DATA SCIENCE JOURNAL

Bruno, I et al 2017 Crystallography and Databases. *Data Science Journal,* 16: 38, pp. 1–17, DOI: https://doi.org/10.5334/dsj-2017-038

REVIEW

Crystallography and Databases

Ian Bruno[1], Saulius Gražulis[2], John R Helliwell[3], Soorya N Kabekkodu[4], Brian McMahon[5] and John Westbrook[6]

[1] Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK

[2] Vilnius University Institute of Biotechnology, Sauletekio al. 7, LT-10257 Vilnius, LT

[3] School of Chemistry, University of Manchester, M13 9PL, UK

[4] International Centre for Diffraction Data, 12 Campus Boulevard, Newtown Square, PA 19073, US

[5] International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK

[6] Department of Chemistry and Chemical Biology, Research Collaboratory for Structural Bioinformatics Protein Data Bank, Center for Integrative Proteomics Research, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, Piscataway, NJ 08854, US

Corresponding author: Brian McMahon (bm@iucr.org)

Crystallographic databases have existed as electronic resources for over 50 years, and have provided comprehensive archives of crystal structures of inorganic, organic, metal–organic and biological macromolecular compounds of immense value to a wide range of structural sciences. They thus serve a variety of scientific disciplines, but are all driven by considerations of accuracy, precise characterization, and potential for search, analysis and reuse. They also serve a variety of end-users in academia and industry, and have evolved through different funding and licensing models. The diversity of their operational mechanisms combined with their undisputed value as scientific research tools gives rise to a rich ecosystem. A session at SciDataCon2016 gave an overview of the largest extant crystallographic databases and their current activities and plans for the future. This review summarizes these presentations and considers them alongside other players in the field, demonstrating their variety, versatility and focus on quality and usefulness.

Keywords: crystallography; structural science; curated databases; data exchange standard; crystallographic information file

1 Introduction

1.1 Rationale

Crystallography is 'the branch of science devoted to the study of molecular and crystalline structure and properties, with far-reaching applications in mineralogy, chemistry, physics, mathematics, biology and materials science' (Online Dictionary of Crystallography, 2017). One of its main applications is the precise determination of atomic-scale structure of chemical compounds. For molecular compounds this provides unparalleled information about composition, configuration and reactivity, and has been successfully used for both 'small' chemical species (ionic compounds, metals, organic molecules and metal–organic complexes) and biological macromolecular complexes (proteins, nucleic acids, ribosomal subunits, viruses). For ionic or metallic compounds, great insight is gained into the physical and chemical properties of materials in the solid state. Since such information is essential to many other applied sciences, it is perhaps not surprising that carefully curated archives of such data have been in existence for many decades – since the 1940s as print compilations (journal articles, collected structure reports, diffraction patterns), and subsequently as electronic databases, first established in the 1960s. These have been of great importance in academic research and industrial applications, and continue to underpin new discoveries in fields as diverse as battery technology, drug design and genomics.

Hence, for over 70 years the International Union of Crystallography (IUCr) and the crystallographic community have been concerned with data archiving and data reuse. A session at SciDataCon2016 provided

the opportunity to present an account of several of the most significant extant crystallographic databases, which we have collected together in Sections 2–4 below. We also provide some information about other crystallographic databases which we could not accommodate within the programming constraints of the conference. Section 1.2 provides an overview of the databases discussed, demonstrating the great variety of different crystallographic science subject areas and specialisms that they cover.

Throughout, we use the term *curated database* to refer to a collection of data sets, generally of a particular type of compound, that have been assessed against various criteria for completeness and correctness, that are intended for long-term archival access, that are well characterized by metadata permitting search and retrieval, and that in most cases are accompanied by software tools allowing for scientific analysis of properties and materials, of which individual data sets are representative. We distinguish these from data repositories, which may contain heterogeneous data described by limited metadata with little scope for analysis across a complete collection.

1.2 Overview

We were motivated to suggest this session for SciDataCon2016 to exemplify how different models of database curation can coexist and be of scientific value. In crystallography, which is inherently multidisciplinary, the databases exhibit variety both in implementation and in their business operations. These have developed around varying community practices and differing technical challenges and requirements. They have also been established at different points in the last half century, during which different approaches to the funding of scientific research have influenced their revenue models. These factors contribute towards a more heterogeneous ecosystem of database services than will be found in many other disciplines.

Table 1 summarizes the characteristics of the main curated crystallographic databases active (to our knowledge) in early 2017. A more complete list of extant databases is maintained by the IUCr (International Union of Crystallography 2017). More information about the historical development of the databases, their approaches to data validation and curation, and examples of their scientific applications may be found in the publication *Crystallographic Databases* (Allen *et al.* 1987) and in special issues of the *Journal of Research of the National Institute of Standards and Technology* (Karen & Mighell 1996) and *Acta Crystallographica Sections B* and *D* (Allen & Glusker 2002).

2 Chemical crystallography

2.1 The Crystallography Open Database – new perspectives

2.1.1 History

Today's connected world crucially depends on the open availability of data on the Internet. The great success of the Protein Data Bank (Berman *et al.* 2003, 2012) and open sequence databases like UniProt (The UniProt Consortium 2015) demonstrates the power of data sharing when it is unhindered by paywalls and copy restrictions.

The Crystallography Open Database (COD) builds on the experience of such open databases and harnesses the power of the community to build an openly-available chemical crystallography database on the net. The COD ingests data in the standard CIF format maintained by the IUCr (Hall *et al.* 1991, Bernstein *et al.* 2016), validates it according to IUCr dictionaries and quality criteria, and offers consolidated data to COD users, again in the standard CIF format. Currently, the COD contains over 376,000 records, spanning the years 1915 to the present.

Numerous people have contributed to the COD since its establishment, and it now contains data from several databases such as the *American Mineralogist* Crystal Structure Database (Rajan *et al.* 2006), CrystalEye (Day *et al.* 2012) and the Predicted Structures Open Database, PCOD (Le Bail 2005). Most structures that have been published in electronic format are represented in the COD, provided they were available on the Internet freely or were donated by authors or their institutions. Some prominent structures published in paper form were also digitized and included.

2.1.2 Scope

The COD stores experimental structures of organic, metal–organic and inorganic compounds and minerals; and collects the result of all types of crystallographic diffraction experiments (X-rays, electrons, or neutrons diffracted from single crystals and powders). In this respect, it is more diverse than most other databases that collect small-unit-cell structures. In recent years, furthermore, quantum mechanics computations using density functional theory and other methods have become powerful enough to yield reliable structural descriptions, either *ab initio* or in conjunction with experimental crystallographic techniques. Such structures are

Table 1: Synoptic description of main databases in crystallography.

Database/operator/URL	Subject area	No. structures	Founded	Operational model	Discussion	Reference
American Mineralogist Crystal Structure Database/Mineralogical Societies of America and Canada/ http://www.geo.arizona.edu/xtal-cgi/test	Minerals	>4000 mineral species	*ca* 2000	Financed by National Science Foundation; data sets freely available	Section 5.3	Downs & Hall-Wallace 2003
Bilbao Crystallographic Server/U. Bilbao/ http://www.cryst.ehu.es/	Properties of crystallographic symmetry groups		1997	Research grant funded; query results and software utilities available free of charge	Section 5.3	Aroyo *et al.* 2006
Cambridge Structural Database (CSD)/Cambridge Crystallographic Data Centre/ https://www.ccdc.cam.ac.uk	Chemical crystallography: organic and metal–organic compounds	>875,000	1965	Revenue received for value-added services and software from individual subscribers, national licences *etc.*; individual data sets freely available	Section 2.2	Groom *et al.* 2016
Crystallography Open Database/ http://www.crystallography.net/	Chemical crystallography: organic, inorganic, metal–organic compounds and minerals, excluding biopolymers	>376,000	2003	Mixed funding model (project, community and commercial user supported); data sets freely available	Section 2.1	Gražulis *et al.* 2009
CrystMet/Toth Information Systems Inc	Metals and intermetallics	>172,000	1960	Commercial; sells copyrighted database of critically evaluated data	Section 5.2	White *et al.* 2002
Incommensurate Structure Database/U. Bilbao/ http://webbdcrista1.ehu.es/incstrdb/	Incommensurate and modulated structures	>140		Research grant funded; data sets available free of charge	Section 5.3	
Inorganic Crystal Structure Database/ FIZ-Karlsruhe/ https://icsd.fiz-karlsruhe.de	Inorganic compounds	>187,000	1978	Revenue received for licensed access to data through copyrighted user interface	Section 5.1	Belsky *et al.* 2002

(contd.)

Magnetic Structures/U. Bilbao/ http://webbdcrista1.ehu.es/magndata/	Magnetic structures (commensurate and incommensurate)	>400		Research grant funded; data sets available free of charge	Section 5.3	Gallego *et al.* 2016
Mincryst/Institute of Experimental Mineralogy/Russian Academy of Sciences/ http://database.iem.ac.ru/mincryst/	Minerals and their structural analogues	>9800	1997	Financed by Russian Foundation of Basic Research; data sets freely available	Section 5.3	
Nucleic Acids Database/Rutgers U./ http://ndbserver.rutgers.edu/	Nucleic acids	>8800	1992	Grant funded; data sets freely available		Narayanan *et al.* 2013
Powder Diffraction File/International Centre for Diffraction Data/ http://www.icdd.com/	Materials science (material identification and characterization)	>890,000 diffraction patters, >330,000 crystal structures	1941	Non-profit database organization; receives no funding. Provides grants for targeting entries not available in open literature. Revenue from value added database/software sales	Section 3.1	Faber & Fawcett 2002
Protein Data Bank/WorldWide PDB/ http://www.pdb.org/	Biological macromolecules	~120,000	1971	Grant-funded international partner organization.*Data sets freely available	Section 4.1	Berman *et al.* 2003

*Funding sources for the wwPDB include: RCSB PDB is supported by NSF (DBI-1338415), NIH, DOE; PDBe by EMBL-EBI, Wellcome Trust (75968, 88944, 104948), BBSRC (BB/G022577/1, BB/J007471/1, BB/K016970/1, BB/K020013/1, BB/M013146/1, BB/M011674/1, BB/M020347/1, BB/M020428/1), EU (284209, 675858) and MRC (MR/L007835/1); PDBj and PDBj-BMRB by JST-NBDC, BMRB by NIGMS (GM109046).

also collected and are stored in a sister database, the TCOD (Theoretical Crystallography Open Database). Since the COD, TCOD and PCOD all use the same CIF framework for data representation, all databases can be searched and processed using the same software tools.

2.1.3 Features

To keep the COD simple and efficient, several ingredients are crucial. Free/Libre Open Source software (F/LOSS) is at the core of COD development, making it easy to reuse data and algorithms. An open data standard – the Crystallographic Information Framework – and its ontologies maintained by the IUCr are essential for efficient data exchange (Hall and McMahon 2016). The COD has maintained stable REST interfaces since its inception (Gražulis *et al.* 2009, 2012). From the very beginning each COD record is assigned a persistent identifier (PID) in the form of a seven-digit COD number, and each structure can be identified by a stable URI in the form http://www.crystallography.net/cod/2000000.html. Special care is taken to keep these URIs stable in accordance with the principle that 'cool URIs do not change' (Berners-Lee 1998). As a result, the COD is ideally suitable for the 21st century's open, connected world, providing stable links to crystal structures on the Web. The stable setup allows instantaneous reuse of COD data in various sites, and provides a mechanism for data citation using stable, unique identifiers for every structure.

The COD contains a simple search-and-retrieve Web interface as demonstrated in **Figure 1**. In addition to web search, it offers download of all data using various protocols (http, svn, rsync) and querying the database directly using a MySQL client. This gives maximum flexibility in cases where the Web interface is not enough, and permits integration of the COD into other websites or standalone programs.

The primary information in the COD is atomic coordinates and crystal descriptions. Moreover, the COD systematically stores experimental data from the diffraction experiment from which the structure was

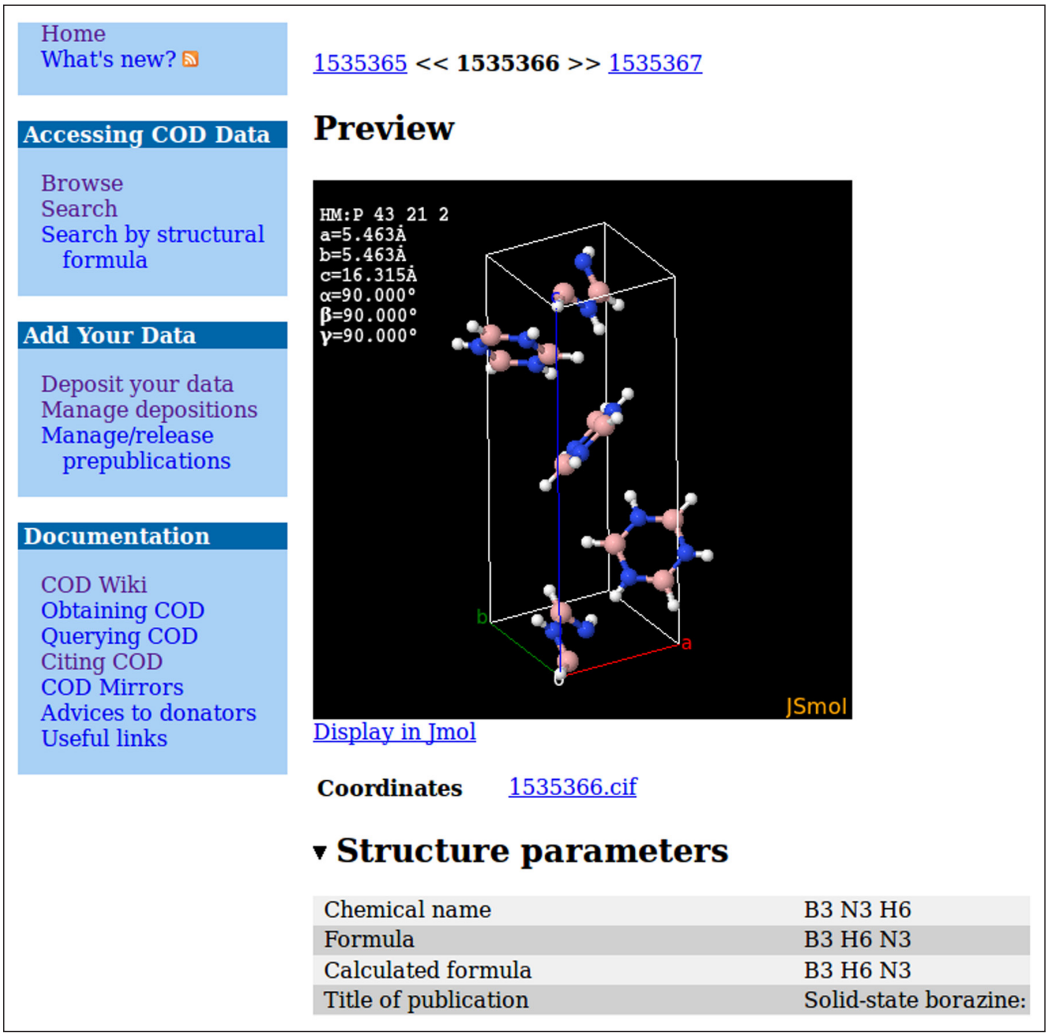

Figure 1: The COD search result interface and structure view page.

determined (F_{obs} and powder trace data) when these are available. COD data entries are manually curated as well as automatically checked. From the ingested data, the COD formulates a new CIF file which is guaranteed to be syntactically correct and contains all original data with attached metadata. Should a change of the COD entry be needed for data curation, the COD meticulously stores all changes in a version control database, currently *Subversion* (Collins-Sussman 2011), thus providing full traceability and data provenance.

The COD enjoys a growing field of applications. Teaching (Gražulis *et al.* 2015), powder identification (Lutterotti *et al.* 2015), and source of data for computational material science (First & Floudas 2013, Pizzi *et al.* 2016) are among the latest areas where the COD has proved useful as a source of data. Structures from the COD were used to produce 3D-printed models that find use in educational work (Moeck *et al.* 2014, Kaminsky *et al.* 2014, Stone-Sundberg *et al.* 2015, Scalfani *et al.* 2016). Moreover, the open nature of the COD allows structure descriptions to be matched against material properties (Pepponi *et al.* 2012), yielding the first open collection of material properties and property–structure relations.

In all cases, the COD has benefited from existing crystallographic CIF dictionaries maintained by the IUCr to describe crystallographic entities. However, it has also created domain-specific CIF dictionaries to describe material properties, computational settings and structural metadata. In this way existing CIF dictionaries have been reused without duplication, allowing the application of existing software and database structure for new fields of inquiry.

The COD continuity and its large collection of crystal data allow its use as a data source and management platform for industrial processes. The recently started SOLSA project allows the COD to provide crystal data for mineral identification in mines in real time. It will ingest public data that is obtained during such identification runs, further enriching the COD collection and increasing identification precision.

In the years to come, the COD will proceed to store much larger files that contain diffraction image data. Storing raw diffraction images currently poses a challenge to meet reasonable management and storage costs, along with enough bandwidth for the distribution of image sets. Here again the COD is pursuing a community backed, distributed and open-source solution: it has started to use a Tahoe-LAFS storage engine (Wilcox-O'Hearn & Warner 2008), which is capable of providing affordable redundant storage up to the order of petabytes.

2.2 The Cambridge Structural Database (CSD)

2.2.1 History

In 1965, a small research group based at the University of Cambridge set out to create a compilation of published organic and metal–organic crystal structure data. This group was led by Dr Olga Kennard acting on a vision she shared with the great polymath J D Bernal that 'collective use of data would lead to the discovery of new knowledge which transcends the results of individual experiments' (Kennard 1997). In time, this small research group became the Cambridge Crystallographic Data Centre (CCDC) and the initial compilation of data evolved into the Cambridge Structural Database (CSD). Much has changed since those early days: the systems used to maintain the Cambridge Structural Database have had to evolve to accommodate new requirements and many external services have been developed to support the continuing quest for new knowledge.

2.2.2 Scope

The CSD contains the results of over 875,000 crystal structure diffraction experiments and is growing by over 60,000 structures per year (Groom *et al.* 2016). Structures are typically submitted to the CSD by researchers in CIF format *via* deposition services provided by the CCDC. These deposition services promote community norms surrounding publication of crystal structure data (Larsen & Kostorz 2011) and make it easy for a researcher to abide by these (Davie & Agbenyega 2015). Many deposited structures end up being associated with a journal article, but increasingly researchers are using the CSD as their primary platform for publication of their crystal structure data (Groom 2016).

A deposited data set is assigned a Digital Object Identifier (DOI), a widely used type of persistent identifier that is actionable and enables interoperability (https://www.doi.org). A CSD DOI will resolve to a landing page where the data can be interactively viewed in a web browser and freely downloaded. The CCDC additionally provides mechanisms that enable linking and discovery of data from journal articles and other resources. Rich metadata is associated with each data set using a combination of automated processing and human validation. This metadata includes a chemical representation of the substance

studied, which is vital for enabling the collection of data to be mined for new knowledge. Encoding these chemical representations using the IUPAC International Chemical Identifier or InChI (Heller *et al.* 2015) enables links between chemical compound records in other resources and crystal structures in the CSD (Day 2014).

2.2.3 Features

The derivation of new knowledge from the data in the CSD is facilitated by a rich set of software applications developed by the CCDC. These enable researchers to search, analyse and visualize the data, and also to readily apply structural knowledge to a range of disciplines including drug discovery (Groom *et al.* 2012) and materials science (Galek *et al.* 2016). As well as enabling research, this area of the CCDC's activity attracts the financial revenue needed to sustain its core community activities from industry and academia without recourse to direct public funding.

The environment in which the CCDC operates has changed significantly since the early beginnings of the CSD in 1965. The rate at which structures are determined has increased rapidly and the data being deposited have become more complex and more diverse. General initiatives around research data management and discovery have resulted in opportunities for new services and greater interoperability between resources.

The CCDC has had to develop its systems over the past half century to adapt to this changing environment whilst remaining sustainable. Internal informatics systems have recently evolved to accommodate increasing throughput and complexity, while external services are developing to ensure that both data and knowledge in the CSD continue to be widely and appropriately available to a broad range of user communities. Although there have been many changes since the early days, one particular feature has been retained: the validation and review undertaken by expert scientists to ensure that data can be reliably used in the discovery of new knowledge today and in the decades to come.

3 Materials Science

3.1 Powder Diffraction File

3.1.1 History

The International Centre for Diffraction Data (ICDD®) Powder Diffraction File is a powerful database for materials identification and has been used extensively by the scientific community for 75 years. Over this period, the database has grown exponentially, housing more than 890,000 diffraction patterns and 330,000 crystal structures in Release 2016. Starting with 1000 hand-written file cards in 1941, the Powder Diffraction File (PDF®) has undergone numerous technological developments in the database as well as search/match and data retrieval methods (Faber & Fawcett, 2002).

The PDF started in 1941 as a collection of X-ray powder diffraction patterns of single-phase material providing chemistry, interplanar spacings (*d*) and relative intensities (*I*). The amount of data that could be presented was limited in the early days as it was published in printed form. In 1969, it was made available to users in the form of magnetic tapes allowing for some additional crystallographic information to be provided. However, data storage was still an issue on account of expensive disk space and slow computer speeds and data retrieval, rather different from today, when one can store and access gigabytes of data very rapidly using a personal computer. During continued development in data storage technology, the PDF started adding much more information than just *d-I* lists to the database. In 2000, it was redesigned using a Relational Database (RDB) concept originally proposed by E F Codd (1970).

3.1.2 Scope

Today, the Powder Diffraction File in Relational Database format contains extensive chemical, physical, bibliographic and crystallographic data including atomic coordinates enabling qualitative and quantitative phase analysis (Kabekkodu *et al.* 2002). In addition to X-ray diffraction patterns, the PDF also contains electron and neutron diffraction patterns enabling scientists to use different diffraction techniques to characterize the material of interest. There have been active recent developments in PDF database capabilities and search/match algorithms.

In 2011, ICDD extended the design of the database to include modulated structures. As these structures need to be described in 3+*n* dimensions (using a superspace approach), the PDF database has had to be modified in a way that does not change the core format that has been used by several search/match software developers for years.

3.1.3 Features

Over the years Relational Database Management Systems (RDBMS) have evolved exponentially offering much more robust and efficient database platforms. This has facilitated the storage in the PDF of significantly more data including chemical, physical, bibliographic, structural classifications, raw powder patterns, electron and neutron diffraction patterns and crystallographic data including atomic coordinates. All of this additional data content becomes essential when addressing a difficult materials characterization problem. During the past few years ICDD expanded the database scope beyond purely crystalline materials to enable the characterization of amorphous, poorly crystalline and nanomaterials.

The PDF now contains raw diffraction patterns of targeted amorphous, polymer, disordered clays and nanomaterials (and has done so for more than 20 years). Comparing collected diffraction data with reference raw diffraction patterns is essential for these materials as a *d-I* list is not sufficient to do a comprehensive phase identification. As of release 2016 of the Powder Diffraction File there are 11,284 raw powder diffraction data archived. These experimental raw data are now beginning to be employed by users of the database for their own analysis. Recently the crystal structure of trandolapril was solved by archived powder diffraction raw data in the PDF (Reid *et al.* 2016). **Figure 2** is an example of a PDF entry with archived raw data.

Quality marks, assigned to all PDF database entries, are extremely important when working with a large database with multiple experimental determinations. It is important to know the quality of the crystal structure or diffraction pattern found in the database to screen the search/match result set. With varying quality of published data in the literature, database editorial review processes had to adopt rigorous data evaluation methods to classify diffraction data based on its quality. All PDF data are thoroughly reviewed, edited and standardized. ICDD performs more than 100 quality and error checks on the data prior to publication. ICDD's editorial group has developed an extensive data validation suite to enrich the editorial process. ICDD corrects all of the resolvable errors found in the original data. An entry gets rejected if the magnitude of an error is large and there is insufficient information to correct the observed error. When the quality mark of a PDF database entry is lowered, those entries are populated with editorial comments describing original data errors and any corrections applied. Data validation, corrections, classification and quality mark assignments are the most important and time consuming step in the editorial process.

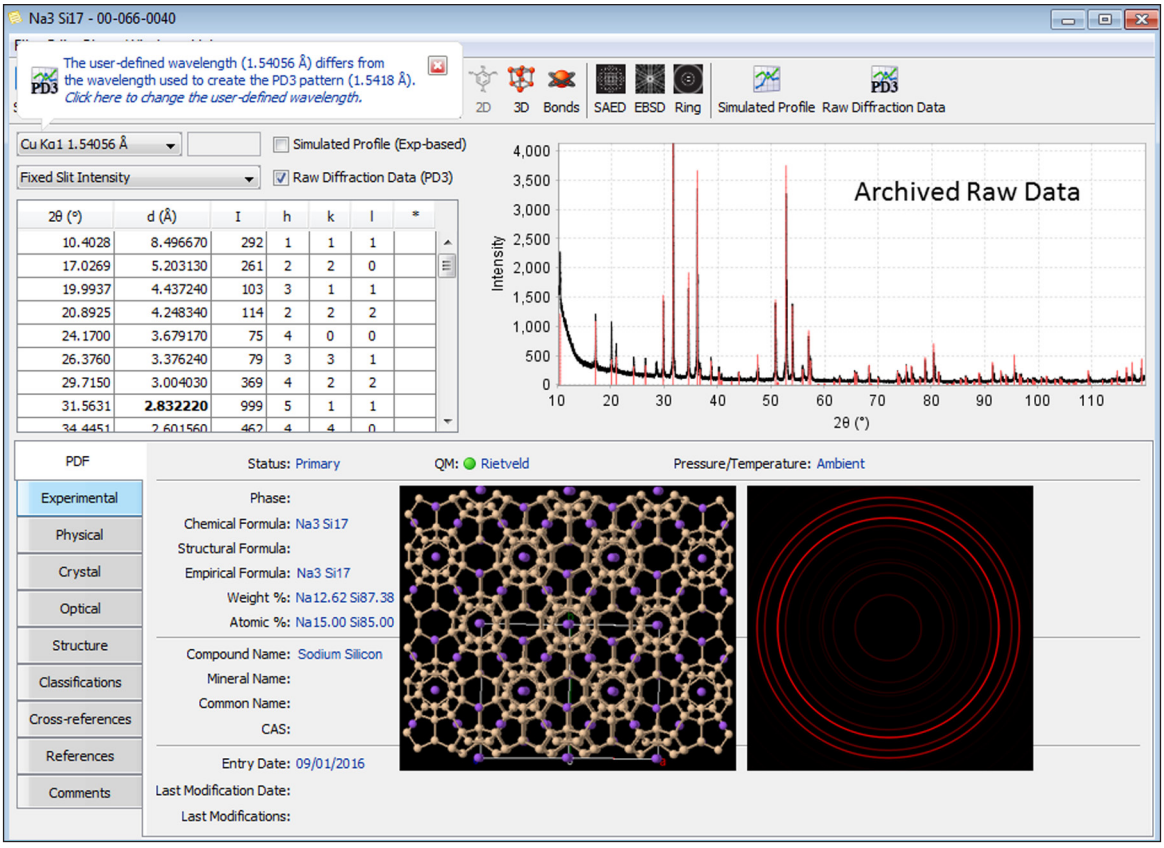

Figure 2: Powder Diffraction File entry with archived raw data.

4 Biological crystallography

4.1 Protein Data Bank

4.1.1 History

The Protein Data Bank (PDB) is the single global repository of experimentally determined 3D structures of biological macromolecules and their complexes. It was established in 1971, becoming the first open-access digital data resource in the biological sciences. The PDB serves a growing structural biology community, which generates the incoming data, and provides open access for researchers and educators/students around the world to ~120,000 archival entries plus value added information central to research and education in the basic and applied biological sciences and medicine. The PDB archive also supports research and teaching efforts in the areas of agriculture, animal husbandry, biological energy production, biochemistry, ecology, protein design, and nanotechnology.

The PDB archive is now managed by the Worldwide Protein Data Bank organization (wwPDB) (Berman *et al.* 2003), which currently includes three founding regional data centres, located in the US (RCSB Protein Data Bank, RCSB PDB: http://rcsb.org), Japan (Protein Data Bank Japan, PDBj: http://pdbj.org) and Europe (Protein Data Bank in Europe, PDBe: http://pdbe.org), plus a global NMR specialist data repository BioMagResBank, composed of deposition sites in the US (BMRB: http://www.bmrb.wisc.edu) and Japan (PDBj-BMRB: http://bmrbdep.pdbj.org). Together, these wwPDB Partners collect, annotate, validate and disseminate standardized PDB data to the public without any limitations on its use.

4.1.2 Scope

Approximately 11,000 new depositions come annually from every inhabited continent, from structural biologists using X-ray crystallography, NMR spectroscopy, and electron microscopy. Worldwide, more than 1 million users come to the PDB every year, as judged by counting unique IP addresses. They download more than 1.5 million structures every day, or more than 500 million per year. They come from 192 of the 195 sovereign nations currently recognized by the United Nations.

4.1.3 Features

The wwPDB Partners are developing the next generation of deposition and annotation tools. A major focus of these tools is validation of submitted atomic coordinates and primary experimental data plus associated metadata. wwPDB biocurators provide depositors with detailed reports that include the results of model and experimental data validation. Validation reports were developed and improved using recommendations from expert Validation Task Forces and Workshops for X-ray (Read *et al.* 2011), electron microscopy (Henderson *et al.* 2012), NMR (Montelione *et al.* 2013), and ligands (Adams *et al.* 2016). As these wwPDB validation reports provide assessments of structure quality using widely accepted standards and criteria, the wwPDB Partners strongly encourage journal editors and referees to request them from authors as part of the manuscript submission and review process. The reports are date-stamped and display the logo of the wwPDB site where the deposition was curated. They contain the same information, regardless of which wwPDB site processed the incoming data. Provision of wwPDB validation reports is already required by Nature Publishing Group Journals, *eLife*, the *Journal of Biological Chemistry*, and IUCr journals as part of their manuscript-submission processes. Once each new PDB entry is released, typically at the time of publication, wwPDB validation reports are made public together with the deposited data.

To analyse how PDB data relate to other publicly available annotations, the RCSB Protein Data Bank (Berman *et al.* 2000) has developed a novel data integration platform that maps 3D structural information across various data sets. This integration bridges from the human genome across protein sequence to 3D structure space. Users can perform simple searches from the top search bar (*e.g.* ID, name, sequence, ligand) or build complex combinations of search parameters using Advanced Search. Information from DrugBank (Wishart *et al.* 2006) is integrated with PDB data to facilitate searches for drugs and drug targets. Other classification systems are used to organize PDB structures in hierarchical trees for browsing and searching (*e.g.* mpstruc, Gene Ontology, Enzyme Classification).

Protein Feature View provides a graphical mapping of PDB entries onto full-length UniProt sequences. It displays observed and unobserved regions in PDB structures, secondary structure, domain architecture, protein disorder, hydrophobicity profiles, genetic variations, and exon boundaries. Information about available homology models from the Protein Model Portal is also displayed.

Gene View supports visualization of the relationship between genomic and 3D structural data. This tool allows browsing of the human genome with PDB data highlighted onto corresponding genomic ranges.

Similar to the Protein Feature View, these data can be correlated with other genomic and protein functional annotations, such as gene structure annotations, DNA repeats, or sequence conservation.

Genome Location to Structure Mapping. A human gene location, *e.g.* the site of a mutation, can be mapped onto a human gene, and from there to the corresponding protein sequence, and 3D structure. This tool can be used to assess the effect of a single nucleotide variation on the 3D structure.

Validation and Ligand Electron Density Maps. Validation reports and corresponding graphical summary indicators for X-ray, and recently NMR and 3DEM structures are distributed by the wwPDB. These reports assess the quality of each structure and highlight specific concerns. They are provided to depositors as part of the wwPDB Deposition and Annotation System to help identify potential problems that should be addressed prior to PDB deposition and publication. For each structure, the RCSB PDB displays the graphical validation summary and links to the corresponding report. Also at the RCSB PDB website, 3D visualization of the Sigma-weighted $2m|F_o| - d|F_c|$ electron density 'mini-maps' around ligands is available to assess the quality of ligand modeling.

4.1.4 Outreach and Education

PDB archival data are also being used to build educational resources that connect between biomolecular structure and function. RCSB PDB has been publishing Molecule of the Month (Goodsell *et al.* 2015) articles online continuously for the past 16 years, presenting a selected set of molecular structures in the context of diverse biological and biomedical themes. In 2011, the RCSB PDB established PDB-101 (http://pdb101.rcsb.org) as an outreach and educational resource aimed at training novice PDB users and introducing the power of structural biology and the PDB to diverse audiences. These resources are used by students and educators in primarily high school and undergraduate studies.

5 Inorganic crystallography and metals

As mentioned in the *Introduction*, several other curated databases offer particular collections of experimentally determined crystal and molecular structures for use by the research community.

5.1 Inorganic Crystal Structure Database

The Inorganic Crystal Structure Database (ICSD) was initiated in 1978 at the Institute of Inorganic Chemistry of the University of Bonn, but has since 1989 been offered as one of the chemical databases managed by the Fachinformationszentrum Karlsruhe, a non-profit organization that is part of the Gottfried Wilhelm Leibniz Scientific Community. ICSD provides the scientific and industrial communities with the world's largest database for completely identified inorganic crystal structures, currently holding more than 187,000 peer-reviewed data entries, including their atomic coordinates, dating back to 1913. These comprise over 2000 crystal structures of the elements, nearly 35,000 records for binary compounds, over 68,000 records for ternary compounds, and more than 68,000 records for quarternary and quintenary compounds. About 80% of entries have been assigned a structure type, and there are currently over 9000 structure prototypes.

In common with the other databases described here, the ICSD is conscious of the need for high-quality data ingest and curation. To quote their website 'As the world's leading provider of scientific information on inorganic crystal structures, we take full responsibility for database production, maintenance and quality control, and we ensure that the ICSD database and our software solutions meet the highest possible quality standards'.

From 2008 to 2013 FIZ Karlsruhe worked on the incorporation of legacy crystal structure data on metallic and intermetallic compounds from their cooperation partner, The National Institute of Standards and Technology (NIST), Gaithersburg, USA, into the ICSD database.

5.2 The Metals Database

The CrystMet database provides critically evaluated structural data and powder patterns for metals, including alloys, intermetallics and minerals. Founded in 1960 at the Los Alamos Laboratory, it became part of the Scientific Numeric Databases system of the National Research Council of Canada, and was subsequently transferred to Toth Information Systems, Inc., from which it is available as a commercial product. This illustrates particularly well the need for curated database systems to be prepared to be adaptable to prevailing funding models if they are to be sustainable over long periods of time.

5.3 Other databases

Also noteworthy are databases of mineral phases (MinCryst) and crystal structures (the *American Mineralogist* Crystal Structure Database); of incommensurate and magnetic structures and of the properties of plane groups, space groups *etc.* that determine crystal symmetry (Bilbao Crystallographic Server http://www.cryst.ehu.es/). These last three are maintained by the Condensed Matter Physics Group at the Basque University of Bilbao, Spain (Aroyo *et al.* 2009).

6 Synergies between databases

This survey has presented individual databases in accordance with the structure of the SciDataCon2016 session that inspired this article. Such a survey illustrates the wide range of scientific disciplines encompassed by crystallography, and allows each database organisation to highlight its particular strengths in handling data sets within its area of expertise. However, such an approach does not adequately convey the extent to which the various databases interoperate and overlap, to provide complementary and comprehensive resources for structural scientists of any discipline. For example, the Cambridge Crystallographic Data Centre (CCDC) and the wwPDB work closely together to improve the description of small chemical compounds (ligands) bound to protein molecules (Davie & Burley 2013, wwPDB 2015, Adams *et al.* 2016, Groom & Cole 2017). The CCDC and FIZ Karlsruhe are embarking on a joint development project to provide shared deposition and access services for crystallographic data across all domains of chemistry (Davie & Mueller, 2017). The *American Mineralogist* Crystal Structure Database is fully embedded in the Crystallography Open Database to allow use of the latter's more general software environment (Gražulis *et al.* 2012). The crystallographic structures of nucleic acids curated by the Nucleic Acids Database are fully incorporated in the Protein Data Bank.

As well as formal alliances between databases, their worth to the community has prompted the establishment at various times of unified or federated access systems, such as the UK National Chemical Database Service (http://cds.rsc.org) funded by the Engineering and Physical Sciences Research Council.

The Bilbao Crystallographic Server is of particular interest because it does not store crystal structures, but instead provides comprehensive information on the symmetry properties of two- and three-dimensional lattices, which determine the packing arrangements of any type of crystalline material. It also supplies software tools that enable the identification of space groups, determine group–subgroup relationships enabling prediction or understanding of phase transitions, and have a wide range of solid-state applications.

An important aspect of interoperability between databases, frequently mentioned in the commentaries above, is the availability of the Crystallographic Information Framework (Hall & McMahon 2016) as a common machine-readable descriptive framework or ontology. Continuing efforts to promote the development of experimental metadata schemas within the same framework (see Section 7) aim to maximize the potential for further interoperability, as does the recommendation of the IUCr's Diffraction Data Deposition Working Group that all distinct data sets should have a unique persistent identifier (Kroon-Batenburg *et al.* 2017). This recommendation currently stops short of specifying a particular globally unique schema, in part because there are already well-established practices of quoting database-specific identifiers (*e.g.* the PDB id for biological macromolecules, the CSD 'refcode' for molecules in the CSD, the COD identifier). We note that these identifiers share with DOI a common encompassing standard for defining identifier syntax and name spaces, namely the uniform resource identifier URI (Berners-Lee *et al.* 2005); thus, compatibility is ensured, existing resolution mechanisms can be applied (*e.g. via* http) to all three identifier groups, and there is no danger of name clashes or ambiguities. Nevertheless, as seen above, the DOI system (International DOI Foundation 2017), adopted by most of the major crystallographic databases, accommodates existing internal mechanisms for ensuring uniqueness, has a well-established infrastructure supported by journal and data publishers, and may be the best mechanism to ensure global uniqueness of persistent identifiers across different disciplines, and hence facilitate interoperability. The cost of registering DOIs, while modest on an individual basis, must also be a factor in the sustainable operation of large-scale databases.

Another contribution towards making databases interoperable is the provision of RESTful services (*i.e.* services that are accessed through uniform resource identifiers, URIs, according to the representational state transfer approach). These have the property of permitting extensive use of database features without the need for specialized query tools – any browser is capable of retrieving the required resource, which may take the form of a complex query. The accounts above describe certain specific implementations, but it is worth noting that most databases with a Web-accessible interface are beginning to provide such services.

7 Future developments

An aspect identified as important by the International Union of Crystallography is the extension of data archiving to the raw diffraction data. Crystallography compares well with other fields such as astronomy and particle physics in ensuring the provision of the underlying data as a formal requirement of research publication, whether it be derived, processed or raw (*i.e.* primary) data. Archiving of raw data is obviously challenging on account of its much greater volume. However, this is not the only challenge; the metadata associated with the stored raw data must be properly characterized so that it can be assessed and understood. Also, individual raw diffraction data sets must be discoverable and reusable by other researchers, whether associated with formal publications or not. The experience of the IUCr community is that existing obstacles towards raw diffraction data archiving have largely dissolved through a combination of positive developments in recent years. These include: establishment of university research data archives *e.g.* at the University of Manchester Library (University of Manchester 2013); centralized initiatives at ESRF (European Synchrotron Radiation Facility, 2016) and Zenodo (OpenAIRE 2017) in Europe; SBGrid Data Bank (https://data.sbgrid.org/; Meyer *et al.* 2016) and the BD2K (Big Data to Knowledge) initiative (Grabowski *et al.* 2016) in the USA. These extend the exemplary approaches of data archiving at the neutron facilities ISIS (Science & Technology Facilities Council 2011) and Institut Laue Langevin (Institut Laue Langevin 2011), and at Store. Synchrotron in Australia (Meyer *et al.* 2014).

The ease of archiving of raw diffraction data sets is a remarkable development of recent years; as an example Zenodo now accepts up to 50 Gbyte data sets in a single DOI deposition. The proper archiving and sharing of raw data is the foundation of any scientific analysis but also it allows sharing at the earliest stage possible of a study. Thus, with raw data accompanied by a preprint, it is hoped that scientific discoveries can be accelerated, especially important with societal challenges; the EU's OPENAIRE project (https://www.openaire.eu/) includes such aspirations. As a practical example of the use of Zenodo for preprint and diffraction images data set deposition and DOI registration see Helliwell and Tanley (2016a, b).

The archives mentioned above have developed extensive details on how to manage raw data and obtain and use appropriate metadata. Nevertheless, specialized metadata appropriate to different experiments and structure models, associated with the various IUCr Commissions' domains of expertise, still need improved definitions. The IUCr is proactively engaging with its constituent research communities *via* workshops, *e.g.* at the 2015 European Crystallographic Meeting and the 2017 Annual Meeting of the American Crystallographic Association (Kroon-Batenburg *et al.* 2017). These activities also link closely with COMCIFS, the IUCr committee responsible for maintenance of the CIF standard. All this activity demonstrates that changes in established practice are possible, building on wonderful new digital storage capacities. Open access to the archived stored data is energetically encouraged, within research funder norms and rules.

8 Concluding remarks

It is seen from the discussion above that there is a great variety of curated databases in the field of crystallography and related structural sciences. Many have a long history and tradition of data sharing, in some cases even dating back to before the era of digital storage and dissemination. Partly on account of this history, they have evolved with different operational models, intended in all cases to ensure sustainability. Income generation to ensure long-term sustainability of curated databases that add value through validation, annotation and enhanced search, is a challenge (Dillo *et al.* 2016). The desire to maximize the availability of research data in accordance with so-called FAIR principles – Findable, Accessible, Interoperable, and Re-usable (Force11 2014) – encourages dissemination of research data under non-restrictive licences, yet a controlled licensing of database software to subscribers may be the most effective approach to sustainability in many cases. The diversity of operational models indicated in **Table 1** illustrates the need for crystallography to accommodate multiple approaches to these challenges [see also the response of Hackert *et al.* (2016) to the Science International (2015) Accord *Open Data in a Big Data World*].

Common to all the databases discussed here is the belief that their greatest value is in providing highly reliable data, where 'reliability' is defined by rigorous validation strategies and quality indicators. Many of the databases actively work with journals and depositors to provide feedback at an early stage, often actually improving the quality of the data that is to be deposited. All are reliant on the data exchange standard CIF, developed by the IUCr to enhance portability, interoperability and full characterization of structural data.

They can be driven by different primary goals. For some, providing open access to as large a collection of data sets within their domain of interest is paramount; for others, the emphasis is on developing search and analysis tools that go far beyond what can be achieved by generic open-source applications. Some concentrate on particular types of chemical compound; others are broadly based. Yet each provides a valuable service to science, and is highly regarded within the research community. This diverse and thriving ecosystem of

databases, underpinned by data exchange standards and the desire to ensure the highest achievable quality, is key to the strength of crystallography's world-wide data infrastructure (Genova *et al.* 2017).

Acknowledgements

We thank the IUCr Diffraction Data Deposition Working Group and its Consultants for their work with us these last 5 years. See http://forums.iucr.org/viewforum.php?f=21. The COD project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No 689868 (SOLSA Project). The CCDC would like to thank all those in the scientific community who have shared their data through the CSD. The ICDD would like to thank ICDD members, editors and volunteers for their contributions to the Powder Diffraction File database. The RCSB PDB is supported by the National Science Foundation (DBI 1338415), National Institutes of Health, and the Department of Energy; PDBe by the Wellcome Trust, BBSRC, MRC, EU, CCP4, and EMBL-EBI; PDBj by JST-NBDC; and BMRB by the National Institute of General Medical Sciences (GM109046). We are grateful to anonymous referees for suggestions and recommendations.

Competing Interests

Ian Bruno is employed by the Cambridge Crystallographic Data Centre which produces and maintains the Cambridge Structural Database. Soorya Kabekkodu is employed by the ICDD which produces and maintains the Powder Diffraction File™. There are no other competing interests.

References

Adams, P D, Aertgeerts, K, Bauer, C, Bell, J A, Berman, H M, Bhat, T N, Blaney, J M, Bolton, E, Bricogne, G, Brown, D, Burley, S K, Case, D A, Clark, K L, Darden, T, Emsley, P, Feher, V A, Feng, Z, Groom, C R, Harris, S F, Hendle, J, Holder, T, Joachimiak, A, Kleywegt, G J, Krojer T, Marcotrigiano, J, Mark, A E, Markley, J L, Miller, M, Minor, W, Montelione, G T, Murshudov, G, Nakagawa, A, Nakamura, H, Nicholls, A, Nicklaus, M, Nolte, R T, Padyana, A K, Peishoff, C E, Pieniazek, S, Read, R J, Shao, C, Sheriff, S, Smart, O, Soisson, S, Spurlino, J, Stouch, T, Svobodova, R, Tempel, W, Terwilliger, T C, Tronrud, D, Velankar, S, Ward, S C, Warren, G L, Westbrook, J D, Williams, P, Yang, H and **Young, J** 2016 Outcome of the First wwPDB/CCDC/D3R Ligand Validation Workshop. *Structure*, 24(4): 502–508. DOI: https://doi.org/10.1016/j.str.2016.02.017

Allen, F H, Bergerhoff, G and **Sievers, R** 1987 *Crystallographic Databases*. Chester, UK: International Union of Crystallography.

Allen, F H and **Glusker, J P** 2002 Preface to Special Issue on Crystallographic Databases. *Acta Crystallogr*. B58(3): unnumbered pages. DOI: https://doi.org/10.1107/S0108768102006638. Also published as *Acta Crystallogr*, D58(6): unnumbered pages. DOI: https://doi.org/10.1107/S0907444902008399

Aroyo, M I, Perez-Mato, J M, Capillas, C, Kroumova, E, Ivantchev, S, Madariaga, G, Kirov, A and **Wondratschek, H** 2009 Bilbao Crystallographic Server: I. Databases and crystallographic computing programs. *Z. Kristallogr*, 221: 15–27. DOI: https://doi.org/10.1524/zkri.2006.221.1.15

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*, B58: 364–369. DOI: https://doi.org/10.1107/S0108768102006948

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

Berman, H M, Henrick, K and **Nakamura, H** 2003 Announcing the worldwide Protein Data Bank. *Nature Structural Biology*, 10: 980. DOI: https://doi.org/10.1038/nsb1203-980

Berman, H M, Westbrook, J, Feng, Z, Gilliland, G, Bhat, T N, Weissig, H, Shindyalov, I N and **Bourne, P E** 2000 The Protein Data Bank. *Nucleic Acids Res*, 28(1): 235–242. DOI: https://doi.org/10.1093/nar/28.1.235

Berners-Lee, T 1998 *Cool URIs don't change*. Available at: https://www.w3.org/Provider/Style/URI.html [Last accessed 10 April 2017].

Berners-Lee, T, Fielding, R and **Masinter, L** 2005 *Uniform Resource Identifier (URI): Generic Syntax*. Available at: https://tools.ietf.org/html/rfc3986 [Last accessed 10 April 2017].

Bernstein, H J, Bollinger, J C, Brown, I D, Gražulis, S, Hester, J R, McMahon, B, Spadaccini, N, Westbrook, J D and **Westrip, S P** 2016 Specification of the Crystallographic Information File format, version 2.0. *Journal of Applied Crystallography*, 49: 277–284. DOI: https://doi.org/10.1107/S1600576715021871

Codd, E F 1970 A Relational Model of Data for Large Shared Data Banks. *Communications of the ACM*, 13(6): 377–387. DOI: https://doi.org/10.1145/362384.362685

Collins-Sussman, B, Fitzpatrick, B W and **Pilato, C M** 2011 *Version Control with Subversion*. Available at: http://svnbook.red-bean.com/en/1.7/svn-book.html [Last accessed 10 April 2017].

Davie, P and **Agbenyega, J** 2015 Collaboration between the CCDC and the IUCr Streamlines Crystallographic Data Deposition into the Cambridge Structural Database. Available at: https://www.ccdc.cam.ac.uk/News/List/post-40/ [Last accessed 10 April 2017].

Davie, P and **Burley, S K** 2013 The Cambridge Crystallographic Data Centre establishes US operations in new partnership with the Rutgers University Center for Integrative Proteomics Research. Available at: https://www.ccdc.cam.ac.uk/News/List/post-25 [Last accessed 10 April 2017].

Davie, P and **Mueller, H** 2017 Alliance Reshapes Crystallography Data Access. Available at: https://www.ccdc.cam.ac.uk/News/List/2017-03-27-alliance-reshapes-crystallography-data-access [Last accessed 10 April 2017].

Day, A 2014 Linking 2D RSC ChemSpider Compounds to 3D CCDC Crystals. Available at: http://www.rsc.org/blogs/escience/2014/12/linking-2-d-rsc-chemspider-compounds-3-d-ccdc-crystals [Last accessed 10 April 2017].

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

Dillo, I, Hodson, S and **de Waard, A (RDA-WDS Interest Group on Income Streams for Data Repositories)** 2016 Income Streams for Data Repositories. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.46693

Downs, R T and **Hall-Wallace, M** 2003 The *American Mineralogist* Crystal Structure Database. *Am. Mineral*, 88: 247–250

European Synchrotron Radiation Facility 2016 ESRF takes the helm in saving data. Available at: http://www.esrf.eu/fr/home/news/general/content-news/general/esrf-takes-the-helm-in-saving-data.html [Last accessed 10 April 2017].

Faber, J and **Fawcett, T** 2002 The Powder Diffraction File: present and future. *Acta Crystallographica Section B: Structural Science*, 58: 325–332. DOI: https://doi.org/10.1107/S0108768102003312

First, E L and **Floudas, C A** 2013 MOFomics: Computational pore characterization of metal–organic frameworks. *Microporous and Mesoporous Materials*, 165: 32–39. DOI: https://doi.org/10.1016/j.micromeso.2012.07.049

Force11 2014 The FAIR data principles. Available at: https://www.force11.org/group/fairgroup/fairprinciples [Last accessed 10 April 2017].

Galek, P T A, Pidcock, E, Wood, P A, Feeder, N and **Allen, F H** 2016 Navigating the Solid Form Landscape with Structural Informatics. In *Computational Pharmaceutical Solid State Chemistry*, 15–35. Hoboken, NJ: John Wiley & Sons, Inc. DOI: https://doi.org/10.1002/9781118700686.ch2

Gallego, S V, Perez-Mato, J M, Elcoro, L, Tasci, E S, Hanson, R M, Momma, K, Aroyo, M I and **Madariaga, G** 2016 *MAGNDATA*: towards a database of magnetic structures. I. The commensurate case. *J. Appl. Cryst.*, 49: 1750–1776. DOI: https://doi.org/10.1107/S1600576716012863

Genova, F, Arviset, C, Almas, B M, Bartolo, L, Broeder, D, Law, E and **McMahon, B** 2017 Building a Disciplinary, World-Wide Data Infrastructure. *Data Science J.*, 16: 16. DOI: https://doi.org/10.5334/dsj-2017-016

Goodsell, D S, Dutta, S, Zardecki, C, Voigt, M, Berman, H M and **Burley, S K** 2015 The RCSB PDB 'Molecule of the Month': Inspiring a Molecular View of Biology. *PLoS Biol*, 13: e1002140. DOI: https://doi.org/10.1371/journal.pbio.1002140

Grabowski, M, Langner, K M, Cymborowski, M, Porebski, P J, Sroka, P, Zheng, H, Cooper, D R, Zimmerman, M D, Elsliger, M-A, Burley, S K and **Minor, W** 2016 A public database of macromolecular diffraction experiments. *Acta Crystallogr*, D72: 1181–1193. DOI: https://doi.org/10.1107/S2059798316014716

Gražulis, S, Chateigner, D, Downs, R T, Yokochi, A F T, Quirós, M, Lutterotti, L, Manakova, E, Butkus, J, Moeck, P and **Le Bail, A** 2009 Crystallography Open Database – an open-access collection of crystal structures. *Journal of Applied Crystallography*, 42: 726–729. DOI: https://doi.org/10.1107/S0021889809016690

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

Gražulis, S, Sarjeant, A A, Moeck, P, Stone-Sundberg, J, Snyder, T J, Kaminsky, W, Oliver, A G, Stern, C L, Dawe, L N, Rychkov, D A, Losev, E A, Boldyreva, E V, Tanski, J M, Bernstein, J, Rabeh, W M and **Kantardjieff, K A** 2015 Crystallographic education in the 21st century. *Journal of Applied Crystallography*, 48: 1964–1975. DOI: https://doi.org/10.1107/S1600576715016830

Groom, C R 2016 New Communications with the New CSD. Available at: http://www.ccdc.cam.ac.uk/Community/blog/2016-03-15-new-communications-with-the-new-csd [Last accessed 10 April 2017].

Groom, C R, Bruno, I J, Lightfoot, M P and **Ward, S C** 2016 The Cambridge Structural Database. *Acta Crystallogr*, B72: 171–179. DOI: https://doi.org/10.1107/S2052520616003954

Groom, C R and **Cole, J C** 2017 The use of small-molecule structures to complement protein–ligand crystal structures in drug discovery. *Acta Crystallogr*, D73: 240–245. DOI: https://doi.org/10.1107/S2059798317000675

Groom, C R, Olsson, T S G, Liebeschuetz, J W, Bardwell, D A, Bruno, I J and **Allen, F H** 2012 5. Mining the Cambridge Structural Database for Bioisosteres. In: Brown, N (ed.), *Bioisosteres in Medicinal Chemistry*, 75–101. Weinheim, Germany: Wiley-VCH Verlag GmbH & Co. KGaA. DOI: https://doi.org/10.1002/9783527654307.ch5

Hackert, M L, Van Meervelt, L, Helliwell, J R and **McMahon, B** 2016 *Open Data in a Big Data World: A position paper for crystallography*. Chester: International Union of Crystallography. Available at: http://www.iucr.org/iucr/open-data [Last accessed 10 April 2017].

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

Hall, S R and **McMahon, B** 2016 The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. *Data Science Journal*, 15: 3. DOI: https://doi.org/10.5334/dsj-2016-003

Heller, S R, McNaught, A, Pletnev, I, Stein, S and **Tchekhovskoi, D** 2015 InChI, the IUPAC International Chemical Identifier. *J. Cheminform*, 7: 23. DOI: https://doi.org/10.1186/s13321-015-0068-4

Helliwell, J R and **Tanley, S W M** 2016a Preprint: Atomic resolution X-ray crystal structure of cisplatin bound to hen egg white lysozyme stored for 5 years 'on the shelf'. DOI: https://doi.org/10.5281/zenodo.155068

Helliwell, J R and **Tanley, S W M** 2016b Raw diffraction data for atomic resolution X-ray crystal structure of cisplatin bound to hen egg white lysozyme stored for 5 years 'on the shelf'. DOI: https://doi.org/10.5281/zenodo.154704

Henderson, R, Sali, A, Baker, M L, Carragher, B, Devkota, B, Downing, K H, Egelman, E H, Feng, Z, Frank, J, Grigorieff, N, Jiang, W, Ludtke, S J, Medalia, O, Penczek, P A, Rosenthal, P B, Rossmann, M G, Schmid, M F, Schroder, G F, Steven, A C, Stokes, D L, Westbrook, J D, Wriggers, W, Yang, H, Young, J, Berman, H M, Chiu, W, Kleywegt, G J and **Lawson, C L** 2012 Outcome of the first electron microscopy validation task force meeting. *Structure*, 20(2): 205–214. DOI: https://doi.org/10.1016/j.str.2011.12.014

Institut Laue Langevin 2011 ILL Data Policy. Available at: https://www.ill.eu/users/ill-data-policy [Last accessed 10 April 2017].

International DOI Foundation 2017 The DOI System. Available at: https://www.doi.org [Last accessed 10 April 2017].

International Union of Crystallography 2017 Data activities in crystallography. Available at: http://www.iucr.org/resources/data [Last accessed 10 April 2017].

Kabekkodu, S N, Faber, J and **Fawcett, T** 2002 New Powder Diffraction File (PDF-4) in relational database format: advantages and data-mining capabilities. *Acta Crystallographica Section B: Structural Science*, 58: 333–337. DOI: https://doi.org/10.1107/S0108768102002458

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 Diffraction*, 29: S42–S47. DOI: https://doi.org/10.1017/s0885715614001092

Karen, V L and **Mighell, A** 1996 NIST Workshop on Crystallographic Databases: Preface. *J. Res. Natl Inst. Stand. Technol*, 101: iii. DOI: https://doi.org/10.6028/jres.101.001

Kennard, O 1997 From Private Data to Public Knowledge. In: Butterworth, I (ed.), *The Impact of Electronic Publishing on the Academic Community*, 159–166. London: Portland Press Ltd.

Kroon-Batenburg, L M J, Helliwell, J R, McMahon, B and **Terwilliger, T C** 2017 Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements. *IUCrJ*, 4: 87–99. DOI: https://doi.org/10.1107/S2052252516018315

Larsen, S and **Kostorz, G** 2011 Publication standards for crystal structures. Available at: http://www.iucr.org/home/leading-article/2011/2011-06-02 [Last accessed 10 April 2017].

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

Lutterotti, L, Chateigner, D, Pillière, H and **Fontugne, C** 2015 *Full-pattern search-match using the Crystallography Open Database: an Internet tool*. Available at: http://www.ecole.ensicaen.fr/~chateign/danielc/abstracts/Lutterotti_abstract_RXMatiere2013_FPSM.pdf [Last accessed 10 April 2017].

Meyer, G R, Aragão, D, Mudie, N J, Caradoc-Davies, T T, McGowan, S, Bertling, P J, Groenewegen, D, Quenette, S M, Bond, C S, Buckle, A M and **Androulakis, S** 2014 Operation of the Australian Store. Synchrotron for macromolecular crystallography. *Acta Crystallogr*, D70: 2510–2519. DOI: https://doi.org/10.1107/S1399004714016174

Meyer, P A, et al. 2016 Data publication with the structural biology data grid supports live analysis. *Nature Commun*, 7: 10882. DOI: https://doi.org/10.1038/ncomms10882

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

Montelione, G T, Nilges, M, Bax, A, Guntert, P, Herrmann, T, Richardson, J S, Schwieters, C D, Vranken, W F, Vuister, G W, Wishart, D S, Berman, H M, Kleywegt, G J and **Markley, J L** 2013 Recommendations of the wwPDB NMR Validation Task Force. *Structure*, 21(9): 1563–1570. DOI: https://doi.org/10.1016/j.str.2013.07.021

Narayanan, B C, Westbrook, J, Ghosh, S, Petrov, A I, Sweeney, B, Zirbel, C L, Leontis, N B and **Berman, H M** 2013 The Nucleic Acid Database: new features and capabilities. *Nucleic Acids Res*, 42: D114–D122. DOI: https://doi.org/10.1093/nar/gkt980

Online Dictionary of Crystallography 2017 Main Page: crystallography. Available at: http://reference.iucr.org [Last accessed 10 April 2017].

OpenAIRE 2017 Zenodo – New and Improved! Available at: https://www.openaire.eu/zenodo-relaunch [Last accessed 10 April 2017]

Pepponi, G, Gražulis, S and **Chateigner, D** 2012 MPOD: A Material Property Open Database linked to structural information. *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms*, 284: 10–14. DOI: https://doi.org/10.1016/j.nimb.2011.08.070

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

Rajan, H, Uchida, H, Bryan, D, Swaminathan, R, Downs, R and **Hall-Wallace, M** 2006 Building the *American Mineralogist* Crystal Structure Database: A recipe for construction of a small Internet database. In: Sinha, A (Ed.), *Geoinformatics: Data to Knowledge*, Geological Society of America. DOI: https://doi.org/10.1130/2006.2397(06)

Read, R J, Adams, P D, Arendall III, W B, Brunger, A T, Emsley, P, Joosten, R P, Kleywegt, G J, Krissinel, E B, Lutteke, T, Otwinowski, Z, Perrakis, A, Richardson, J S, Sheffler, W H, Smith, J L, Tickle, I J, Vriend, G and **Zwart, P H** 2011 A new generation of crystallographic validation tools for the protein data bank. *Structure*, 19(10): 1395–1412. DOI: https://doi.org/10.1016/j.str.2011.08.006

Reid, J R, Kaduk, J A and **Vickers, M** 2016 The crystal structure of trandolapril, $C_{24}H_{34}N_2O_5$: an example of the utility of raw data deposition in the powder diffraction file. *Powder Diffraction*, 31(3): 205–210. DOI: https://doi.org/10.1017/S0885715616000294

Scalfani, V F, Williams, A J, Tkachenko, V, Karapetyan, K, Pshenichnov, A, Hanson, R M, Liddle, J M and **Bara, J E** 2016 Programmatic conversion of crystal structures into 3D printable files using *Jmol*. *J. Cheminform*, 8: 66. DOI: https://doi.org/10.1186/s13321-016-0181-z

Science International 2015 *Open Data in a Big Data World*. Paris: International Council for Science. Available at: http://www.icsu.org/science-international/accord [Last accessed 10 April 2017].

Science & Technology Facilities Council 2011 ISIS data management policy. Available at: http://www.isis.stfc.ac.uk/user-office/data-policy11204.html [Last accessed 10 April 2017].

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 their anisotropic physical properties. *Crystal Research and Technology*, 50: 432–441. DOI: https://doi.org/10.1002/crat.201400469

The UniProt Consortium 2015 UniProt: a hub for protein information. *Nucleic Acids Research*, 43: D204–D212. DOI: https://doi.org/10.1093/nar/gku989

University of Manchester 2013 University of Manchester Research Data Management Policy. Available at: http://www.library.manchester.ac.uk/using-the-library/staff/research/services/research-data-management/policy [Last accessed 10 April 2017].

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*, B58: 343–348. DOI: https://doi.org/10.1107/S0108768102002902

Wilcox-O'Hearn, Z and **Warner, B** 2008 Tahoe: The Least-authority Filesystem. 21–26. Available at: https://gnunet.org/sites/default/files/lafs.pdf.

Wishart, D S, Knox, C, Guo, A, C, Shrivastava, S, Hassanali, M, Stothard, P, Chang, Z and **Woolsey, J** 2006 DrugBank: a comprehensive resource for *in silico drug* discovery and exploration. *Nucleic Acids Res.*, 34 Suppl. 1 (Database issue): D668–D672. DOI: https://doi.org/10.1093/nar/gkj067

wwPDB 2015 Data correspondences between the PDB and CSD archives now available. Available at: http://wwpdb.org/news/news?year=2015#29-July-2015 [Last accessed 10 April 2017].

How to cite this article: Bruno, I, Gražulis, S, Helliwell, J R, Kabekkodu, S N, McMahon, B and Westbrook, J 2017 Crystallography and Databases. *Data Science Journal,* 16: 38, pp. 1–17, DOI: https://doi.org/10.5334/dsj-2017-038

Submitted: 14 October 2016 **Accepted:** 04 July 2017 **Published:** 07 August 2017

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS