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.

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.