last accessed 2016-10-16]

Baerlocher C., McCusker L. and Olson D., 2007 *Atlas of Zeolite Framework Types, 6th revised edition*. Elsevier, .

Baldi, P. 2011 Data-driven high-throughput prediction of the 3-D structure of small molecules: review and progress. A response to the letter by the Cambridge Crystallographic Data Centre. *Journal of chemical information and modeling*, 51: 3029. DOI: 10.1021/ci200460z

Belsky, A., Hellenbrandt, M., Karen, V. L. and Luksch, P. 2002 New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Crystallographica Section B*, 58: 364-369. DOI: 10.1107/S0108768102006948

Berman, H., Kleywegt, G., Nakamura, H. and Markley, J. 2012 The Protein Data Bank at 40: Reflecting on the Past to Prepare for the Future. *Structure* , 20: 391-396. DOI: 10.1016/j.str.2012.01.010

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. DOI: 10.1016/S0006-3495(92)81649-1

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. and Bourne, P. E. 2000 The Protein Data Bank. *Nucleic Acids Research*, 28: 235-242. DOI: 10.1093/nar/28.1.235

Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. and Westrip, S. P. 2016 Specification of the Crystallographic Information File format, version 2.0. *Journal of Applied Crystallography*, 49(1): 277-284. DOI: 10.1107/s1600576715021871

Bienfait, B. and Ertl, P. 2013 JSME: a free molecule editor in JavaScript. *Journal of Cheminformatics*, 5: 24. DOI: 10.1186/1758-2946-5-24

Bolton E. E., Wang Y., Thiessen P. A. and Bryant S. H. 2008 *Chapter 12 PubChem: Integrated Platform of Small Molecules and Biological Activities*. In: Wheeler R. A. and Spellmeyer D. C. (Ed.), *Annual Reports in Computational Chemistry*, Elsevier: Oxford, UK. DOI: 10.1016/S1574-1400(08)00012-1

Breternitz, J. and Gregory, D. 2015 The Search for Hydrogen Stores on a Large Scale; A Straightforward and Automated Open Database Analysis as a First Sweep for Candidate Materials. *Crystals*, 5: 617-633. DOI: 10.3390/cryst5040617

Bruno, I. and Groom, C. 2014 A crystallographic perspective on sharing data and knowledge. *Journal of Computer-Aided Molecular Design*, : 1-8. DOI: 10.1007/s10822-014-9780-9

Bruno, I. J., Shields, G. P. and Taylor, R. 2011 Deducing chemical structure from crystallographically determined atomic coordinates. *Acta Crystallogr Sect B Struct Sci*, 67(4): 333–349. DOI: 10.1107/s0108768111024608

Caliste, D., Pouillon, Y., Verstraete, M., Olevano, V. and Gonze, X. 2008 Sharing electronic structure and crystallographic data with ETSFIO. *Computer Physics Communications* , 179: 748-758. DOI: 10.1016/j.cpc.2008.05.007

Chateigner, D., Grazulis, S., Pérez, O., Pepponi, G. and Lutterotti, L. 2015 COD, PCOD, TCOD, MPOD… open structure and property databases. , : .

Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A. I., Sweeney, B., Zirbel, C. L., Leontis, N. B. and Berman, H. M. 2014 The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res.*, 42: D114-D122. DOI: 10.1093/nar/gkt980

Collins-Sussman B., Fitzpatrick B. W. and Pilato C. M., 2004 *Version Control with Subversion: Next Generation Open Source Version Control*. O'Reilly Media, .

Collins-Sussman B., Fitzpatrick B. W. and Pilato C. M. 2011 Version Control with Subversion.. Available at http://svnbook.red-bean.com/

Davison W. 2015 *Rsync*. Available at http://samba.anu.edu.au/rsync/ [Last accessed 2016-11-06 13:42 EET]

Day, N., Downing, J., Adams, S., England, N. W. and Murray-Rust, P. 2012 CrystalEye: automated aggregation, semantification and dissemination of the world's open crystallographic data. *Journal of Applied Crystallography*, 45: 316-323. DOI: 10.1107/S0021889812006462

Downs, R. T. and Hall-Wallace, M. 2003 The American Mineralogist crystal structure database. *American Mineralogist*, 88: 247-250.

Eger T., Scheufen M. and Meierrieks D. 2013 *The Determinants of Open Access Publishing: Survey Evidence from Germany*. Available at http://ssrn.com/abstract=2232675

Eysenbach, G. 2006 Citation Advantage of Open Access Articles. *PLoS Biol*, 4(5): e157. DOI: 10.1371/journal.pbio.0040157

Faber, J. and Fawcett, T. 2002 The Powder Diffraction File: present and future. *Acta Crystallographica Section B*, 58(3 Part 1): 325-332. DOI: 10.1107/S0108768102003312

Falcioni, M. and Deem, M. W. 1999 A biased Monte Carlo scheme for zeolite structure solution. *Journal of Chemical Physics*, 110(3): 1754-1766. DOI: 10.1063/1.477812

Fielding R. T. 2000 Architectural Styles and the Design of Network-based Software Architectures.. Available at https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm

Fitzgerald P. M. D., Westbrook J. D., Bourne P. E., McMahon B., Watenpaugh K. D. and Berman H. M. 2006 *4.5. Macromolecular dictionary (mmCIF)*. In: Hall S. R. and McMahon B. (Ed.), *International Tables for Crystallography*, International Union of Crystallography. DOI: 10.1107/97809553602060000745

Fuentes-Cobas, L., Chateigner, D., Pepponi, G., Muñoz-Romero, A., Ramírez-Amparan, G., Templeton-Olivares, I., Sánchez-Aroche, D., Hernández-Montes, J., Márquez de la Mora Núñez, A. and López-Carrasco, M. 2014 Implementing Graphic Outputs for the Material Properties Open Database (MPOD). *Acta Cryst*, 70: C1039. DOI: 10.1107/S2053273314089608

Funatsu, K., Miyabayashi, N. and Sasaki, S. 1988 Further development of structure generation in the automated structure elucidation system CHEMICS. *Journal of Chemical Information and Modeling*, 28(1): 18–28. DOI: 10.1021/ci00057a003

Galperin, M. Y. and Cochrane, G. R. 2010 The 2011 Nucleic Acids Research Database Issue and the online Molecular Biology Database Collection. *Nucleic Acids Research*, 39(Database): D1-D6. DOI: 10.1093/nar/gkq1243

Gildea, R. J., Bourhis, L. J., Dolomanov, O. V., Grosse-Kunstleve, R. W., Puschmann, H., Adams,

P. D. and Howard, J. A. K. 2011 iotbx.cif: a comprehensive CIF toolbox. *Journal of Applied Crystallography,* 44: 1259-1263. DOI: 10.1107/S0021889811041161

Gilliland, G. L., Tung, M. and Ladner, J. E. 2002 The Biological Macromolecule Crystallization Database: crystallization procedures and strategies. *Acta Crystallographica Section D,* 58(6 Part 1): 916-920. DOI: 10.1107/S0907444902006686

Gonze X., Almbladh C.-O., Cucca A., Caliste D., Marques M., Freysoldt C., Olevano V., Pouillon Y., Sottile F. and Verstraete M. 2008 Specification of file formats for ETSF Specification version 3.3. Second revision for this version (SpecFF ETSF3.3).. Available at http://www.etsf.eu/system/files/SpecFFETSF_v3.3.pdf

Gražulis S. 2016 *TCOD mailing list*. Available at http://lists.crystallography.net/cgi-bin/mailman/listinfo/tcod [Last accessed 2016-04-13]

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. *Journal of Applied Crystallography,* 42: 726-729. DOI: 10.1107/S0021889809016690

Gražulis, S., Daškevič, A., Merkys, A., Chateigner, D., Lutterotti, L., Quirós, M., Serebryanaya, N. R., Moeck, P., Downs, R. T. and Le Bail, A. 2012 Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Research,* 40: D420-D427. DOI: 10.1093/nar/gkr900

Gražulis, S., Merkys, A., Vaitkus, A. and Okulič-Kazarinas, M. 2015 Computing stoichiometric molecular composition from crystal structures. *Journal of Applied Crystallography,* 48: 85-91. DOI: 10.1107/S1600576714025904

Groom, C. R., Bruno, I. J., Lightfoot, M. P. and Ward, S. C. 2016 The Cambridge Structural Database. *Acta Crystallographica Section B,* 72(2): 171-179. DOI: 10.1107/S2052520616003954

Hall, S. R., Allen, F. H. and Brown, I. D. 1991 The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallographica Section A,* 47: 655-685. DOI: 10.1107/S010876739101067X

Harnad, S. and Brody, T. 2004 Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine,* 10(6): .

Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Hajjem, C. and Hilf, E. R. 2008 The Access/Impact Problem and the Green and Gold Roads to Open Access: An Update. *Serials Review,* 34(1): 36-40. DOI: 10.1080/00987913.2008.10765150

Hester, J. R. 2006 A validating CIF parser: it PyCIFRW. *Journal of Applied Crystallography,* 39: 621-625. DOI: 10.1107/S0021889806015627

Hewett J. 2006 *LHC Factoids*. Available at http://blogs.discovermagazine.com/cosmicvariance/2006/09/27/lhc-factoids/ [Last accessed 2016-10-16]

IUCr 2016 *CIF Version 1.1 Working specification*. Available at http://www.iucr.org/resources/cif/spec/version1.1 [Last accessed 2016-11-06 14:55 EET]

Kaminsky, W. 2007 From CIF to virtual morphology: new aspects of predicting crystal shapes as part

of the WinXMorph program. *J. Appl. Cryst.*, 40: 382-385. DOI: 10.1107/S0021889807003986

Lafuente B., Downs R. T., Yang H. and Stone N. 2015 *The power of databases: the RRUFF project*. In: Armbruster T. and Danisi R. M. (Ed.), *Highlights in Mineralogical Crystallography*, W. De Gruyter.

Le Bail, A. 2005 Inorganic structure prediction with it GRINSP. *Journal of Applied Crystallography*, 38: 389-395. DOI: 10.1107/S0021889805002384

Long, F., Gražulis, S., Merkys, A. and Murshudov, G. N. 2014 A new generation of CCP4 monomer library based on Crystallography Open Database. *Acta Cryst. A*, 70: C338.

Long F., Nicholls R. A., Emsley P., Gražulis S., Merkys A., Vaitkus A. and Murshudov G. N. 2016a ACEDRG: A stereo-chemical description generator for ligands..

Long F., Nicholls R. A., Emsley P., Gražulis S., Merkys A., Vaitkus A. and Murshudov G. N. 2016b Validation and extraction of stereochemical information from small molecular databases..

Mallinson P. R. and Brown I. D. 2006 *3.5*. In: (Ed.), *International Tables for Crystallography, Vol. G*, Chester: International Union of Crystallography.

Merkys, A., Vaitkus, A., Butkus, J., Okulič-Kazarinas, M., Kairys, V. and Gražulis, S. 2016 it COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *Journal of Applied Crystallography*, 49(1): 292-301. DOI: 10.1107/S1600576715022396

Mesirov, J. P. 2010 Computer science. Accessible reproducible research.. *Science (New York, N.Y.)*, 327: 415-6. DOI: 10.1126/science.1179653

Moeck, P., Stone-Sundberg, J., Snyder, T. J. and Kaminsky, W. 2014 Enlivening a 300 level general education class on nanoscience and nanotechnology with 3D printed crystallographic models. *J. Mater. Edu.*, 36: 77-96.

Moeck P., Čertík O., Upreti G., Garrick W. and Fraundorf P. 2005 Crystal structure visualizations in three dimensions with database support. **909E**,.

Mohamed F. R. 2016 *Nomad meta info*. Available at https://gitlab.rzg.mpg.de/nomad-lab/nomad-meta-info/wikis/home [Last accessed 2016-02-18]

Moreau L., Freire J., Futrelle J., McGrath R. E., Myers J. and Paulson P. 2008 *The Open Provenance Model: An Overview*. In: Freire J., Koop D. and Moreau L. (Ed.), *Provenance and Annotation of Data and Processes*, Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-89965-5_31

Moreau L., Freire J., Futrelle J., McGrath R. E., Myers J. and Paulson P. 2007 The Open Provenance Model.. Available at http://eprints.soton.ac.uk/264979/

Mounet N., Schwaller P., Cepellotti A., Merkys A., Castelli I., Gibertini M., Pizzi G. and Marzari N. 2016 High-throughput prediction of two-dimensional materials..

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T. and Hutchison, G. R. 2011 Open Babel: An open chemical toolbox. *Journal of Cheminformatics*, 3: 33. DOI: 10.1186/1758-2946-3-33

Pence, H. E. and Williams, A. 2010 ChemSpider: An Online Chemical Information Resource. *Chemical Education Today*, 87: 1123-1124. DOI: 10.1021/ed100697w

Pepponi, G., Gražulis, S. and Chateigner, D. 2012 MPOD: A Material Property Open Database

linked to structural information. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 284(0): 10 - 14. DOI: 10.1016/j.nimb.2011.08.070

Perez-Mato, J., Gallego, S., Tasci, E., Elcoro, L., de la Flor, G. and Aroyo, M. 2015 Symmetry-Based Computational Tools for Magnetic Crystallography. *Annual Review of Materials Research*, 45(1): 217-248. DOI: 10.1146/annurev-matsci-070214-021008

Piwowar, H. A. and Vision, T. J. 2013 Data reuse and the open data citation advantage. *PeerJ*, 1: e175. DOI: 10.7717/peerj.175

Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. and Kozinsky, B. 2016 AiiDA: automated interactive infrastructure and database for computational science. *Computational Materials Science*, 111: 218-230. DOI: 10.1016/j.commatsci.2015.09.013

PPARC 2006 *'Maiden Flight' for LHC Computing Grid Breaks Gigabyte-per-Second Barrier*. Available at http://phys.org/news/2006-02-maiden-flight-lhc-grid-gigabyte-per-second.html [Last accessed 2016-10-16]

Protein Data Bank 1971 Protein Data Bank. *Nature New Biology*, 233: 223. DOI: 10.1038/newbio233223b0

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. (Ed.), *Geoinformatics: Data to Knowledge*, Geological Society of America. DOI: 10.1130/2006.2397(06)

Rauber A., Asmi A., van Uytvanck D. and Pröll S. 2015 Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC).. Available at https://rd-alliance.org/group/data-citation-wg/wiki/wgdc-dynamic-data-citation-recommendations.html

Rauber A., Asmi A., van Uytvanck D. and Pröll S. 2016 *Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use*. Available at https://www.rd-alliance.org/group/data-citation-wg.html

Sadowski, P. and Baldi, P. 2013 Small-Molecule 3D Structure Prediction Using Open Crystallography Data. *Journal of Chemical Information and Modeling*, 53: 3127-3130. DOI: 10.1021/ci4005282

Sondergeld P., Plachinda P., Gledhill G., Dempsey A., Harshfield H., DeStefani E., Lerud R. and Moeck P. 2016 *Open Access Crystallography*. Available at http://nanocrystallography.research.pdx.edu [Last accessed 2016-09-14]

Stone-Sundberg, J., Kaminsky, W., Snyder, T. and Moeck, P. 2015 3D printed models of small and large molecules, structures and morphologies of crystals, as well as of their anisotropic physical properties. *Cryst. Res. Technol.*, : 1-11. DOI: 10.1002/crat.201400469

Toby, B. H., Von Dreele, R. B. and Larson, A. C. 2003 CIF applications. XIV. Reporting of Rietveld results using pdCIF: GSAS2CIF. *Journal of Applied Crystallography*, 36: 1290-1294.

Todorov, G. and Bernstein, H. J. 2008 it VCIF2: extended CIF validation software. *Journal of Applied Crystallography*, 41: 808-810. DOI: 10.1107/S002188980801385X

Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. and Murshudov, G. N. 2004 it REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallographica Section D*, 60(12): 2184-2195. DOI:

10.1107/S0907444904023510

Villars, P., Berndt, M., Brandenburg, K., Cenzual, K., Daams, J., Hulliger, F., Massalski, T., Okamoto, H., Osaki, K., Prince, A., Putz, H. and Iwata, S. 2004 The Pauling File, Binaries Edition . *Journal of Alloys and Compounds* , 367(1–2): 293 - 297. DOI: 10.1016/j.jallcom.2003.08.058

Villars, P., Onodera, N. and Iwata, S. 1998 The Linus Pauling file (LPF) and its application to materials design . *Journal of Alloys and Compounds* , 279: 1 - 7. DOI: 10.1016/S0925-8388(98)00605-7

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 Crystallographica Section B*, 58: 343-348. DOI: 10.1107/S0108768102002902

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. , : .

Hahn T. (Ed.), 2006 *International Tables for Crystallography*. International Union of Crystallography, . DOI: 10.1107/97809553602060000100