Data Science Skills in Publishing: for authors, editors and referees
A Satellite Workshop to the 32nd European Crystallographic Meeting

Organised by CommDat (the IUCr Committee on Data)

There is a trend towards ensuring that modern science research data are findable, accessible, interoperable and reusable (FAIR). However, this is something that crystallographers have been achieving for many decades, during which excellent crystallographic databases have always exploited the best available hardware for digital archiving. FAIR is necessary but not sufficient, as physicists would say, as the archived data should also be true facts. So FACT and FAIR are needed for reproducibility. The crystallographic community has developed automatic checking software by pooling its experiences from hundreds of thousands of crystal structure analyses into validation procedures with numerous data file checks on both coordinates and processed diffraction data sets. Alarm alerts can then be scrutinised by journal editors and referees. With such exemplary procedures is there anything to be improved? Crystallographers conclude that there is. Firstly the IUCr journal *Acta Cryst. C: Structural Chemistry* has always required submission of article with validation report with underpinning data files. Thus the specialist subject expertise of referees can involve their own direct calculations to supplement the automatic checks before article and data set acceptance as versions of record by the editor. This has inspired others to look to improve their own crystallographic disciplines and journals to follow the *Acta Cryst. C* standard. Secondly the digital archives have enhanced their capacity in recent years owing to amazing hardware advances so that even the Gigabyte-sized raw data sets can also be preserved as versions of record. A reader of a publication can thereby revisit even the earliest calculation decisions of the authors of a publication. As the Royal Society of London puts it: science is about not taking someone's word and so, instead, the science is always in the data. FACT and FAIR, indeed scientific objectivity itself, is possible. This Workshop will address the state of the art in the field and the data science skills hoped for, indeed to be expected, of all those involved in publishing crystallography results, and of results from all the cognate methods such as scattering, microscopy and spectroscopy.

Timetable

Session I: The *checkCIF* paradigm

Chair: Annalisa Guerri

8.25 am Introduction to the Workshop
John R. Helliwell
School of Chemistry, University of Manchester, M13 9PL, UK

8.30 am Data refereeing and editing in chemical crystallography; the *Acta Cryst. C* experience
Anthony Linden
Department of Chemistry, University of Zurich, Switzerland

9.00 am The vital role of Crystallographic Information Files in chemical and biological crystallography to underpin the databases' validation reports
Brian McMahon
IUCr, 5 Abbey Square, Chester CH1 2HU, UK

9.30 am *PLATON* and raw diffraction data opportunities for chemical crystallography publishing
Ton Spek
Utrecht University, The Netherlands

10.00 am Coffee break

Session II: Beyond chemical crystallography

Chair: Brian McMahon

10.30 am The role of raw powder diffraction data in peer review; past, present and future
Miguel Aranda
Departamento de Química Inorgánica, Universidad de Málaga, 29010 Málaga, Spain
ALBA Synchrotron, Carrer de la Llum 2–26, 08290 Barcelona, Spain

11.00 am Diffraction Data Deposition and Publication
Kay Diederichs[1] and Manfred Weiss[2]
[1] *University of Konstanz, D-78457 Konstanz, Germany*
[2] *Macromolecular Crystallography (HZB-MX), Helmholtz-Zentrum Berlin, Albert-Einstein-Str. 15 D-12489 Berlin-Adlershof, Germany*

11.30 am Raw data opportunities for biological crystallography publishing
Loes Kroon-Batenburg
Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands

12.00 noon Lunch break

Session III: Enhancing the scientific record

Chair: Simon Coles

1.00 pm Correcting the public record of chemical crystallography science
Simon Coles[1], Suzanna Ward[2] and Carl Schwalbe[2,3]†
[1] *Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK*
[2] *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK*
[3] *School of Life & Health Sciences, Aston University, Birmingham B4 7ET, UK*

1.30 pm Correcting the public record of biological crystallography science
Mariusz Jaskólski
Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland

2.00 pm Overview of the role of data reviews and tutorial reviews in improving crystallographic science training
Petra Bombicz
Research Laboratory of Chemical Crystallography, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar Tudósok körútja 2, H-1117 Budapest, Hungary

2.30 pm Break

Session IV: Future prospects

Chair: Brian McMahon

2.45 pm Towards a human and machine-readable scientific literature
Simon Billinge
Department of Applied Physics and Applied Mathematics, Columbia University, 200 Mudd, 500 W 120th Street, New York, NY 10027, USA

3.15 pm *IUCrData* – update on data publication and practices at the IUCr
Gillian Holmes
IUCr, 5 Abbey Square, Chester CH1 2HU, UK

3.45 pm Overview of the new opportunities in and a harmonisation of peer review of 'data with validation report with article narrative' practices
John R. Helliwell
School of Chemistry, University of Manchester, M13 9PL, UK

4.00 pm General Discussion

4.30 pm Tea

5.00 pm Close of Workshop

6.00 pm ECM32 Opening Ceremony
University of Vienna

Posters

Data and Data Science at the Royal Society of Chemistry
Rita Giordano, Colin Batchelor and John Boyle
Royal Society of Chemistry, Thomas Graham House (290), Science Park, Milton Road, Cambridge CB4 0WF, UK

Using the CSD to increase data science skills in the publication of crystallographic data
Suzanna C. Ward, Natalie T. Johnson and Amy Sarjeant
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK

Towards data standards for pressure measurement
Kamil Filip Dziubek
LENS – European Laboratory for Non-Linear Spectroscopy, Via Nello Carrara 1, 50019 Sesto Fiorentino, Italy

† *We are sorry to record that Carl Schwalbe, who was originally to give this presentation, died on 1 August 2019. We are grateful to Simon Coles and Suzanna Ward for reporting on work in this area that they undertook with Carl.*

We are also grateful to Kay Diederichs for agreeing to present on behalf of Manfred Weiss, who has had to withdraw at short notice.

Abstracts

Introduction to the Workshop

John R. Helliwell

School of Chemistry, University of Manchester, M13 9PL, UK
Email: john.helliwell@manchester.ac.uk

I will provide an introduction to the Workshop. The origins were some 20 years ago when I witnessed daily an *Acta Cryst. C* Co-editor undertaking her editorial work evaluating, with referees, each submitted article's narrative, *checkCIF* report, underpinning data and the authors' finalised model before letting them pass as 'version of record'. To reach that acceptance stage modifications were often required in those data as well as the article because of the specific nature of each study and/or the relative inexperience of the authors. I was Editor-in-Chief and thought that such exemplary work within *Acta Cryst. C* should obviously be replicated across all of crystallography. In the mid-2000s with IUCr Chester we surveyed this topic and considered archiving and use of raw diffraction data [1].

[1] Strickland, P., McMahon, B. & Helliwell, J. R. (2008). Integrating research articles and supporting data in crystallography. *Learned Publishing,* **21**, 63–72.

Data refereeing and editing in chemical crystallography: the *Acta Cryst. C* experience

Anthony Linden

Department of Chemistry, University of Zurich, Switzerland
Email: anthony.linden@chem.uzh.ch

The IUCr journals have a continuous history of archiving data in one form or another; in the early days, tables of observed and calculated structure factors were published as part of the paper. The utility of such data was rather limited; considerable effort was required to convert the archived record into something that could be used. Most other publishers eventually turned their back on the archiving of structure factors, relegating this duty to the authors, which has undoubtedly led to the unfortunate loss of very many data sets. The advent of CIF and then electronic validation of CIF content during the 1990s was a valuable step forward, which facilitated easier archiving and more consistent evaluation by reviewers and Co-editors. Even then, IUCr Journals remained almost the only publisher using structure factors during reviewing until the CCDC began accepting them about 10 years ago. The advent of area detector diffractometers in the 1990s gave us raw frame data which contains much information beyond that which is routinely extracted. Access to such data has been helpful during some reviews, but was only asked for in special circumstances, mainly because of logistics. The capacity and accessibility of digital archives are now reducing the hurdle to reviewing and preserving raw data. In the IUCr Journals experience, authors often show resistance when new requirements are introduced. Busy authors will embrace the FAIR principles if the repositories, and the deposition, access and data extraction procedures are as routine, transparent and automatic as possible.

The vital role of Crystallographic Information Files in chemical and biological crystallography to underpin the databases' validation reports

Brian McMahon

IUCr, 5 Abbey Square, Chester CH1 2HU, UK
Email: bm@iucr.org

The Crystallographic Information File (CIF) was introduced in 1991 to facilitate data exchange between computer programs [1], but it soon became apparent that its most important feature was the *precise* definition of the data items that were manipulated by the programs. With standardisation of terminology across essentially all crystallographic software came the prospect of validating any structure, and indeed, to some extent, the experimental data upon

which it was modelled. Structural databases already carried out extensive validation in their curating of stored data sets [2], but the development of protocols such as *checkCIF* [3] permitted extensive validation and evaluation of quality by the journals and indeed by the submitting author. The recent incorporation of software methods in the CIF dictionaries *via* the DDLm protocol [4] opens the door to even greater automation in both quality control and information retrieval from any solved crystal structure.

[1] Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The crystallographic information file (CIF): a new standard archive file for crystallography. *Acta Cryst.* A**47**, 655–685.

[2] See, for example, Karen, V. L. & Mighell, A. (1996). Special Issue: NIST Workshop on Crystallographic Databases. *J. Res. Natl Inst. Sci. Technol.* **101**, 205–381; Allen, F. H. & Glusker, J. P. (2002). Crystallographic Databases. Joint special issue. *Acta Cryst.* B**58**, 317–422; *Acta Cryst.* D**58**, 879–920.

[3] Spek, A. L. (2009). Structure validation in chemical crystallography. *Acta Cryst.* D**65**, 148–155.

[4] Spadaccini, N. & Hall, S. R. (2012). DDLm: a new dictionary definition language. *J. Chem. Inf. Model.* **52**, 1907–1916.

PLATON and raw diffraction data opportunities for chemical crystallography publishing

Ton Spek

Utrecht University, The Netherlands
Email: a.l.spek@uu.nl

A published crystal structure represents its author's interpretation of the underlying experimental diffraction data. Qualifiers such as 'best attainable' and 'sufficient for the purpose of the study' are often lost or neglected by the users of the archived data. It is good scientific practice not only to archive the pertinent results of a structure determination but also to archive the primary data and procedures followed to obtain the reported results. This gives the option to investigate unusual reported results or to use and improve the analysis of the data for a purpose unrelated to that of the original author. The experimental data might be unique and not easily obtained again.

Historically, archival was realised with the deposition of printed F_{obs}/F_{calc} tables along with a published paper. That practice was later dropped by most journals in view of its limited practicality, even after the upcoming option for the deposition of 'FCF' files in computer readable format. More recently, with the introduction of the computer readable CIF style of crystal structure deposition, it became mandatory to include the unmerged reflection data in that file. This now offers the option to improve on the published results, to investigate unusual issues, to detect errors or to flag reports based on faked data.

A significant issue is still the processing of the diffraction images into the set of *hkl* data. A lot of issues with a reported structure may be resolved only with the availability of the diffraction images. Knowledge about the presence of streaks, unindexed diffraction spots *etc.* in the images may be key information. Either archiving the diffraction images themselves or providing an automatic report on special features in the images along with the details of the image processing might be made mandatory.

The role of raw powder diffraction data in peer review; past, present and future

Miguel Aranda

Departamento de Química Inorgánica, Universidad de Málaga, 29010 Málaga, Spain
Email: g_aranda@uma.es
ALBA Synchrotron, Carrer de la Llum 2–26, 08290 Barcelona, Spain

Scientific data in our community can be classified in three broad categories: raw, reduced and derived data. IUCr has been very active in promoting the sharing of reduced and derived data for decades in independently verified databases. The need for raw data sharing is clearly increasing, being nowadays technically feasible and likely cost-effective.

The powder diffraction (PD) community is a subgroup of the crystallographic community dealing with several goals, mainly (1) average crystal structure determination, (2) quantitative phase analyses, (3) microstructural analyses, and (4) local structure determination and quantitative analyses of nanocrystalline materials. For PD, derived data for objectives (2) and (3) and to a large extent (4) cannot be incorporated in 'standard' databases. Derived data are not independently validated, and therefore, in my opinion, the need for sharing raw PD data is even more compelling than that of sharing raw single-crystal diffraction data.

So, if raw diffraction data sharing is approaching, we have the responsibility to ensure that this action is useful. Hence, and as stated by John Helliwell in the introduction of this Workshop, two conditions must be fulfilled. On the one hand, and from the computing point of view, the shared data must be findable, accessible, interoperable and reusable — *i.e.* comply with FAIR standards. However, this is necessary but not sufficient. On the other hand, and from the point of view of the scientific community involved, the shared data must have sufficient quality. They must be true facts and the 'FACT and FAIR' term has been coined.

By incorporating raw PD data 'check/validation' in the peer review process, the FACT nature of the raw data could be established. Or at least, a minimum quality level could be ensured. Some ideas (and experiences) will be developed in the meeting, including the use of shared raw PD data by meticulous reviewers. Furthermore, some ideas will be floated including the possibility of IUCr Journals requesting (confidentially) the raw data and the control file to check/verify a minimum quality of the raw data and that of the derived data. This endeavor is enormous but not unsurmountable thanks to our deep-rooted collaborative spirit. The IUCr database of referees should be updated to incorporate also the skills/experience in data analysis software, in order to ensure that the process is sustainable, *i.e.* the time required for reviewing the raw data is not excessive. How can this be automated? This is a matter for future discussion(s). Finally, it could be that use of the pair distribution function, where there is a very dominant analysis software, is a good option to test this.

Acknowledgements: Supported by MinCIU research grants, BIA2014-57658 and BIA2017-82391-R

Data accessibility. Since 2017, our research group is freely sharing all diffraction (and tomographic) data at Zenodo. The DOIs for the data sets will be given wherever appropriate.

Diffraction Data Deposition and Publication

Kay Diederichs[1*] and Manfred Weiss[2]

[*] Presenting.

[1] *University of Konstanz, D-78457 Konstanz, Germany*
[2] *Macromolecular Crystallography (HZB-MX), Helmholtz-Zentrum Berlin, Albert-Einstein-Str. 15, D-12489 Berlin-Adlershof, Germany*

Recently, several current and former editors of IUCr journals have suggested that editors should begin to encourage authors of manuscripts describing a macromolecular structure to supplement their manuscript with the underlying diffraction data, *i.e.* the unprocessed diffraction images. An important question to address here is, how can this improve the publication process and what mechanisms need to be put in place to maximize the impact?

Data processing, *i.e.* the step from raw diffraction images to a reduced list of merged or unmerged intensities, is nowadays done mostly automatically. Quite frequently, authors do not even visually inspect diffraction images anymore. This means that several things might happen. Weak reflections indicating a superlattice or an incommensurate structure might be overlooked, the data might be processed assuming too high symmetry or two low symmetry depending on the thresholds set in the automated processing pipelines, the presence or absence of twinning might be misjudged, *etc.* The data processing results are then commonly reported in a crystallographic *Table 1*, the main function of which is to support the notion that the data set has been processed expertly and competently, and that the data quality is sufficient to support the claims made in the paper.

However, given the potential pitfalls of automated data processing and the serious limitations imposed by *Table 1*, it seems like a good idea to give editors, referees and later on readers of the paper access to the underlying diffraction images, just in case some doubts arise during the paper handling process and after. But the data alone will not suffice. Along with the data and the associated metadata it will be necessary to provide some mechanisms of analysis, whether this be a checklist or something else. Some ideas along these lines will be discussed.

Raw data opportunities for biological crystallography publishing

Loes Kroon-Batenburg

Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands
Email: l.m.j.kroon-batenburg@uu.nl

Automation in macromolecular X-ray crystallography has proceeded to a level where high-throughput data collection and structure solution is feasible. In general, structural biologists can get X-ray diffraction data from their crystals in less than a minute at synchrotron installations, and since recently the data can be collected in remote sessions, without the need to travel to the synchrotron beamline. To check the usefulness of the data, they are auto-processed through a multidata-processing pipeline, often resulting in a .mtz type file containing reflection data ready for structure solution. The user then takes these reflection data home and goes on to solving the structure without having a further look at the diffraction images. How many users actually copy the data to their home computers and inspect their diffraction images in detail? And how many users make the data available to the public?

Validation tools are set in place to assess the quality of the structural model in the light of the data, which itself is given statistical metrics. There could be various reasons for wanting to reassess the data processing. For example, the raw data sets could be valuable for developing new methods of structure determination; they could be important for validating the interpretation of structural features; or for allowing data analysis at higher resolution than used in the original work, searching for the presence of multiple lattices present in a crystal, or deducing details of correlated motions or disorder from the diffuse scattering that is largely ignored in determining Bragg peak positions and characteristics. Funding agencies are adopting Open Science philosophies, requiring researchers to make available all scientific data following the FAIR principles.

The current state of technology is such that raw crystallographic diffraction data can be preserved and made public via different types of archive, such as at universities, discipline-specific repositories (Integrated Resource for Reproducibility in Macromolecular Crystallography, Structural Biology Data Grid), general public data repositories (Zenodo, ResearchGate) and centralized neutron and X-ray facilities. Some synchrotron facilities have committed themselves to storing raw diffraction images for ever (probably meaning at least 10 years) and also to release the data to the public after an embargo period. Formulation of improved metadata descriptors for the raw data types is in progress and is an essential prerequisite for the reusability of the data.

Correcting the public record of chemical crystallography science

Simon J. Coles[1*], Suzanna Ward[2*] and Carl Schwalbe[2,3]†

* Presenting.

[1] *Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, UK*
[2] *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK*
[3] *School of Life & Health Sciences, Aston University, Birmingham B4 7ET, UK*

† Obit. 1 August 2019.

Crystal structures are rightly deemed the 'gold standard' for small-molecule structure determination. The comprehensive tests embodied in *checkCIF* should highlight errors prior to deposition or publication. Nevertheless, errors appear even in some recent structures [1]. While the increasingly frequent deposition of structure factors along with CIFs enables retrospective detection and correction, many entries in the Cambridge Structural Database lack them; and geometrical features can provide the best clue to problems.

Missing symmetry used to occur regularly, but diligent checking by Richard Marsh and colleagues raised awareness among authors frightened of 'getting Marshed'. They found that 10% of structures described in space group *P*1 in both 1999 and 2004 needed revision, but the proportion had dropped below 2% in 2010–2013. Feeble interaction with X-rays makes H atoms difficult to locate [1]. Modern software conveniently places them in calculated positions without requiring careful inspection of a difference electron density map. Misplaced H atoms in carboxylic acids

are detectable by a number of geometrical clues: C—OH bonds shorter than C=O, clashing H atoms, OH groups lacking hydrogen-bond acceptors. *checkCIF* effectively picks up such clues. Imidazole rings with one N: and one N—H group as found in the amino acid histidine and the drug cimetidine are more ambiguous because the electron density of unprotonated N: resembles that of N—H. Furthermore, N—H in one ring often forms an easily reversible hydrogen bond to N: of an adjacent ring. Descriptors based on distances from the intervening C atom to N: and NH along with bond angles at these N atoms are the best guide. Even more difficult are 1,2,4-triazole structures as in the parent compound the C=N and C—N distances differ by only 0.009 Å, and only bond angles provide a clue to protonation state.

[1] Schwalbe, C. H. (2018). *Crystallogr. Rev.* **24**, 217–235.

Correcting the public record of biological crystallography science

Mariusz Jaskólski

Department of Crystallography, Faculty of Chemistry, A. Mickiewicz University Center for Biocrystallographic Research, Institute of Bioorganic Chemistry, Polish Academy of Sciences, Poznań, Poland
Email: mariuszj@amu.edu.pl

A number of reports have warned that a growing fraction of biomedical research cannot be reproduced. Structural biology, traditionally seen as the gold standard in life science, is not without sin as problematic models end up in the PDB (Protein Data Bank), contaminate the literature and often frustrate other efforts, such as drug design. The errors are most frequent at the macro-/small-molecule interface and are manifested by suboptimal/unsatisfactory/wrong modeling of ligands in electron density maps. We have analyzed several classes of PDB models and found errors of various severity, as well as serious ripple effects in the scientific literature. In many cases the models could be significantly corrected and redeposited in the PDB. We always contacted the original authors, with reaction ranging from hostility to fruitful cooperation. The open questions include the treatment of the new deposits and retraction of the invalidated papers.

Overview of the role of data reviews and tutorial reviews in improving crystallographic science training

Petra Bombicz

Research Laboratory of Chemical Crystallography, Research Centre for Natural Sciences, Hungarian Academy of Sciences, Magyar Tudósok körútja 2, H-1117 Budapest, Hungary
Email: bombicz.petra@ttk.mta.hu

Crystallography Reviews publishes a range of review types, including tutorial reviews and, most recently, data reviews. These offer direct educational training and, for the latter especially, insights for readers at all career stages to the collections of published crystallographic data. A unique feature of the full reviews is that their total length can be around 25 000 words and occasionally even longer, up to 35 000 words or so, *i.e.* on a scale not published by other journals. This places them at a sort of typical book size (although obviously the range of book word lengths varies a lot). The data reviews in particular illustrate the current and past publication practices of our crystallographic fields and subfields. Most telling are general data weaknesses and, inevitably, specific error cases that are found; identifying possibilities of better practice are described. After a general introduction I will describe some *Crystallography Reviews'* tutorial and data reviews with special relevance to improving the publishing of data with articles. Some specific authors' conclusions will be mentioned.

Towards a human and machine-readable scientific literature

Simon Billinge

Department of Applied Physics and Applied Mathematics, Columbia University, 200 Mudd, 500 W 120th Street, New York, NY 10027, USA
Email: sb2896@columbia.edu

The scientific literature has served humankind valiantly for more than 350 years, resulting in scientific and technological revolutions in areas from energy and food to healthcare and the environmnent, resulting in a life for modern humans that would be completely unrecognizable to people in the late 17th century when the first journals were founded. This happened following many millennia of human development over which life changed little. What was the recipe for such success? The basic recipe is to create incentives to share knowledge freely, and to have safeguards in place that ensure standards of quality and veracity in the knowledge base are maintained. The latter was provided by peer review, a somewhat strained and maligned activity these days, but its importance in getting us to where we are today should not be undervalued, and moving forward, not undermined. As we move into the 21st century this structure is under strain. The sheer amount of information being generated is unprecedented, making quality assurance a challenge, and some of the incentives have become distorted, such that weaving good yarns in high profile journals is valued over careful and thorough scholarship. Leaving aside the last point, a new development is that increasingly it is not only humans that are the target audience of the scientific journals, but also computers. In the past, sharing knowledge openly meant writing words on paper, placing numbers in tables and representing data in pretty figures that conveyed meaning to a human. We are gradually training machines to read our literature through the process of Natural Language Processing, but I would rather take a different view which is to ask whether we can make our journals in a language and form that may be read by machines in the first place. This is not to say that they won't also be read by humans, but I see a critical need to develop journals whose product, the journal articles, may be read by both humans and machines, without the need to teach the machines how to read a human language. This is similar to the FACT and FAIR concepts at the heart of this workshop, but this way of thinking can go beyond parallel universes of databases and journal articles to something that blends these together and results in a more reproducible and more easily shared and reused scientific literature for the 21st century.

IUCrData – update on data publication and practices at the IUCr

Gillian Holmes

IUCr, 5 Abbey Square, Chester CH1 2HU, UK
Email: checkin@iucr.org

IUCrData, the IUCr's open-access data publication, was launched in 2016 with the aim of providing short descriptions of crystallographic data sets and data sets from related scientific disciplines. The first phase of this venture provides a home for the short crystal structure reports previously published in *Acta Cryst. E: Crystallographic Communications*. It provides a mechanism to publish structures that might otherwise not be available to the scientific community, as well as ensuring that peer-reviewed information on uniquely determined structures is readily accessible.

The IUCr has always seen the need for archiving data and literature together, to preserve the complete record of the science. The journals require deposition of supporting data sets for published crystal structures. These data sets are thoroughly assessed as part of the publication process. IUCr policy has always been to allow free access to data, including coordinates, anisotropic displacement parameters and structure factors. The development of CIF and *checkCIF* has served to make critical crystallographic data more consistently reliable and accessible at all stages of the information chain.

One possible route for the development of *IUCrData* as a service is motivated by the desire to maximise the potential of the IUCr data archive. It would be useful to develop an application programming interface (API) keyed on DOI that could load data in multiple formats, including CIF and alternative formats, and retrieve data sets matching specific criteria. IUCr Chester is actively collaborating in one pilot project to create a prototype data-harvesting API that may guide the direction of development effort in related fields.

Overview of the new opportunities in and a harmonisation of peer review of 'data with validation report with article narrative' practices

John R. Helliwell

School of Chemistry, University of Manchester, M13 9PL, UK
Email: john.helliwell@manchester.ac.uk

Observing the work of an *Acta Cryst. C* Co-editor directly for several years (my wife and colleague Dr Madeleine Helliwell) inspired me to look to improve the IUCr biological journals to follow the *Acta Cryst. C* standard. My effort as Editor-in-Chief at the IUCr Congress in Geneva in 2002 was not approved at the Open Commission of Biological Macromolecules meeting. Later I realised, in my refereeing, that I should insist on having from the authors, via the editor, the to-be-released PDB coordinates and structure factors as well as the PDB Validation Report and article. [1] If I did not receive these then I would not undertake the refereeing task, of course promptly informing the editor. My whole refereeing experience was transformed, being much more thorough and satisfying. Basically I find that the Validation Report is invaluable in checking the general aspects of a PDB deposition and my refereeing of the data, with my own calculations, addresses the specific aspects of a study. Meanwhile a new theme involves the raw data owing to amazing hardware advances. So Gigabyte-, even Terabyte-, sized raw data sets can be preserved. A reader of a publication can revisit the earliest calculation decisions of the authors. My reaction to this new refereeing opportunity over the last two years has been to recommend that authors improve their raw data processing, rather than my moving their diffraction images onto my computer and reprocessing them [1]. Overall, IUCr Journals policy on raw diffraction data is being informed by the IUCr Commissions' reactions to the IUCr Diffraction Data Deposition Working Group's Final Report [2]. The first such formal change to Notes for Authors is for biological crystallography [3].

[1] Helliwell, J. R. (2018). Data science skills for referees: biological X-ray crystallography. *Crystallogr. Rev.* **24**, 263–272.

[2] https://www.iucr.org/resources/data/dddwg/final-report (2017). Authored by the Members of the DDDWG.

[3] Helliwell, J. R., Minor, W., Weiss, M. S., Garman, E. F., Read, R. J., Newman, J., van Raaij, M. J., Hajdu, J. & Baker, E. N. (2019). Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography. *IUCrJ*, **6**, 341–343.

Posters

Data and Data Science at the Royal Society of Chemistry

Rita Giordano, Colin Batchelor and John Boyle

Royal Society of Chemistry, Thomas Graham House (290), Science Park, Milton Road, Cambridge CB4 0WF, UK

We show how the Royal Society of Chemistry deals with crystallographic data during peer review in the context of our evolving data publishing strategy. We discuss some recent prototypes of broader data sharing in chemistry publication and give examples of recent research by the Data Science team and how they support the publishing process.

Using the CSD to increase data science skills in the publication of crystallographic data

Suzanna C. Ward, Natalie T. Johnson and Amy Sarjeant

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK
Email: ward@ccdc.cam.ac.uk, njohnson@ccdc.cam.ac.uk

Every day, a vast number of datasets are generated by researchers globally across many different disciplinary domains. Collectively, these constitute an ocean of data containing a wealth of novel insights just waiting to be discovered. FAIR data principles have been established by the data community in order to ensure data is made intelligible to humans and machines so that these insights can be revealed.

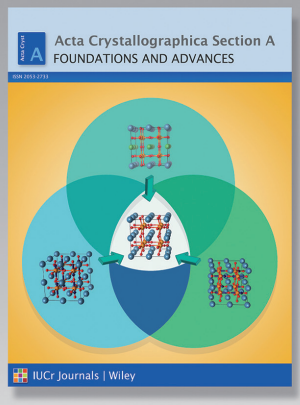
Thankfully, crystallographers recognised the value of collecting and sharing datasets early on and the first data repositories were set up over half a century ago. Today there are a handful of established crystallographic databases that collectively share the entire collection of published structures enabling scientists to derive valuable insights from this data worldwide. The Cambridge Structural Database (CSD) is one such database containing over one million organic and metal–organic crystal structures. This wealth of information can be used to check the validity of new structures and the CSD is already being used by crystallographers worldwide in small-molecule and protein structure validation. One key area where databases can help the crystallographic community set standards is through interactive deposition processes. At the CCDC we have made changes to our online deposition service to facilitate this and in partnership with FIZ Karlsruhe have extended this service to cover data in the Inorganic Crystal Structure Database (ICSD). This poster will explore the role databases can play to help the crystallographic community comply with FAIR data principles and to set new standards from structure validation to publication.

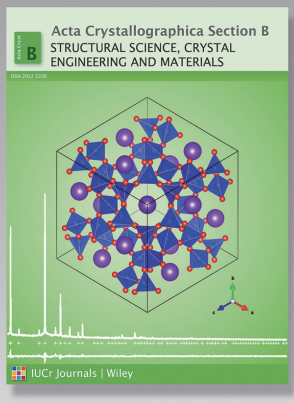
Towards data standards for pressure measurement

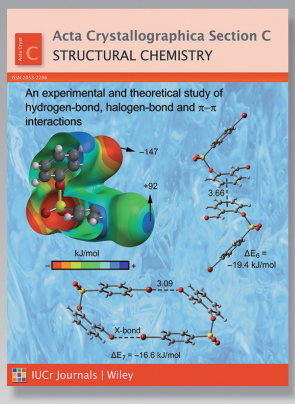
Kamil Filip Dziubek

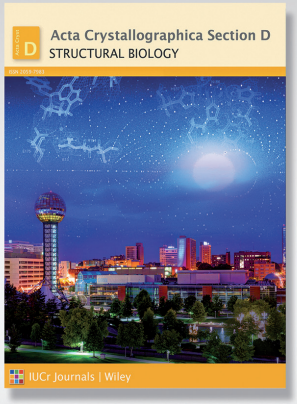
LENS – European Laboratory for Non-Linear Spectroscopy, Via Nello Carrara 1, 50019 Sesto Fiorentino, Italy
Email: dziubek@lens.unifi.it

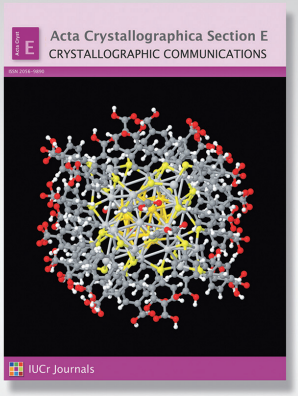
There are two thermodynamic parameters, temperature and pressure, that can be varied independently in non-ambient crystallography experiments. Temperature measurement usually is routine, as it simply corresponds to the reading of a thermocouple positioned at a cryostream nozzle, inside a cryostat or a furnace. By contrast, pressure cannot be determined directly. This requires employing pressure sensors placed together with a sample inside an experimental chamber (usually in a diamond anvil cell). There are two main types of such pressure gauges: based on a shift of the principal fluorescence spectrum line with pressure (*e.g.* ruby) or on a pressure-tuned compression of a unit-cell volume, as determined by X-ray diffraction (*e.g.* gold). In both cases, however, some associated data (as details of the lattice parameters refinement, the center wavelength of the main peak of fluorescence at ambient and high pressure, applied pressure scale or equation of state, *etc.*) are needed to calculate the pressure value. These data items can be therefore organised, classified and stored within Crystallographic Information Framework (CIF). This presentation aims to bring forth most important data indispensable for correct pressure measurement and gives some suggestions about specific CIF data types and structures suitable to host these data.

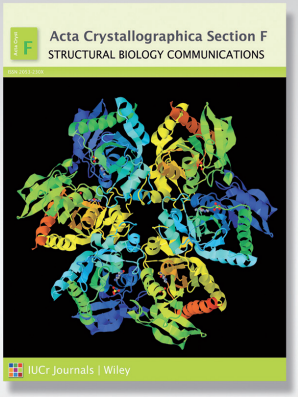


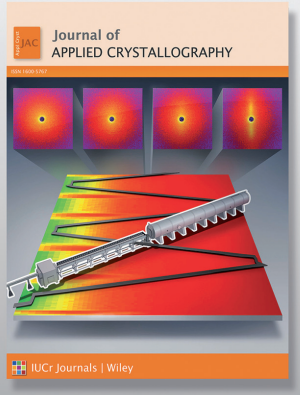


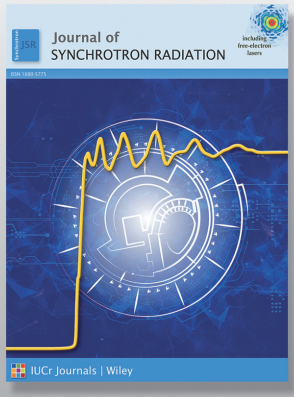


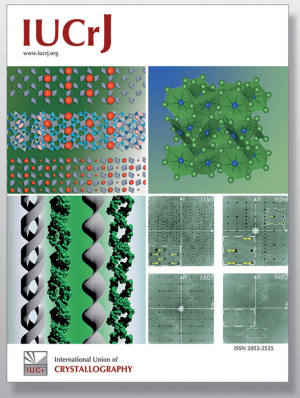












journals.iucr.org