1

Crystallography Open Database: History, Development, and Perspectives

Saulius Gražulis[1], Andrius Merkys[1], Antanas Vaitkus[1], Daniel Chateigner[2], Luca Lutterotti[3], Peter Moeck[4], Miguel Quiros[5], Robert T. Downs[6], Werner Kaminsky[7], and Armel Le Bail[8]

[1] *Vilnius University, Institute of Biotechnology, Department of Protein-DNA Interactions, Saulėtekio al. 7, 10257 Vilnius, Lithuania*
[2] *Normandie Université, Université de Caen Normandie, CRISMAT-CNRS, ENSICAEN, IUT-Caen, boulevard du Maréchal Juin, 6, 14050, Caen Cedex, France*
[3] *University of Trento, Department of Industrial Engineering, Via Sommarive 9, 38123, Trento, Italy*
[4] *Portland State University, Department of Physics, 1719 SW 10th Avenue, Portland, OR 97201, USA*
[5] *Universidad de Granada, Departamento de Química Inorgánica, Facultad de Ciencias, Avenida de Fuentenueva, 18071, Granada, Spain*
[6] *University of Arizona, Department of Geosciences, 1040 E 4 Street, Tucson, AZ 85721, USA*
[7] *University of Washington at Seattle, Department of Chemistry, 4000 15th Avenue NE 36 Bagley Hall, Seattle, WA 98195-1700, USA*
[8] *Université du Maine, Institut des Molécules et des Matériaux du Mans, Département des Oxydes et Fluorures, CNRS UMR 6283, 72085 Le Mans, France*

1.1 Introduction

Science is crucially based on observational data. As an example of an ancient data-driven discovery, the observation of equinox precession by Hipparchus around 130 BCE comes to mind [1] – 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 Timocharis 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 terabytes (TB) [2] to as much as 15 petabytes (PB) of data per year [3, 4].

In the field of crystallography, the need of long-term data preservation was recognized 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 [5] and created the crystallographic information file/framework (CIF) standard for crystallographic

Materials Informatics: Methods, Tools and Applications,
First Edition. Edited by Olexandr Isayev, Alexander Tropsha, and Stefano Curtarolo.

Table 1.1 Material property and structure databases available online.

No.	Database	Approx. no. of records	License	Current URL	Est.	References
1.	MPOD	300	Public domain	http://mpod.cimav.edu.mx	2010	[10]
2.	RRUFF	47 000	Open access	http://rruff.info/	2015	[11]
3.	AMCSD	20 000	Open access	http://rruff.geo.arizona.edu/AMS/amcsd.php	2003	([12]; [13])
4.	IZA Zeolite database	176[a)]	Open access	http://www.iza-structure.org/databases/	1996	[14]
5.	Bilbao server	—	—	http://www.cryst.ehu.es	1997	[15]
6.	B-IncStrDB (Bilbao Incommensurate Structures Database)	140	Open access	http://webbdcrista1.ehu.es/incstrdb/	2010	[16]
7.	MAGNDATA (Bilbao Magnetic Structure Database)	428	Open access	http://webbdcrista1.ehu.es/magndata/	2015	[17]
8.	NDB	8 600	Open access	http://ndbserver.rutgers.edu/	1992	([18]; [19])
9.	COD	400 000	Public domain	http://www.crystallography.net/cod	2003	([20]; [21])
10.	PCOD	1 000 000	Public domain	http://www.crystallography.net/pcod	2003	[22]
11.	TCOD	2 000	Public domain	http://www.crystallography.net/tcod	2013	[23]
12.	CSD	800 000	Subscription based	http://www.ccdc.cam.ac.uk/solutions/csd-system/components/csd/	1965	[24]
13.	ICSD	200 000	Subscription based	https://icsd.fiz-karlsruhe.de/	1987	[25]
14.	PDF	380 000	Subscription based	http://www.icdd.com/products/pdf4.htm	1941	[9]
15.	CRYSTMET	170 000	Subscription based	http://www.TothCanada.com,[b)] https://cds.dl.ac.uk/cgi-bin/news/disp?crystmet	1996	[26]
16.	Linus Pauling file	290 000	Subscription based; free of charge queries accepted[c)]	http://paulingfile.com, http://crystdb.nims.go.jp/index_en.html	1995	([27]; [28])
17.	PDB	124 000	Open access	http://www.rcsb.org/pdb	1971	([29]; [30])
18.	BMCD	43 000	Open access	http://xpdb.nist.gov:8060/BMCD4	1995	[31]

a) The number of unique zeolite framework types that had been approved and assigned a 3-letter code by the Structure Commission of the IZA.

b) The page at the http://www.TothCanada.com advertised in [26] seems no longer operational, but the access for subscribers is advertised at https://cds.dl.ac.uk/cgi-bin/news/disp?crystmet.

c) Free of charge queries are offered at http://crystdb.nims.go.jp/index_en.html, but "no reproduction, republication or distribution to third parties of any content is permitted without written permission of NIMS."

data exchange [6, 7], which is constantly maintained to address new challenges in data management [8]. Crystal diffraction data has been accumulated systematically in a number of databases since as early as 1941 [9], archived in various crystallographic databases (Table 1.1), the largest ones being the Crystallography Open Database (COD) [21], the Cambridge Structural Database (CSD) [24], the Inorganic Crystal Structure Database (ICSD) [25], the Pauling File [28], the Protein Data Bank (PDB) [30], the Powder Diffraction File (PDF) from International Centre for Diffraction Data (ICDD) [9], and the CRYSTMET [26]. Several other databases that focus on specific aspects of crystallographic data exist; the structures they mention are usually included in one or several above-mentioned databases. References to these specialized databases will be given in the following text.

Before 2003, of the above-mentioned crystal structure archives, only the 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 nonsubscribers, as well as requiring purchase of a license for systematic data searches and, occasionally, restricting publication of derived data [32, 33]. The advent of the Web, ubiquitous computing, and advantages of open linked data prompted a group of crystallographers to initiate the COD, offering crystal structures for chemical crystallography on similar grounds as the PDB provides them for macromolecular crystallography. Currently, 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 [34], they still require acquisition of licenses for systematic data searches.

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 the fields of crystallography, chemistry, material identification, and teaching.

1.2 Open Databases for Science

Over the years, various researchers found that open access to articles consistently increases citations of these publications [35–39]. Similar trends are observed for data in the field of bioinformatics [40], 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 user of data, the absence of paywalls and use restrictions provides the convenience of one-click access to data. Finally, there are ethical considerations – most published research were funded by public money, and the society members whose taxes were used to produce scientific results have reasonable expectations that these results would be available to them without demand of extra payment and without restrictions. Understandably, then, many funding agencies require that researchers whom they have supported publish their results under open-access licenses for both publications and data.

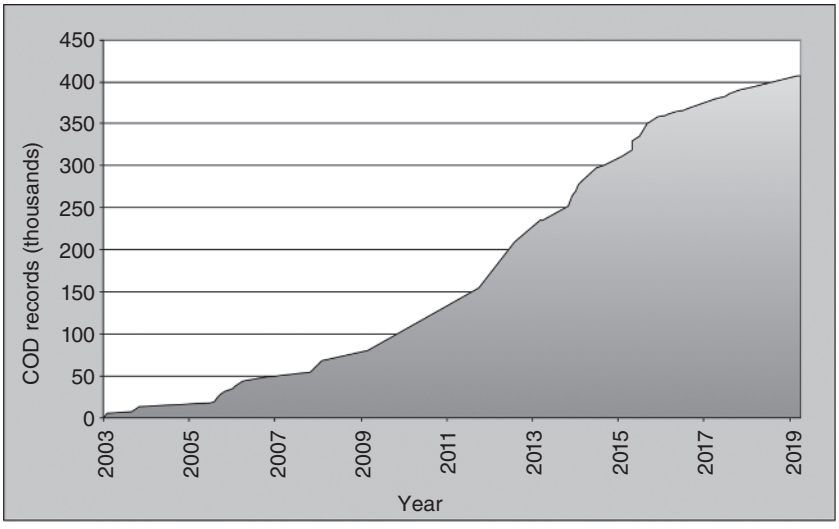

Figure 1.1 COD record number growth.

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

A list of scientific databases in the field of biosciences can be found in the Nucleic Acids Research [41], and crystallographic databases are listed by the IUCr (http://www.iucr.org/resources/data).

The COD incorporates a continuously increasing number of determined crystal structures, reaching >367 000 entries at the time of writing this chapter (Figure 1.1). The equivalent of COD for structures obtained from first-principle calculation and/or optimization is Theoretical Crystallography Open Database (TCOD), started in 2013, with consequently a more modest number of entries around 2000 (Figure 1.2). However such entries require long calculation times and one can expect larger increases in the years to come.

The COD was founded in February 2003 as a grassroot initiative – its establishment was proposed in a letter published at the Structure Determination by Powder Diffractometry (SDPD) 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

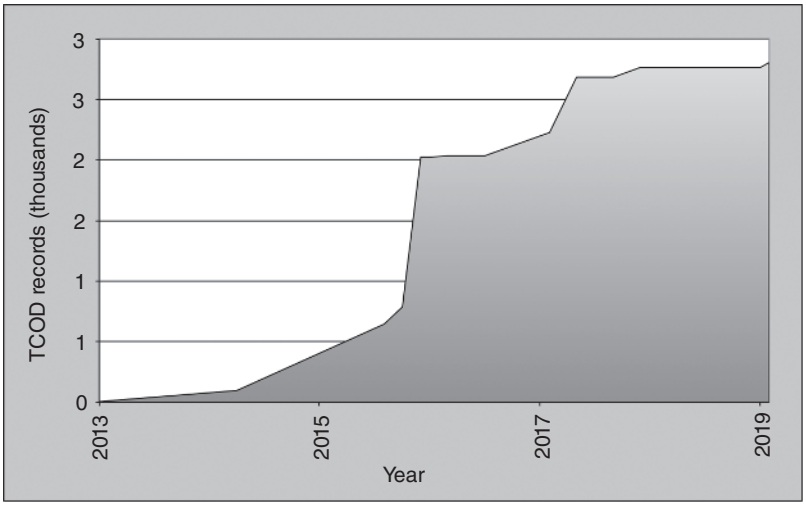

Figure 1.2 TCOD record number growth.

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 American Mineralogist Crystal Structure Database (AMCSD) [13] data set (all the crystal structures previously published in the *American Mineralogist* that were made freely accessible from the websites of the Mineralogical Society of America). The necessary MySQL/PHP scripts were written by Hareesh Rajan. In the meantime, Daniel Chateigner joined, and less than three 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 and 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 Laboratoire des Oxydes et Fluorures, Université du Maine (LdOF), 850 CIFs from the CRISMAT). The CIF2COD computer program (FORTRAN) was built on the basis of CIF2SX with the 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 the 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 website. The decision had to wait for the next IUCr Executive Committee meeting in August 2003. After four months, the number of entries in the COD reached 12 000, essentially by donations, from individuals or laboratories and the AMCSD.

Then came the sad news. Michael Berndt died on 30 June 2003, after a long, serious illness at the age of 39. Lachlan Cranswick went missing on 18 January 2010 at the age of 41 and his body was found later in the water on the Ottawa River, near Deep River. Despite the losses, the COD team continued to implement his plan and to work on the database. Five years after its founding, the COD passed a major milestone in 2008, by archiving the 50 000th entry. To attain completion, the COD should add much more than 40 000 new entries per year and also digitize older data that were published in print form. The required growth rate of the COD was attained in 2011 (Figure 1.1), when automated procedures for crystallographic data collection were implemented. Nevertheless, a lot of work remains to be done and the COD welcomes contributions from all crystallographers in order to accelerate its completion. During the past 10 years, the COD Advisory Board underwent some variations, departures, and new admissions, and the list of coauthors of this chapter reflects the current situation, presenting the main actors of the COD development until now.

1.3 Building COD

The COD collects all published crystal structures with small- to medium-sized unit cells. To facilitate this process, the CIF framework is employed. Currently, COD uses the CIF 1.1 [7] 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 is open access – all data are readily available on the Internet. COD data records are identified by stable Uniform Resource Identifiers (URIs) and accessible via the REpresentational State Transfer (REST) interface. The COD main page on the Web (http://www.crystallography.net/) states, "All data on this site have been placed in the public domain by the contributors," which we assume binding for COD Advisory Board, data maintainers, and contributors. All deposited data, unless embargoed by depositors for a fixed amount of time as a "prepublication deposition," are immediately available after the deposition 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. Since not all submitted files can be guaranteed to conform to the formal CIF definition [42], an error-correcting CIF parser [43] is employed. This ensures that all COD CIFs can be automatically parsed and supports unassisted COD data processing.

1.3.1 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 the PDB [44] and the Nucleic Acid Database (NDB) [18, 19].

As an experimental database, the COD collects structures determined by any experimental method. However, there are sister databases, the PCOD and the TCOD, which aim to collect predicted and theoretically determined structures, respectively (see Section 1.3.4 for a more comprehensive description).

COD structures may be refined using just X-ray data and first physical principles (using full-matrix least-squares methods), 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 combined with first principles using density functional theory (DFT)).

1.3.2 Data Sources

The COD acquires most of its structures (over 90%) from peer-reviewed scientific publications. The rest is deposited by authors either as personal communications or as prepublication 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 prepublication depositions are in addition checked for conformance to the IUCr data criteria.[1] The COD permits both manual deposition

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

by crystallographers using a Web interface (http://www.crystallography.net/cod/deposit) and an automated deposition using various Web-inspecting engines. Automated Web searches are conducted on journals that publish openly accessible crystallographic supplementary data. Data are also automatically extracted from open-access publications. Data from other crawlers, such as CrystalEye [45], and other open databases (e.g. AMCSD [13]) are incorporated into the COD on a regular basis, either using automated or semi-manual procedures. Such a strategy permits broad coverage of published structures with little resources required; it leverages the power of Internet automation while 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 paywalls 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 into open databases virtually impossible, not for technical but for purely organizational reasons. The barriers are not even related to intellectual property, since published data and facts of nature are not copyrightable. We thus urge everyone who sees virtue in open scientific data exchange and has benefited from open-access database to approach every publisher and ask them to provide underlying publication data for deposition to open-access databases or to deposit her or his crystallographic data directly to the COD.

1.3.3 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 [46] 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 Write Once Read Many (WORM) principle, which ensures that once data is written it is never changed irreversibly, becomes a necessity for scientific databases.

Collecting and preserving scientific data is an important endeavor, 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.

Table 1.2 Error classes routinely addressed by the COD maintainers.

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

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 and affects the data usability in varying degrees (see Table 1.2).

The initial step of data management in the COD is the detection and correction of syntactic errors. This kind of discrepancies is especially important since it renders the files unreadable and limits the possibility of any further data maintenance. Crystallographic structures in the COD are stored as CIFs, 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 [43]. 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. During the initial deposition semantic errors are recognized, automatically corrected, and reported to the depositor and in case an automatic correction is not possible, these errors are 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 workflow ensures that the overall semantic validity of the COD data set will only increase.

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 recognizing most 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 units other than Kelvin or data items used to describe the density of a 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 workflow 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 formulas, 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 modeling errors, unmarked disordered sites, or redundant atoms; several non-P1 structures, which 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, and typographical errors in coordinates and cell constants possible.

Not all structures, however, can be successfully corrected. To inform the user and enable the recognition of such entries in automated analyses, a warning or an error flag is added to the CIF manually. Currently there are around 20 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/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 the 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 with 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 of one another or deliberately suboptimal versions of some properly refined structure (the suboptimal structures are sometimes published to support the space group or refinement parameter choices). We must note, however, that not all duplicates are marked in the COD at the moment. 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 resulted in 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.

1.3.3.1 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, and at later stages, for peer review and for replications of conclusions reached. Curated databases, however, change over time, and databases like the 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 the 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 the COD data.

Currently, a Subversion server [47, 48] is used to register versions of the 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 a scientific database of the medium size, such as the COD. The text nature of

the CIF format makes them particularly well suited for tracking with revision control systems.

Since the introduction of Subversion, all COD data curation history is available, and any state of the database can be restored. As an additional advantage, Subversion also records movement of files in the file tree and rename operations, thus providing full data provenance of each COD CIF 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 Research Data Alliance (RDA) Recommendations for Data Citation [49, 50]. Therefore, the COD will implement a possibility to query every revision of COD database online (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.

1.3.3.2 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 as follows. A COD entry 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 represent the publication authors' *intent*. To that end, data in COD entries can be enhanced during the data curation; additional 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 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 orthogonalization matrices relating the used Cartesian frame and the crystal axes must be given). Another instance is the melting point of a crystalline material that must be given in Kelvin. 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 the COD context.

1.3.3.3 Quarterly Releases

The COD follows a continuous release policy – each commit to the COD database is immediately available on the Web and in the public Subversion repository. Each such commit introduces a new COD revision. The COD content is mostly updated on a daily basis, 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 the latest 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 are 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.

1.3.4 Sister Databases (PCOD, TCOD)

The growing need for COD-like databases for other than experimental structures has sparked the creation of two sister databases: the Predicted Crystallography Open Database (PCOD) for predicted structures and the Theoretical Crystal Structure Database (TCOD) for theoretically constructed structures. Predicted Crystallography Open Database (PCOD) (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 Geometrically Restrained INorganic Structure Prediction (GRINSP) software [22]. As the COD passed a major milestone by archiving the 50 000th entry in 2008, the 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 Zeolite Framework Solution (ZEFSA II) [51]. As a fork of the COD, the PCOD has inherited most of its features, such as stable unique data identifiers, data versioning, and Web and MySQL interfaces for searching. An automatic deposition service remains to be implemented in the PCOD.

The 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 crystal 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, the 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, a 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, the TCOD contains more than 2000 entries.

1.4 Use of COD

1.4.1 Data Search and Retrieval

Open-access Web resources pave the way for unprecedented applications that interconnect and reuse data hosted by many different organizations without the

need of coordination between them. Key elements for such cooperation are the interfaces for data access. Commonly used architectural style for both human- and machine-usable Web interfaces is REST, according to which RESTful interfaces are built [52], which use common HTTP requests to stable URLs for data retrieval.

1.4.1.1 Data Identification

Each entry in the COD consists of a CIF data block, listing the atomic positions of the crystal of interest, and an optional data block for diffraction data (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 1 000 000 to 9 999 999 – 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, which 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 with atomic positions, and .hkl for the diffraction data file). For example, files of entry 2002916 can be accessed via http://www.crystallography.net/cod/2002916.html, http://www.crystallography.net/cod/2002916.cif, and http://www.crystallography.net/cod/2002916.hkl.

1.4.1.2 Web Search Interface

Data can be searched on the Web using simple Web forms that use the COD MySQL database as a fast search index (Figure 1.3):

The COD server returns found results as a paginated HTML table (Figure 1.4). From this page, results can be downloaded in bulk as an archive. COD currently supports ZIP archives for downloaded data. The result table can be downloaded as a comma-separated value (CSV) file, and the list of selected structures can be obtained as a text file, either one COD number or one COD URI per line.

1.4.1.3 RESTful Interfaces

The same search interface can also be accessed programmatically using the COD RESTful API. The base URL for carrying out searches 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 Figure 1.5.

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

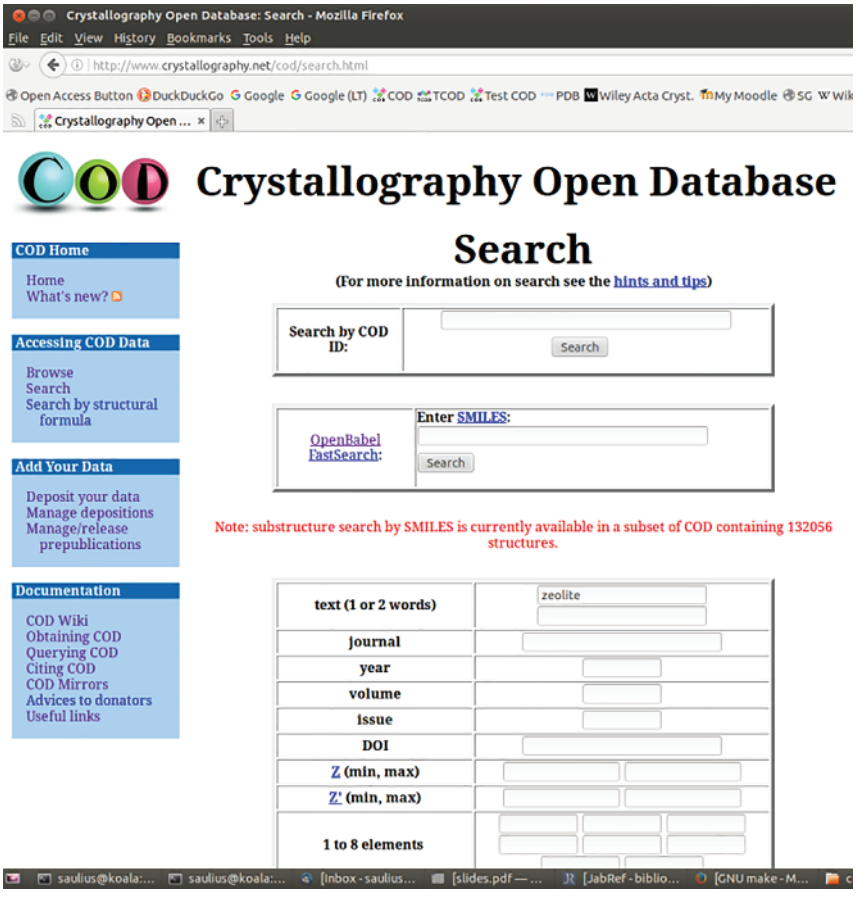

Figure 1.3 COD search Web interface form.

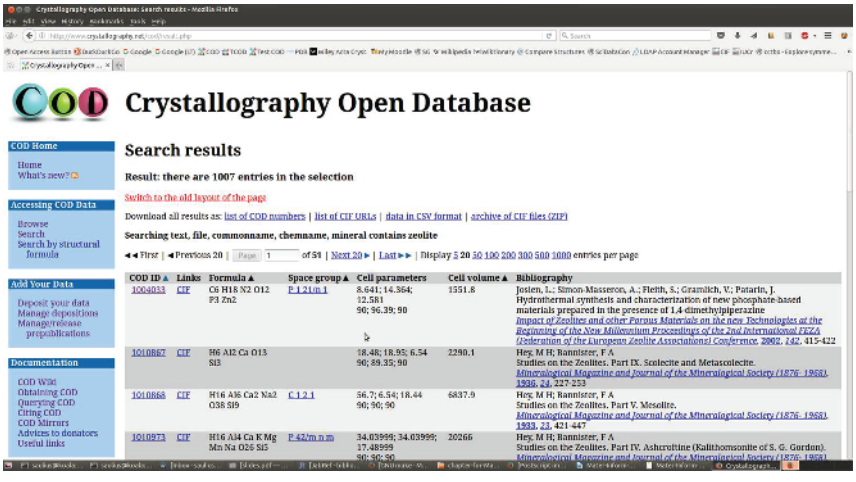

Figure 1.4 COD search result page, obtained as of 05 November 2016 from the query shown in Figure 1.3.

```
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
```

Figure 1.5 Example of the COD programmatic search interface.

- *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 the CIF format; however, additional output formats can be requested.

1.4.1.4 Output Formats

A combination of search parameters results in logical conjunction (OR operation). The output format can also be controlled using HTTP GET or POST parameter "format," with one of the following values: "html," "csv," "zip," and "json." In addition, "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 entries, 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.

1.4.1.5 Accessing COD Records

As presented in Section 1.4.1.1, each entry in the COD is identified by unique seven-digit number. COD presents the following URLs for access to the 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 identifier. An example of a query made using these identifiers from the Unix-style command line is shown in Figure 1.6.

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 must be provided via HTTP POST:

- *username*: depositor's username
- *password*: depositor's password
- *user_email*: depositor's e-mail address

```
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
```

Figure 1.6 Retrieving a specific COD structure using the stable COD URI identifier.

- *cif*: contents of to-be-deposited CIF
- *hkl*: contents of to-be-deposited diffraction data file (optional)
- *deposition_type*: type of deposition, either "published," "prepublication," or "personal"

1.4.1.6 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 cannot expose a full data query language such as SQL at the moment. To alleviate this limitation, the COD exposes a read-only version of the COD MySQL database for queries. When accessed as the "cod_reader" user, this database grants SELECT privilege to that user without asking a password to enable full use of the SQL query language. A special "sql.crystallography.net" host is dedicated for such queries. An example of such query using the Linux "mysql" command line client is illustrated in Figure 1.7.

The structure of the "data" view can be queried using standard SQL commands (Figure 1.8). 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/xml/documents/database-description/database-description.xml).

```
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 |
+---------+-------------------------+------+--------+--------+---------+---------+
```

Figure 1.7 Querying the COD MySQL database.

```
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    |       |
+-----------+---------------+------+-----+---------+-------+
```

Figure 1.8 Finding column definitions of the COD "data" view.

When querying data using SQL, the 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 the 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 the 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 Figure 1.9. This query method is recommended for the most material structure searches in the 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 Figure 1.9.

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 the local COD CIF tree is checked out in the directory ~/struct/cod/cif) is presented in Figures 1.10 and 1.11. Fetching coordinates from a copy on a local file system is of course much

```
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
```

Figure 1.9 Filtering out structures from the COD MySQL queries.

```
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 1.10 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 "'${HOME}/struct/cod/cif/'" \
                substr($0,1,1)"/"substr($0,2,2)"/"substr($0,4,2)"/" \
                $1".cif"}' \
    | xargs cat \
    > Si.cif
```

Figure 1.11 Preparing coordinates for an SQL query using a locally installed COD copy.

faster but requires preparation and maintenance of the up-to-date COD copy. In Section 1.4.1.8, we describe how to build such a COD copy.

1.4.1.7 Alternative Implementations of COD Search on the Web

Since the COD is openly accessible on the Web and all data are free for download, anyone can implement an alternative Web-based search engine for the COD, and indeed such sites have been implemented already. The oldest is probably the http://nanocrystallography.research.pdx.edu/ Web page that uses a subset of the COD for teaching purposes. The COD database access is provided by the STFC Chemical Database Service at Sci-Tech Daresbury (https://cds.dl.ac.uk/) on their page (https://cds.dl.ac.uk/cgi-bin/news/disp?COD). Another chemist-oriented

search tools existing at the moment are the MolView online molecular viewer by Herman Bergwerf (http://molview.org/) and the DataWarrior stand-alone Java program by Thomas Sander (http://www.openmolecules.org/), to mention just two mature open-source projects. Other similar endeavors exist on the Web as well.

In addition, the Web base abstractors of chemical information such as PubChem [53] and ChemSpider [54] 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 online COD servers. If a subset of the 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 use the full potential of open databases, something that is completely impossible with closed archives of data.

1.4.1.8 Installing a Local Copy of the COD

Since the 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 the 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 Figure 1.12; for other platforms, alternative SVN clients can be used (for example, TortoiseSVN (www.tortoisesvn.net/) for Windows).

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 Figure 1.13.

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

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

Figure 1.12 Obtaining (checking-out) a working copy of the COD data using the command line "svn" Subversion client.

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

Figure 1.13 Cloning COD data directory with GIT and GIT SVN.

"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 [55]. 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 synchronize a local tree with the COD database are provided in Figure 1.14.

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 synchronize 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 no knowledge about the Subversion repository transactions and cannot ensure that a complete revision is transferred. If an update of the 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.

To install the COD MySQL database, one has to obtain dumps of the COD MySQL tables and source them into a MySQL database in a MySQL server. Dumps can be either checked out from Subversion repository (commands shown in Figure 1.15) or downloaded and extracted from the COD quarterly releases (commands shown in Figure 1.16). One should note, however, that the first method is the most effective, since the latter requires downloading of a whole archive of a quarterly release (3–4 GB as of 2016, size depending on compression).

```
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/
```

Figure 1.14 Using the "rsync" program to download and update the COD file collection.

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

Figure 1.15 Checking out the COD MySQL dumps from the Subversion repository.

```
sh% wget http://www.crystallography.net/archives/cod-cifs-mysql.zip
sh% unzip cod-cifs-mysql.zip mysql/\*; cd mysql
```

Figure 1.16 Extracting the COD MySQL dumps from a ZIP archive of quarterly release.

Downloaded table schemata (*.sql) and tab-separated value lists (*.txt) have then to be loaded into an empty MySQL database. A script "cod-load-mysql-dump.sh," which is included in MySQL dumps, creates COD "data" table, provided that the user has root access to an empty database "cod" on a local machine. The same script can be used to update already existing MySQL database. However, the script should be used with attention, as it blanks the "data" table before loading in the data, so all local changes to the table between subsequent updates will be lost.

1.4.1.9 File System-Based Queries

When all COD files are available on a local disk, another kind of COD queries becomes possible, namely, queries of 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 online to an existing one.

Being ASCII-encoded text files, CIFs can be searched using the Unix "grep" and other tools. A query in Figure 1.17 will find all CIFs that contain the 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 that "diamond" in this case can be a name of a mineral, a name of a program, or something else).

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 organize 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 [42] 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: the COD employs an error-correcting parser from the "cod-tools" package [43] that has C, Perl, and Python bindings; other parsers have been proposed by various authors [56–58].

For quick composition of different processing tools, however, one can employ simple command line utilities to extract values. The "cod-tools" package [43] contains utility "cifvalue," which is written entirely in C and permits fast extraction

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

Figure 1.17 Search COD CIFs using "grep." Options of this command are supported by the GNU "grep" utility on the Ubuntu 12.04 operating system or higher.

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

Figure 1.18 Use of "find," "xargs" and "cifvalues" from the "cod-tools" package to extract requested data from CIFs.

of requested CIF values and their printout in a space-separated-value form that is then easily processed by "awk," "perl," and "R," in most spreadsheet programs and a multitude of other tools. An example in Figure 1.18 shows how to use "cifvalue," in conjunction with the aforementioned "find" and "xargs" programs, to extract molecular weight, unit cell volume, and melting point data from the COD collection.

1.4.1.10 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 provides the means of recognizing structures unfit for conventional use. 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 planning 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 the 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 to be done by the depositor; thus the COD cannot guarantee that all mineral structures in the database are marked as such.

As described in the Section 1.3.3, CIFs of entries with issues are marked with special data items to be recognized 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 their number 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 meant to use the labels for this purpose or if he or she 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, respectively, would be recognized 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 which are mixtures of two or more different chemical types. For example, the grunerite structure in the COD entry 9000000 contains four iron–magnesium sites, which 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 [43].

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, the 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]).

1.4.2 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 [20], 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.
- Prepublication 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 its maintainers. Since 2010 all depositions have been directed to the novel interface, thus saving many man-hours of effort.

1.5 Applications

1.5.1 Material Identification

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 the 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.

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, etc.) 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 software 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, ignoring the other phases (inorganics, metal–organics, etc.), except if the user can afford all databases and corresponding software.

Another approach resolving the mentioned drawbacks of classical databases is clearly provided using the COD. Since the COD records all structures independently of their "classification" as inorganics or other classes, the search–match results extend to a wider range 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. Additionally, 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 [90].

The COD and its sister databases are 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 the COD in the 2000s in its search–match software, rapidly followed by Panalytical (Highscore+ software), Bruker (Eva), and Rigaku (PDXL). More recently an employee at the 3D Systems Corp. used it for the 3D printing of crystallographic models, and Kagaku Benran incorporated the COD in his Crystallography Handbook.

1.5.2 Applications for the Mining Industry

The usefulness of the COD for mineral identification proved very useful for practical applications in mining. In the SOLSA[2] (Sonic Drilling coupled with Automated Mineralogy and chemistry On-Line-On-Mine-Real-Time)[3] project

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

that started in 2016, the COD is used as an essential data provider for identifying minerals for characterization of the drill cores. The COD is also planned as a vehicle of the subsequent data dissemination, storing results of crystallographic investigations of drill cores. All properties of the COD are essential here – open-access regime permits efficient distribution and fast access to data; the well-established CIF framework provides a sound foundation for describing measurement results, and the RESTful interface enables easy integration. The COD codebase has also been reused to launch the Raman Open Database (ROD) in order to properly store Raman spectroscopy measurements as well as to interlink them with the crystal structures in the COD [59]. It is anticipated that other results of the SOLSA project will be made openly available to the community after the project is completed.

1.5.3 Extracting Chemical Information

Many of the potential users of the COD are chemists so they will be more interested in the chemical features of any crystallized compound than in the purely crystallographic facts. For organic and metal–organic chemists, the chemical features of the compound are mostly defined by the statement of how atoms are directly bonded 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 CIFs uploaded to COD and it needs to be deduced from atom coordinates and/or the bond list (if present). 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 [60] (http://www.daylight.com/dayhtml/doc/theory/theory.smiles.html) and an open specification established afterward (http://opensmiles.org); both are essentially identical). This format represents 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 writable/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) by 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 (as 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, boranes, etc.). 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 is however far from being a trivial task. We are using the Open Babel toolbox [61] (http://openbabel.org) which, in principle, has the ability for performing the CIF to SMILES conversion, but the result is not optimal 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 and hence does not handle properly chemical species placed on symmetry elements of the crystal, nor does it considers the possible disorder. To circumvent these problems, algorithms and corresponding software have been developed by the COD maintainers [62].

But even if we have a set of atoms chemically representing 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 convention. 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 force every atom to have its usual valence. The number of bonds that an atom can form is also limited 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 above-mentioned reasons, the obtained crude SMILES usually represent accurately organic compounds (easily recognizable by the absence of square brackets) that may be accepted without further treatment, but not metal–organic compounds, for which one very frequently finds 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 or unreleased software and by not well-disclosed algorithms that are used by commercial databases [63].

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 already been used to set around 35 000 links between the COD and the open chemical database ChemSpider (http://www.chemspider.com) [54], 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 are the main way of defining families of compounds. The COD website implements such searches

by allowing the user to introduce the fragment also in SMILES format and then use the 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 the COD website using graphical interfaces written in the JavaScript [64] http://www.molinspiration.com/jme/ language. 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 own 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 (https://github.com/timvdm/OpenSMARTS; http://www.daylight.com/dayhtml/doc/theory/theory.smarts.html), which allows for more versatile and sophisticated searches than SMILES.

1.5.4 Property Search

Modern methods of computational chemistry can greatly reduce the efforts in fields such as the 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 [65]. A group of researchers has embarked on the screening for crystal structures with periodic layered compounds in order to identify novel graphene-like compounds in both the COD and the ICSD [66].

1.5.5 Geometry Statistics

In order to simplify and encourage similar research on the basis of the COD, we are developing a database for the geometry of the 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 [67, 68]. We have chosen a "fuzzy" descriptor as a balance between too strict matching, which 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 data set and the describing

models, we have launched a Web interface. Currently, browsing is implemented using atom descriptors as mentioned earlier.

One of the possible uses of our geometry database is the detection of common geometric features. The semi-automated search for artifacts and outliers in the crystal structures is another possible use. Furthermore, derived statistical distributions from our database could be used to generate force fields in modeling, as well as constraints or restraints for the refinement of crystal structures. This particular approach is being used to compile a dictionary of constraints for macromolecular structure refinement using the REFMAC5 refinement program [69, 70].

1.5.6 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 [71]. To aid the field of atomistic simulations, Pizzi et al. have developed the AiiDA framework [72], based on the Open Provenance Model [73, 74]. AiiDA can 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 the TCOD. The current pipeline allows seamless importing of experimental data from the COD to AiiDA for further atomistic simulations while 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.

1.5.7 Applications in College Education and Complementing Outreach Activities

Crystallographic open-access databases have been built from 2004 onward for educational purposes at Portland State University. 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 College in Minnesota, United States, has been adopted for this purpose [75].

In recent years, we augmented our educational activities with 3D-printed crystallographic models [76, 77]. The key to these activities was a Windows executable program by Werner Kaminsky [78] 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 [78] and longitudinal representation surfaces for anisotropic crystal physical properties [78].

While the CIF dictionaries contain provisions to encode crystal morphologies in *.cif files directly so that it can be read into Werner's program [79], the

developers of the Material Properties Open Database [10] 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 [80]. Selected 3D print files and CIF-encoded crystal morphologies are available for download at the above-mentioned educational project of Portland State University.

1.6 Perspectives

1.6.1 Historic Structures

As of August 2016, most of the structures in the COD are published in the "CIF era" (1990s onward), with the contribution of older structures equal to only 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 digitalize and deposit them as CIFs. Therefore, we have produced a few dozens of such entries manually, but the laborious nature of such task prevents the conversion from attaining speed. Nevertheless, the collection of historic structures can be speeded 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.

1.6.2 Theoretical Data in (T)COD

Over the last 25 years, the CIF format has become the standard for the reporting and archiving of the results of experimental crystal structure solutions. It was adopted and used by the crystallographic journals as well as the structural databases. New CIF dictionaries are being developed to define ontologies in such fields as macromolecular crystallography [6], powder diffraction [81], and electron density studies [82]. However, much effort is still needed to consolidate the knowledge in the field of theoretical materials science, which is expanding rapidly currently. Nevertheless, there are a few disjoint attempts, namely, European Theoretical Spectroscopy Facility (ETSF) ([83, 84], and NoMaD [85]. 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, the TCOD puts emphasis on the provenance of the results and reproducibility by devising a special dictionary for related metadata – cif_tcod [86]. The TCOD, accompanied with a huge collection of experimental structures in the COD [21], opens an immediate potential for the cross-validation of experimental and theoretical data.

1.6.3 Conclusion

The 16 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. In its history the COD was online most

of the time, except for a very few short technical glitches. Its volume grew constantly over time, and it enjoys an 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 not only 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 [87–89].

Acknowledgments

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 no. 689868), and the Conseil Régional de Normandie (COMBIX project, 2013–2014, Chair of Excellence of LL). We thank Dr. Peter Murray-Rust for providing information about CrystalEye.

References

1 Authors of Wikipedia (2016). Hipparchus. https://en.wikipedia.org/wiki/Hipparchus (accessed 16 October 2016).
2 Annis, J., Bakken, J., Holmgren, D., et al. (1999). The Sloan Digital Sky Survey data acquisition system, and early results. *Real Time Conference, 1999. Santa Fe 1999. 11th IEEE NPSS 14–18 June 1999*. IEEE. DOI: https://doi.org/10.1109/RTCON.1999.842551
3 Hewett J. (2006). LHC factoids. http://blogs.discovermagazine.com/cosmicvariance/2006/09/27/lhc-factoids/ (accessed 16 October 2016).
4 PPARC (2006). 'Maiden Flight' for LHC computing grid breaks gigabyte-per-second barrier. http://phys.org/news/2006-02-maiden-flight-lhc-grid-gigabyte-per-second.html (accessed 16 October 2016).
5 Hahn, T. (ed.) (2006). *International Tables for Crystallography. Vol. A: Space-group Symmetry*. Dordrecht, The Netherlands: Published for the International Union of Crystallography by Springer https://doi.org/10.1107/97809553602060000100.
6 Fitzgerald, P.M.D., Westbrook, J.D., Bourne, P.E. et al. (2006). Macromolecular dictionary (mmCIF). In: *International Tables for Crystallography* (ed. S.R. Hall and B. McMahon). International Union of Crystallography https://doi.org/10.1107/97809553602060000745.
7 Hall, S.R., Allen, F.H., and Brown, I.D. (1991). The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Crystallogr. Sect. A* 47: 655–685. https://doi.org/10.1107/S010876739101067X.
8 Bernstein, H.J., Bollinger, J.C., Brown, I.D. et al. (2016). Specification of the Crystallographic Information File format, version 2.0. *J. Appl. Crystallogr.* 49 (1): 277–284. https://doi.org/10.1107/s1600576715021871.

9 Faber, J. and Fawcett, T. (2002). The Powder Diffraction File: present and future. *Acta Crystallogr. Sect. B* 58 (3 Part 1): 325–332. https://doi.org/10.1107/S0108768102003312.

10 Pepponi, G., Gražulis, S., and Chateigner, D. (2012). MPOD: A Material Property Open Database linked to structural information. *Nucl. Instrum. Methods Phys. Res., Sect. B* 284: 10–14. https://doi.org/10.1016/j.nimb.2011.08.070.

11 Lafuente, B., Downs, R.T., Yang, H., and Stone, N. (2015). The power of databases: the RRUFF project. In: *Highlights in Mineralogical Crystallography* (ed. T. Armbruster and R.M. Danisi), 1–29. W. De Gruyter.

12 Downs, R.T. and Hall-Wallace, M. (2003). The American Mineralogist crystal structure database. *Am. Mineral.* 88: 247–250.

13 Rajan, H., Uchida, H., Bryan, D. et al. (2006). Building the American Mineralogist crystal structure database: a recipe for construction of a small Internet database. In: *Geoinformatics: Data to Knowledge* (ed. A. Sinha). Geological Society of America https://doi.org/10.1130/2006.2397(06).

14 Baerlocher, C., McCusker, L., and Olson, D. (2007). *Atlas of Zeolite Framework Types*, 6th revised edition. Amsterdam – London – New York – Oxford – Paris – Shannon – Tokyo: Elsevier.

15 Aroyo, M.I., Perez-Mato, J.M., Orobengoa, D. et al. (2011). Crystallography online: Bilbao Crystallographic Server. *Bulg. Chem. Commun.* 43 (2): 183–197.

16 Aroyo, M.I., Perez-Mato, J.M., Capillas, C. et al. (2006). Bilbao Crystallographic Server: I. Databases and crystallographic computing programs. *Z. Kristallogr. – Cryst. Mater.* 221 (1): 15–27. https://doi.org/10.1524/zkri.2006.221.1.15.

17 Perez-Mato, J., Gallego, S., Tasci, E. et al. (2015). Symmetry-based computational tools for magnetic crystallography. *Ann. Rev. Mater. Res.* 45 (1): 217–248. https://doi.org/10.1146/annurev-matsci-070214-021008.

18 Berman, H.M., Olson, W.K., Beveridge, D.L. et al. (1992). The nucleic acid database: a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.* 63: 751–759. https://doi.org/10.1016/S0006-3495(92)81649-1.

19 Coimbatore Narayanan, B., Westbrook, J., Ghosh, S. et al. (2014). The nucleic acid database: new features and capabilities. *Nucleic Acids Res.* 42: D114–D122. https://doi.org/10.1093/nar/gkt980.

20 Gražulis, S., Chateigner, D., Downs, R.T. et al. (2009). Crystallography Open Database – an open-access collection of crystal structures. *J. Appl. Crystallogr.* 42: 726–729. https://doi.org/10.1107/S0021889809016690.

21 Gražulis, S., Daškevič, A., Merkys, A. et al. (2012). Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. *Nucleic Acids Res.* 40: D420–D427. https://doi.org/10.1093/nar/gkr900.

22 Le Bail, A. (2005). Inorganic structure prediction with it GRINSP. *J. Appl. Crystallogr.* 38: 389–395. https://doi.org/10.1107/S0021889805002384.

23 Chateigner, D., Grazulis, S., Pérez, O., et al. (2015). COD, PCOD, TCOD, MPOD … open structure and property databases. http://www.ecole.ensicaen.fr/~chateign/danielc/abstracts/Chateigner_abstract_JNCO2013.pdf (accessed 19 April 2019).

24 Groom, C.R., Bruno, I.J., Lightfoot, M.P., and Ward, S.C. (2016). The Cambridge Structural Database. *Acta Crystallogr. Sect. B* 72 (2): 171–179. https://doi.org/10.1107/S2052520616003954.

25 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 Crystallogr. Sect. B* 58: 364–369. https://doi.org/10.1107/S0108768102006948.

26 White, P.S., Rodgers, J.R., and Le Page, Y. (2002). CRYSTMET: a database of the structures and powder patterns of metals and intermetallics. *Acta Crystallogr. Sect. B* 58: 343–348. https://doi.org/10.1107/S0108768102002902.

27 Villars, P., Onodera, N., and Iwata, S. (1998). The Linus Pauling file (LPF) and its application to materials design. *J. Alloys Compd.* 279: 1–7. https://doi.org/10.1016/S0925-8388(98)00605-7.

28 Villars, P., Berndt, M., Brandenburg, K. et al. (2004). The Pauling File, Binaries Edition. *J. Alloys Compd.* 367 (1–2): 293–297. https://doi.org/10.1016/j.jallcom.2003.08.058.

29 Protein Data Bank (1971). Protein Data Bank. *Nat. New Biol.* 233: 22–23. https://doi.org/10.1038/newbio233223b0.

30 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. https://doi.org/10.1016/j.str.2012.01.010.

31 Gilliland, G.L., Tung, M., and Ladner, J.E. (2002). The Biological Macromolecule Crystallization Database: crystallization procedures and strategies. *Acta Crystallogr. Sect. D* 58 (6 Part 1): 916–920. https://doi.org/10.1107/S0907444902006686.

32 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. *J. Chem. Inf. Model.* 51: 3029. https://doi.org/10.1021/ci200460z.

33 Sadowski, P. and Baldi, P. (2013). Small-molecule 3D structure prediction using open crystallography data. *J. Chem. Inf. Model.* 53: 3127–3130. https://doi.org/10.1021/ci4005282.

34 Bruno, I. and Groom, C. (2014). A crystallographic perspective on sharing data and knowledge. *J. Comput. Aided Mol. Des.* 28 (10): 1015–1022. https://doi.org/10.1007/s10822-014-9780-9.

35 Eger, T., Scheufen, M. and Meierrieks D. (2013). The determinants of Open Access Publishing: survey evidence from Germany. http://ssrn.com/abstract=2232675 (accessed 19 April 2019).

36 Eysenbach, G. (2006). Citation advantage of open access articles. *PLoS Biol.* 4 (5): e157. https://doi.org/10.1371/journal.pbio.0040157.

37 Harnad, S. and Brody, T. (2004). Comparing the impact of Open Access (OA) vs. Non-OA articles in the same journals. *D-Lib Magaz.* 10 (6): https://doi.org/10.1045/june2004-harnad.

38 Harnad, S., Brody, T., Vallières, F. et al. (2008). The access/impact problem and the green and gold roads to open access: an update. *Serials Rev.* 34 (1): 36–40. https://doi.org/10.1080/00987913.2008.10765150.

39 Zucker, L.G., Darby, M.R., Furner, J., et al. (2006). Minerva unbound: knowledge stocks, knowledge flows and new knowledge production. NBER Working Paper Series. http://www.nber.org/papers/w12669 (accessed 19 April 2019).

40 Piwowar, H.A. and Vision, T.J. (2013). Data reuse and the open data citation advantage. *PeerJ* 1: e175. https://doi.org/10.7717/peerj.175.

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

42 IUCr (2016). CIF version 1.1 working specification. http://www.iucr.org/resources/cif/spec/version1.1 (accessed 06 November 2016, 14:55 EET).

43 Merkys, A., Vaitkus, A., Butkus, J. et al. (2016). COD::CIF::Parser: an error-correcting CIF parser for the Perl language. *J. Appl. Crystallogr.* 49 (1): 292–301. https://doi.org/10.1107/S1600576715022396.

44 Berman, H.M., Westbrook, J., Feng, Z. et al. (2000). The Protein Data Bank. *Nucleic Acids Res.* 28: 235–242. https://doi.org/10.1093/nar/28.1.235.

45 Day, N., Downing, J., Adams, S. et al. (2012). CrystalEye: automated aggregation, semantification and dissemination of the world's open crystallographic data. *J. Appl. Crystallogr.* 45: 316–323. https://doi.org/10.1107/S0021889812006462.

46 Dalle O. (2012). On reproducibility and traceability of simulations. *Proceedings of the Winter Simulation Conference. Winter Simulation Conference, 9–12 December 2012.* IEEE.

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

48 Collins-Sussman B., Fitzpatrick B.W. and Pilato C.M. (2011). Version control with subversion. http://svnbook.red-bean.com/ (accessed 19 April 2019).

49 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).* RDA https://rdalliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf (accessed 19 April 2019).

50 Rauber, A., Asmi, A., van Uytvanck, D. and Pröll, S. (2016). Identification of reproducible subsets for data citation, sharing and re-use. https://www.ieeetcdl.org/Bulletin/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf (accessed 19 April 2019).

51 Falcioni, M. and Deem, M.W. (1999). A biased Monte Carlo scheme for zeolite structure solution. *J. Chem. Phys.* 110 (3): 1754–1766. https://doi.org/10.1063/1.477812.

52 Fielding, R.T. (2000). Architectural styles and the design of network-based software architectures. University of California, Irvine. https://www.ics.uci.edu/~fielding/pubs/dissertation/top.htm (accessed 19 April 2019).

53 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: *Annual Reports in Computational Chemistry* (ed. R.A. Wheeler and D.C. Spellmeyer). Oxford, UK: Elsevier https://doi.org/10.1016/S1574-1400(08)00012-1.

54 Pence, H.E. and Williams, A. (2010). ChemSpider: an online chemical information resource. *Chem. Educ. Today* 87: 1123–1124. https://doi.org/10.1021/ed100697w.

55 Davison W. (2015). Rsync. http://samba.anu.edu.au/rsync/ (accessed 06 November 2016, 13:42 EET).

56 Gildea, R.J., Bourhis, L.J., Dolomanov, O.V. et al. (2011). iotbx.cif: a comprehensive CIF toolbox. *J. Appl. Crystallogr.* 44: 1259–1263. https://doi.org/10.1107/S0021889811041161.

57 Hester, J.R. (2006). A validating CIF parser: PyCIFRW. *J. Appl. Crystallogr.* 39: 621–625. https://doi.org/10.1107/S0021889806015627.

58 Todorov, G. and Bernstein, H.J. (2008). it VCIF2: extended CIF validation software. *J. Appl. Crystallogr.* 41: 808–810. https://doi.org/10.1107/S002188980801385X.

59 El Mendili, Y., Vaitkus, A., Merkys, A. et al. (2019). Raman Open Database: first interconnected Raman-XRD open-access resource for material identification. *J. Appl. Crystallogr.* 52: 618–625. https://doi.org/10.1107/S1600576719004229.

60 Funatsu, K., Miyabayashi, N., and Sasaki, S. (1988). Further development of structure generation in the automated structure elucidation system CHEMICS. *J. Chem. Inf. Model.* 28 (1): 18–28. https://doi.org/10.1021/ci00057a003.

61 O'Boyle, N.M., Banck, M., James, C.A. et al. (2011). Open Babel: an open chemical toolbox. *J. Cheminf.* 3: 3. https://doi.org/10.1186/1758-2946-3-33.

62 Gražulis, S., Merkys, A., Vaitkus, A., and Okulič-Kazarinas, M. (2015). Computing stoichiometric molecular composition from crystal structures. *J. Appl. Crystallogr.* 48: 85–91. https://doi.org/10.1107/S1600576714025904.

63 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. https://doi.org/10.1107/s0108768111024608.

64 Bienfait, B. and Ertl, P. (2013). JSME: a free molecule editor in JavaScript. *J. Cheminf.* 5: 2–4. https://doi.org/10.1186/1758-2946-5-24.

65 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. https://doi.org/10.3390/cryst5040617.

66 Mounet, N., Gibertini, M., Schwaller, P., et al. (2016). High-throughput prediction of two-dimensional materials. https://doi.org/10.1038/s41565-017-0035-5.

67 Long, F., Nicholls, R.A., Emsley, P., et al. (2016). ACEDRG: a stereo-chemical description generator for ligands. https://doi.org/10.1107/s2059798317000067.

68 Long, F., Nicholls, R.A., Emsley, P., et al. (2016). Validation and extraction of stereochemical information from small molecular databases. https://doi.org/10.1107/s2059798317000079.

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

70 Vagin, A.A., Steiner, R.A., Lebedev, A.A. et al. (2004). it REFMAC5 dictionary: organization of prior chemical knowledge and guidelines for its use. *Acta Crystallogr. Sect. D* 60 (12): 2184–2195. https://doi.org/10.1107/S0907444904023510.

71 Mesirov, J.P. (2010). Computer science. Accessible reproducible research. *Science* (New York, NY) 327: 415–416. https://doi.org/10.1126/science.1179653.

72 Pizzi, G., Cepellotti, A., Sabatini, R. et al. (2016). AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sci.* 111: 218–230. https://doi.org/10.1016/j.commatsci.2015.09.013.

73 Moreau, L., Freire, J., Futrelle, J. et al. (2008). The open provenance model: an overview. In: *Provenance and Annotation of Data and Processes* (ed. J. Freire, D. Koop and L. Moreau). Berlin, Heidelberg: Springer https://doi.org/10.1007/978-3-540-89965-5_31.

74 Moreau, L., Freire, J., Futrelle, J. et al. (2007). *The Open Provenance Model*. University of Southampton http://eprints.soton.ac.uk/264979/ (accessed 19 April 2019).

75 Moeck, P., Čertík, O., Upreti, G. et al. (2005). Crystal structure visualizations in three dimensions with database support. *MRS Proc.* 909E: 3.5.1–3.5.6. https://doi.org/10.1557/PROC-0909-PP03-05.

76 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. Educ.* 36: 77–96.

77 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. https://doi.org/10.1002/crat.201400469.

78 Kaminsky, W., Snyder, T., Stone-Sundberg, J., and Moeck, P. (2015). 3D printing of representation surfaces from tensor data of KH_2PO_4 and low-quartz utilizing the WinTensor software. *Z. Kristallogr.* 230: 651–656. https://doi.org/10.1515/zkri-2014-1826.

79 Kaminsky, W. (2007). From CIF to virtual morphology: new aspects of predicting crystal shapes as part of the WinXMorph program. *J. Appl. Crystallogr.* 40: 382–385. https://doi.org/10.1107/S0021889807003986.

80 Fuentes-Cobas, L., Chateigner, D., Pepponi, G. et al. (2014). Implementing graphic outputs for the Material Properties Open Database (MPOD). *Acta Cryst.* 70: C1039. https://doi.org/10.1107/S2053273314089608.

81 Toby, B.H., Von Dreele, R.B., and Larson, A.C. (2003). CIF applications. XIV. Reporting of Rietveld results using pdCIF: GSAS2CIF. *J. Appl. Crystallogr.* 36: 1290–1294.

82 Mallinson, P.R. and Brown, I.D. (2006). Classification and use of electron density data. In: *International Tables for Crystallography*, vol. G (ed. S.R. Hall and B. McMahon). International Union of Crystallography. https://doi.org/10.1107/97809553602060000107.

83 Caliste, D., Pouillon, Y., Verstraete, M. et al. (2008). Sharing electronic structure and crystallographic data with ETSFIO. *Comput. Phys. Commun.* 179: 748–758. https://doi.org/10.1016/j.cpc.2008.05.007.

84 Gonze X., Almbladh C.-O., Cucca A., et al. (2008). Specification of file formats for ETSF Specification version 3.3. Second revision for this version (SpecFF ETSF3.3). European Theoretical Spectroscopy Facility. http://www.etsf.eu/system/files/SpecFFETSF_v3.3.pdf (accessed 19 April 2019).

85 Mohamed F.R. (2016). Nomad meta info. https://gitlab.rzg.mpg.de/nomad-lab/nomad-meta-info/wikis/home (accessed 18 February 2016).

86 Gražulis S. (2016). TCOD mailing list. http://lists.crystallography.net/cgi-bin/mailman/listinfo/tcod (accessed 13 April 2016).

87 Gražulis, S., Sarjeant, A.A., Moeck, P. et al. (2015). Crystallographic education in the 21st century. *J. Appl. Crystallogr.* 48 (6): 1964–1975. https://doi.org/10.1107/S1600576715016830.

88 Kaminsky, W., Snyder, T., Stone-Sundberg, J., and Moeck, P. (2014). One-click preparation of 3D print files (*.stl, *.wrl) from *.cif (crystallographic information framework) data using Cif2VRML. *Powder Diffr.* 29: S42–S47. https://doi.org/10.1017/S0885715614001092.

89 Moeck, P., Kaminsky, W., Fuentes-Cobas, L. et al. (2016). 3D printed models of materials tensor representations and the crystal morphology of alpha quartz. *Symmetry: Cult. Sci.* 27: 319–330.

90 Lutterotti, L., Pillière, H., Fontugne, C., Boullay, P., and Chateigner, D. (2019). Full-profile search–match by the Rietveld method. *J. Appl. Crystallogr.* 52: 587–598. https://doi.org/10.1107/S160057671900342X.