Zeitschrift: Bulletin / Vereinigung der Schweizerischen Hochschuldozierenden = Association Suisse des Enseignant-e-s d'Université

Herausgeber: Vereinigung der Schweizerischen Hochschuldozierenden

Band: 46 (2020)

Heft: 2

Artikel: Open science and reproducibility of the scientific record

Autor: Helliwell, John R.

DOI: https://doi.org/10.5169/seals-966117

Open science and reproducibility of the scientific record

John R Helliwell*

Abstract

It is true to say that open science is a definite trend and reproducibility a growing concern. The former should however counter the latter. There are then complex features to the science research landscape. So, where is the process of science working well and where is it less than optimal? I will explore this landscape from my science base in crystallography and describe our ongoing initiatives. I will describe what can already be copied from crystallography as examples of best practice. The experiences of other fields of science that have come to my attention on these matters, such as at joint workshops, I will describe with some key conclusions.

1. Introduction

A great fraction of funded research is by the taxpayer and most funding agencies now espouse the principle that this research should be open. Also published research is generally assumed to be trustworthy, i.e. reproducible.

The USA National Academies of Science, Engineering and Medicine (2019) published an authoritative report on *Reproducibility and Replicability in Science*, a 256 page analysis and survey including a final chapter of 20 pages on Confidence in Science. Whilst naturally USA focused, it also had international participation in the study. Data and software transparency, i.e. openness, featured prominently in the Report in realising the best processes for achieving as high a scientific standard as possible. They made the important distinction of the two terms:-

- *Reproducibility* meaning obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis (as the original researchers).
- *Replicability* meaning obtaining consistent results across (different) studies aimed at answering the same scientific question, each of which (research team) has obtained its own data.

In between these two procedures there is possible what I call robustness checks so that an existing set of data can be assessed with a different software package. Thus a variance of results estimate is obtained indicating further the reliability of a study.

These National Academy of Science definitions are at the core of the scientific method and the definition of science itself (Latin word *scientia* meaning to *know*). If we *know* something it is no longer subjective. The closest we can come to objectivity as our evidence of what we know is the primary experimental data (the raw data). The slight caveat is that a piece of apparatus has to be calibrated by an experienced person. The publication contains a narrative, and although it includes a summary of what was done, it is necessarily the words and interpretation of the authors, i.e. subjective. Processing raw data with different softwares, thereby deriving models such as a crystal structure in my scientific field, has this possible *variance of results in the processed and derived data*. In the absence of a suitable word I offer a new word 'subjectovity', namely between the objectivity of the raw data and the subjectivity to some degree of the writer of the analysis software. Variation in analysis software performance can arise in cases such as estimating weak signals of various kinds and their uncertainty values. Sharing data to compare software packages can

* Department of Chemistry, University of Manchester, M13 9PL, UK.

E-mail: *John.helliwell@manchester.ac.uk*
https://www.research.manchester.ac.uk/portal/en/researchers/john-helliwell(8e39a79d-aa6d-4301-8b62-bf443581c009).html

ORCID 0000-0002-0520-7540

John R Helliwell, DSc (Physics, University of York), DPhil (Molecular Biophysics, Oxford University) is Emeritus Professor of Chemistry at The University of Manchester, where he served as Professor of Structural Chemistry from 1989 to 2012. Academic teaching from 1979 till 1988 was at the Universities of Keele and York. Researcher in the fields of Crystallography, Biophysics, Structural biology, Structural Chemistry and Data science. Also based at the SRS Daresbury Laboratory, in various periods of appointment between 1979 to 2008, including in 2002 as Director of Synchrotron Radiation Science based at the Council for the Central Laboratory of the Research Councils (CCLRC), Daresbury Laboratory. He is a Fellow of the Institute of Physics, the Royal Society of Chemistry, the Royal Society of Biology, and the American Crystallographic Association, an Honorary Member of the British Crystallographic Association and of the British Biophysical Society. He is a Corresponding member of the Royal Academy of Sciences & Arts of Barcelona, Spain and Honorary Member of the National Institute of Chemistry, Slovenia. His awards include the European Crystallographic Association Eighth Max Perutz Prize 2015, the American Crystallographic Association Patterson Award 2014, and the 'Professor K Banerjee Endowment Lecture Silver Medal' of the Indian Association for the Cultivation of Science (IACS) 2001. He published over 200 scientific research papers and several books, e.g. "Macromolecular Crystallography with Synchrotron Radiation" Cambridge University Press (1992), published in paperback 2005 and "Macromolecular Crystallization and Crystal Perfection" (by N E Chayen, J R Helliwell and E H Snell), Oxford University Press - International Union of Crystallography Monographs on Crystallography (2010).

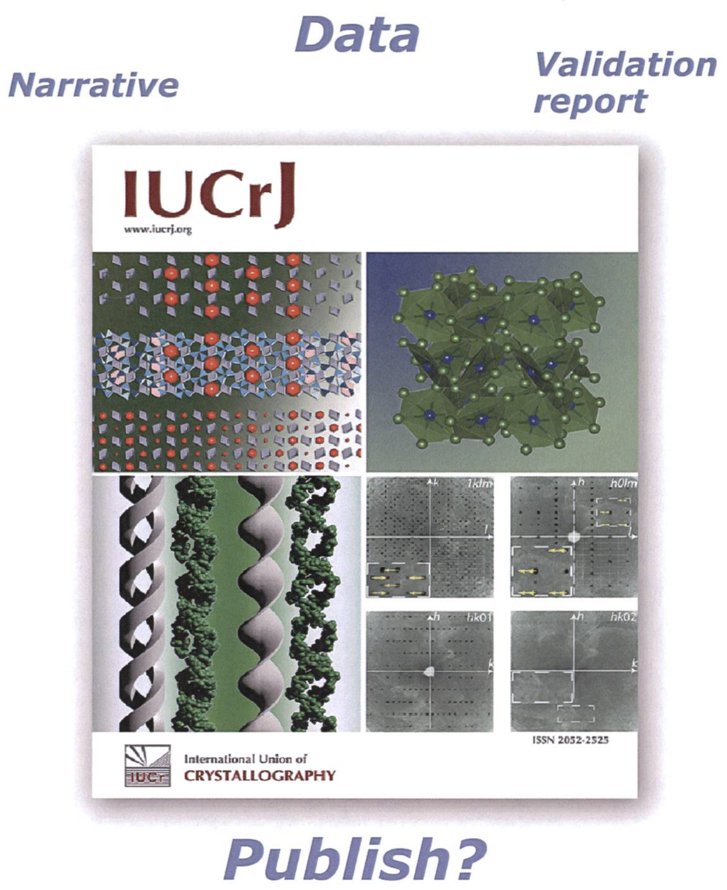

Figure 1. *The Editor as the gatekeeper of reproducible science knowledge and on which basis replicability by others, and ultimately wisdom, can rely.*

improve them and reduce variations in analysis. An early example of software improvement for processing of raw data (oscillation camera diffraction images recorded on film) was the comparative "round robin" project of Helliwell *et al.* (1981).

The National Academies Report also cogently states:

> Unhelpful sources of non-replicability can be minimized through initiatives and practices aimed at improving research design and methodology through training and mentoring, repeating experiments before publication, rigorous peer review, utilizing tools for checking analysis and results, and better transparency in reporting.

2. The open science and reproducibility landscape in crystallography

In my own field of crystallography our community prides itself on its archiving of processed diffraction data and associated derived atomic coordinates. It is therefore one of the pioneers of the FAIR principles of data use (Findable, Accessible, Interoperable and Reusable, Wilkinson *et al.* 2016). An inspirational approach for striving for reproducibility is the International Union of Crystallography's (IUCr) chemical crystallography journals' refereeing processes. IUCr publishes nine journals; an example front cover is from IUCrJ (see Figure 1). Thus an article narrative, a *checkCIF* report (more than 400 general checks of the data and metadata), and the underpinning processed diffraction data and derived coordinates data are assessed together by referees and editor. This is done in order to arrive at the *proper version of record* for each of these aspects of a study. In other areas of my science, this *openness* during the pre-publication peer review has not been so easy to secure, in my experience due to anxieties of the authors who can insist on confidentiality at that stage. Also, besides other areas of crystallography, how far might this exemplary *open* chemical crystallography procedure be extended into all other areas of science; e.g. do we just need to advertise it?

In another direction, we can ask:- are our crystallography laboratory leaders, who tend to be the ones asked for referee reports, still data analysis active? Clearly then they need to be if they agree to be a referee. Continual professional development should then include data analysis as a core learning objective. So, the IUCr Committee on Data has started running data science skills workshops such as the one at the European Crystallography Meeting held in Vienna in 2019.[1] The Vienna Workshop of course accepted participants who were at any stage of their careers. Aiming more at early career crystallographers the Associazione Italiana Cristallografia, along with several IUCr Committees, ran the School on Crystallographic information linked with data.[2]

A natural development is to extend the approach of refereeing of processed and derived crystallographic data with submitted articles to peer review of the primary, *i.e.* raw diffraction data. Thus IUCr Journals (Strickland, Helliwell and McMahon 2008) stated:

> The IUCr is beginning to consider longer-term approaches to archiving the raw data, since these often contain additional information that is not fully utilized in the process of data reduction.

I define raw data as follows:

> Those primary experimental data that have emanated from a properly calibrated device, *i.e.* detector, and which have been properly corrected for detector non-uniformities, non-linearities or spatial distortions, *etc.* Where it is a totally novel detector, of some level of unquantified performance and stability, then the data before such calibrations have been applied might need to be preserved.

1 *https://www.iucr.org/resources/data/commdat/vienna-workshop*
2 *http://www.cristallografia.org/contenuto/aics2019-crystallographic-information-fiesta/3311*

Raw data sets can however be very large compared with processed and derived data based on those raw data. Therefore this approach is pushing the limits of data archiving capacities and use of a referee's time. This also requires an expanded suite of metadata. Experiences with making diffraction image data available in crystallography are described by Kroon-Batenburg and Helliwell (2014). From 2011 to 2017 I chaired the IUCr Diffraction Data Deposition Working Group (DDDWG) and since 2017 I am Chairman of the IUCr's Committee on Data. The final report of the DDDWG[3] had as its three top recommendations:

- Authors should provide a permanent and prominent link from their article to the raw data sets which underpin their journal publication and associated database deposition of processed diffraction data (e.g. structure factor amplitudes and intensities) and coordinates, and which should obey the 'FAIR' principles, that their raw diffraction data sets should be Findable, Accessible, Interoperable and Re-usable.[4]
- A registered Digital Object Identifier (DOI) should be the persistent identifier of choice (rather than a Uniform Resource Locator, URL) as the most sustainable way to identify and locate a raw diffraction data set.
- An archive of raw diffraction data sets for currently unsolved crystal structures should be pursued.

The first two are quite general and can be applied to any field. The third can be generalised by saying that in all fields of science there are likely to be problematic cases for which opening up the data for these would unleash the knowledge and expertise of a whole research community and maybe lead to a research breakthrough.

The IUCr has provided for open discussion of all matters to do with diffraction data deposition; the DDDWG from 2011 to 2017 and Committee on Data from 2017 onwards can be found at *https://forums.iucr.org/*. Individual postings have reached several thousands of downloads each.

A well-meaning effort by funding agencies is to require preprints so as to speed up research and discovery but does not include, thus far, attachment of the data, which is de facto problematic for ensuring reproducibility. However, if preprint servers would strongly encourage a preprint to include a DOI to the underpinning archived raw data that would make preprints much more effective.

Some not-for-profit crystallography data repositories, namely the Cambridge Structure Database and the International Center for Diffraction Data, have survived through a user pays approach but is that open data? Open access has nuanced meanings here. The Protein Data Bank (PDB) and the Crystallographic Open Database (COD) are free to use databases including downloads of the whole PDB or COD. Single structures are open in the CSD. The completeness of databases, derived from the scientific literature, also varies. The value of a download of a whole database is evident for example with the PDB Redo project, now more than ten years old, which provides updated and optimised versions of existing entries of the PDB from its own DataBank. It provides then a remediation of historical entries, or of deficient in some way entries, or users can optimise their own structure model using the PDB Redo server even before deposition (Joosten *et al.* 2009, 2014). Joosten *et al.* 2009 also envisaged ultimately a role for the primary data in that process noting that

> the Joint Center for Structural Genomics archive (*http://www.jcsg.org*) is an excellent prototype of what can be achieved in this respect, and its often-acknowledged value to software developers is a clear indication of the potential benefits of such an extended deposition scheme through the dual improvements it would enable in both the results for structures already solved and in the ability to solve new more difficult ones in the future thanks to better tested and better validated software advances.

A detailed presentation and overview of the trust in crystallographic data, in powerpoint slides style, is available (McMahon, Helliwell and Hester 2018).

3. The situation in some other science fields and special cases

Allison *et al.* (2016), who are a group of researchers working on obesity, nutrition and energetics, have emphasised how pre-publication peer review of an article with data is far superior to their post-publication peer review.

Anonymity of medical data is obviously needed for respecting proper data ethics. This leads to the oft-quoted statement about open science and open data that they should be as open as possible but closed where necessary. In the reporting of clinical studies, medical journals can then employ their own data statistician in the editorial office so as to vet the data accompanying the article. A discussion of this was

3 *https://www.iucr.org/resources/data/dddwg/final-report*
4 *https://www.force11.org/group/fairgroup/fairprinciples*

held in 2018 at a workshop entitled "Transparency, Recognition and Innovation in Peer Review in the Life Sciences".[5] This Workshop considered the following issues:

- Should journal peer review become a transparent and citable form of scholarly communication?
- Should scientists receive credit for peer review and, if so, how might this be achieved?
- What are best practices in peer review, how can they be spread? How can we train scientists in scholarly review?
- Is it possible to overcome inefficiencies and redundancies in peer review?
- Should reviewers be expected to review supporting datasets and code?
- Using new tools (e.g. preprints and the internet), are there new models for feedback/evaluation that could augment traditional peer review?

In this Life Science workshop there were commonalities of approach to those I have described in detail for crystallography. The responses of attendees to some of those questions depended strongly on the age of the person. A most prominent point made was by early career researchers that they wished to retain their anonymity. Otherwise they feared retribution from their seniors.

My own contribution as a blog in preparation for the Life Sciences workshop can be found in Helliwell (2017). My conclusions included:

> *So, what should publishers and funders be doing?*
> Clearly, as well as requiring data availability after publication, they should make clear in their policies that they value peer reviewers who sincerely try to attest to a submitted article's conclusions by truly having access to the underpinning data.
>
> *What should professional associations be doing?*
> They should ensure training courses are given which are properly examined at the end of a course. These should be available to early, mid and late career researchers. I realise now, in my retraining and reskilling I have undertaken in my retirement, that in my late career I had become a manager of research with old, rather out of date, data skills. Continual professional development for a researcher is important at all stages of one's career not least if they are going to accept to do **proper peer review including the data.**

My article attracted one comment Lawrence (2018):

> Great piece John, and completely agree with you that it is very hard to see how a referee can judge the conclusions of a paper without seeing the underpinning data. Just to add a point of clarity about Wellcome Open Research (and the other Open Research platforms that F1000 operate – and to declare my (conflict of interest) COI, I am Managing Director of F1000): you are absolutely correct that we require the data to be made openly available for all our articles, but because our peer review process takes place after publication, this means that all our referees can indeed look at the data as part of their evaluation process. We also specifically prompt our referees to do exactly that in the referee report form, with some examples of really valuable reanalyses and subsequent insights.

Another Workshop organised by the Research Data Alliance (RDA) Europe that I took part in brought various science disciplines together with the theme for the meeting of "*The opportunities and challenges of open data for research*". The Workshop looked at three specific issues:

> How do researchers share and work with other people's data within their discipline and across disciplines?
> What are the incentives and rewards to researchers for sharing their data?
> What infrastructure tools and services are of particular use to researchers in their disciplines?

These themes RDA stated *"are inspired by the emerging European Cloud Initiative, which is drawing together research infrastructures and e-infrastructures across Europe."* This Workshop allowed the sharing of current and best practice between disciplines including plasma fusion research and materials research as well as crystallography which I spoke on.

Also the developing world researcher may prefer to avoid openness of research data linked for example with their agriculture if that openness leads to a developing world data theft by the more developed world's researchers who may be better funded with better facilities. This is itself a complex landscape for realising fairness and responsible practice (Ferris and Rahman 2017).

In the neurosciences is the recent startling report of Miyakawa (2020), Editor of Molecular Brain. He

5 *https://asapbio.org/peer-review/agenda*

described the situation of manuscripts that were submitted to him since 2017. He stated:

> As an Editor-in-Chief of *Molecular Brain*, I have handled 180 manuscripts since early 2017 and have made 41 editorial decisions categorized as "Revise before review," requesting that the authors provide raw data. Surprisingly, among those 41 manuscripts, 21 were withdrawn without providing raw data, indicating that requiring raw data drove away more than half of the manuscripts. I rejected 19 out of the remaining 20 manuscripts because of insufficient raw data. Thus, more than 97% of the 41 manuscripts did not present the raw data supporting their results when requested by an editor, suggesting a possibility that the raw data did not exist from the beginning, at least in some portions of these cases.

4. Future developments to increase new and better science results and publication

Science in general is experiencing considerable expansion of data flows. This is true in crystallography as we have, and welcome, the ever-improving synchrotron radiation and X-ray laser, as well as neutron facilities, and enhancement of detectors that is still ongoing. This means that simple prioritising of which data to analyse would lead to delays. But the funding agencies are pressing all researchers with policies demanding quicker results and certainly within 3 years after the measurement. Artificial intelligence (AI) and machine learning (ML) must be brought increasingly to bear to assist the researcher to comply with such new policies. The alternative, again with funding agency pressure, is that the researchers' data must be made open after a period, such as 3 years since measurement (see for example the European Synchrotron Radiation Facility Data Policy[6]). This policy is in line with European Space Agency requirements with which my PhD student and I had to comply in our microgravity experiments. Indeed PhD students may need protection of their own data during their training but maybe for longer than 3 years if they are delayed for some reason in completing their studies.

It is a challenging world to the researcher. In section 2 I also described the situation where making the raw data open could lead to making a breakthrough.

5. Evidence of my progress with pre-publication peer review

In my article (Helliwell 2018) I documented my experiences with applying the chemical crystallography refereeing approach to my particular field of biological crystallography over a 1½ years period: "I have successfully refereed approximately twelve such article submissions in this time on this basis. These were from seven different journals and four different publishers." In one case, I received the following letter text from the editor of a specialist chemistry journal, which I have anonymised:

> Dear Reviewer,
> This is just a short letter of thanks for your review of the manuscript by Prof. xxx for our Journal of yyy. We recognize how much time careful reviewing requires. Your time and effort are appreciated.
> The quality and prestige of the journal can only be enhanced by the knowledge and advice of reviewers such as you.
> Many thanks!
> Sincerely,

In another case, the journal policy was to publish the peer review reports, also a good initiative. The pre-publication peer review reports on Langan *et al.* (2018) are therefore available to view at Nature Communications (2018). Not long after, Nature (2018) published its editorial entitled *"Referees should exercise their rights: Peer reviewers should not feel pressured to produce a report if key data are missing"*.

6. The role of international organisations in commitment to standards

A leading organisation for all matters to do with scientific data is the International Science Council (ISC)'s Committee on Data, 'CODATA'.[7] An important recent document in terms of open science in general and data in particular is *Open Data in a Big Data World* (2015) published jointly by the International Council for Science (ICSU), the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC). The IUCr endorsed this accord, which provided a detailed and reasoned account of responsibilities and opportunities for all stakeholders in the modern data-rich scientific enterprise. The endorsement was supported by a position paper which provided details of how the crystallographic community in general, and the IUCr in particular, respond to the challenges laid out in the Accord (Hackert *et al.* 2016) and offers the view:

> Science is best served when access barriers to data (and publications) are low. A major barrier to access is cost, and the phrase 'open access' is often used to characterize access to data

6 *https://www.esrf.eu/datapolicy*

7 The ISC is the organisation arising from the recent merger of the International Council for Science (ICSU) and the International Social Science Council (ISSC).

> and publications that involve no charge to the end-user. However, the maintenance of the highest levels of quality in collecting, evaluating, storing and curating data is a very expensive component of the scientific process, and care must be taken to understand how to obtain the maximum benefit from public funding of science.

And:

> Crystallography has a diverse ecosystem of disciplinary databases, data repositories, experimental facilities and publishers (Bruno et al 2017). Several of these are sustained through subscription-based access; but the other side of the coin is that they ingest, evaluate and publish data and information at no charge to the author/depositor, and without imposing any additional charge on the public purse. At the present time, this variety of approaches to sustainability and quality assurance serves this discipline well.

7. So, who pays and who will accept to do refereeing of 'article with data'?

So, I get asked these two questions that I have reiterated in my title of this section. I could have equally well entitled it feasibility of all that I advocate here. Let's do an options' analysis:

- A favourite of the Russian army general Mikhail I. Kutuzov, as described in detail in Tolstoy's War and Peace, is do nothing or rather in his case save lives by the strategy of retreat, which he did even after the victory at Borodino. In achieving final victory over Napoleon's army the Russian winter did the trick and Napoleon retreated. So, will *irreproducibility of science* go away if we do nothing? I think not and if trust is increasingly lost will taxpayers continue to fund us scientists, through Government funding agencies as their proxy? Do nothing is not a realistic option.

The other options are then a matter of scale and level of commitment:

- Journals could, rather simply actually, strongly encourage the review of the submitted article with data approach. Increasingly the handling scientific editor's name is included at the start of the publication so editor pride can be harnessed. In addition a gold star on a publication would indicate a full review of article with data had been done. This would also offer an alternative to the "gold star by impact factor" chasing by authors to get published in the highest impact factor journal, and which can cause considerable delays in publication. Deans of Universities could encourage this new vision instead of demanding high impact factor publications only from their staff, a trend at least in the UK. Internationally the DORA (Declaration on Research Assessment[8]) pursues the same goal of not slavishly following journal impact factors.

- Funding agencies could provide a funding pulse for say three years to journals who would actively seek to employ a data specialist so that the burden would not be on referees but on the in house journal staff member. In that time frame hopefully journals pursuing such a methodology of pre-publication review would then self-fund the cost of the extra staff person. The UK's JISC (Joint Information Systems Committee, a membership organisation, owned by the sectors they serve such as UK higher education[9]), offers funding pulses such as helping a publisher promote open access publications.

- Funding agencies could guide its researchers away from journals that do not undertake article with data pre-publication review. A list, not unlike the list of predatory journals, could be readily assembled.

- Funders could also recognise laboratories that do early data sharing in urgent situations like the covid-19 pandemic. A marvellous example of data sharing with an explanatory, detailed, blog post, before even a preprint, is from the FraserLab and collaborators (2020).

The above four points offer options in a plan of action harnessing an increased openness to referees for enhanced reproducibility in science. In sections 2 and 4 I described the situation of opening up the raw data to a whole community where the research was at a stop or proceeding too slowly or speed was of the essence. An example of the latter situation is immediately to hand with the covid-19 pandemic and where in crystallography a great deal of prompt action and data sharing worldwide has been on the way (Arago *et al.* 2020, Kramer 2020, Blakeley and Schorber 2020, FraserLab2020), including prompt reanalyses based on those shared data being offered (Wlodawer *et al.* 2020). In such research challenges working together as a whole community has so much more to offer than a fragmentation of individual parts.

8 https://sfdora.org/
9 https://www.jisc.ac.uk/about

Acknowledgements
As Editor in Chief of IUCr Journals 1996 to 2005 I saw in detail the exemplary checks undertaken by Acta Crystallographica Section C Coeditors, which included Dr Madeleine Helliwell, using the checkCIF report as well as the chance to scrutinize with their own calculations the underpinning processed structure factors and derived atomic coordinates of each submitted article. Thus I am very grateful to all colleagues involved with Acta Crystallographica Section C at the time notably Professor Syd Hall, University of Perth, Australia and also to Dr Madeleine Helliwell for many discussions.

As IUCr Representative to CODATA (2011 onwards) and to ICSTI (the International Council for Science and Technical Information) from 2005 to 2014 I met colleagues from many science disciplines and associated agencies which greatly enriched my experience of the ecosystem that is the amazing world of science. ■

References

Allison, D. B., Brown, A. W., George, B. J. & Kaiser, K. A. (2016). *Reproducibility: A tragedy of errors.* Nature, 530, 27–29.

Aragao, David, Brandao-Neto, Jose, Carbery, Anna, Crawshaw, Adam, Dias, Alexandre, Douangamath, Alice, ... Winter, Graeme. (2020). *Raw diffraction data for structure of SARS-CoV-2 main protease with Z44592329* (ID: mpro-x0434 / PDB: 5R83) [Data set]. Zenodo. *http://doi.org/10.5281/zenodo.3730610* [This is just one example of around 50 such raw data set depositions from this research team.]

Blakeley, M. P. and Schorber, H. (2020). CERN Courier *https://cerncourier.com/a/neutron-sources-join-the-fight-against-covid-19/*

Bruno, I., Gražulis, S., Helliwell, J.R., Kabekkodu, S.N., McMahon, B. and Westbrook, J. (2017). *Crystallography and Databases. Data Science Journal*, 16, 38. DOI: *http://doi.org/10.5334/dsj-2017-038*

Ferris L and Rahman Z. *Responsible Data in Agriculture* [2017; version 1; not peer reviewed]. *F1000Research* 6:1306 (document) (*https://doi.org/10.7490/f1000research.1114555.1*)

FraserLab and collaborators (2020). "*Identifying new ligands for the SARS-CoV-2 Macrodomain by Fragment Screening and Multi-temperature Crystallography*" https://fraserlab.com/macrodomain/ with processed and derived data available at *https://zenodo.org/record/3932380#.Xwg2Euco_tS* .

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* available at *https://www.iucr.org/iucr/open-data*

Helliwell, J. R. (2017). *Should reviewers be expected to review supporting datasets and code?* Blog post for the meeting entitled "*Transparency, Reward, and Innovation in Peer Review in the Life Sciences*" *https://asapbio.org/helliwell-datasets*

Helliwell, J. R. (2018). Data science skills for referees: I biological X-ray crystallography, *Crystallography Reviews*, 24, 263-272.

Helliwell, J. R., Achari, A., Bloomer, A.C., Bourne, P.E., Carr, P.D., Clegg, G.A., Cooper, R., Elder, M., Greenhough, T.J., Shaanan, B., Smith, J.M.A., Stuart, D.I., Stura, E.A., Todd, R., Wilson, K.S., Wonacott, A.J. and Machin, P.A. (1981). *Protein crystal oscillation film data processing:A comparative study* Acta Cryst. A37, C311–C312.

Joosten, R.P., Womack, T., Vriend, G. and Bricogne, G. (2009). *Re-refinement from deposited X-ray data can deliver improved models for most PDB entries*. Acta Cryst. D65, 176–185.

Joosten, R.P., Long, F., Murshudov, G.N. and Perrakis, A. (2014). *The PDB_REDO server for macromolecular structure model optimization.* IUCrJ 1(Pt 4):213-220.

Kramer, D. (2020). *World's physics instruments turn their focus to COVID-19* Physics Today *https://physicstoday.scitation.org/doi/10.1063/PT.3.4470*

Kroon-Batenburg, L. M. J. and Helliwell, J. R. (2014). *Experiences with making diffraction image data available: what metadata do we need to archive?* Acta Cryst. D70, 2502–2509.

Langan, P.S., Vandavasi, V.G., Weiss, K.L., Afonine, P. V., El Omari, K., Duman, R., Wagner, A. and Coates, L. (2018). *Anomalous X-ray diffraction studies of ion transport in K+ channels*. Nat. Commun 9, 4540. *https://doi.org/10.1038/s41467-018-06957-w*

Lawrence, R. (2018). *Comment to Helliwell, J. R. (2017). https://asapbio.org/helliwell-datasets#comments*

Miyakawa, T. (2020) Editorial: *No raw data, no science: Another possible source of the reproducibility crisis Molecular Brain.* *https://molecularbrain.biomedcentral.com/articles/10.1186/s13041-020-0552-2* DOI: 10.1186/s13041-020-0552-2.

McMahon, B., Helliwell, J. R. and Hester, J. (2018). *The element of trust: validating and valuing crystallographic data. https://www.iucr.org/__data/assets/pdf_file/0012/141222/6-mcmahon.pdf*

National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press. *https://doi.org/10.17226/25303*.

Nature Communications (2018) *Peer review report on Langan et al (2018) https://static-content.springer.com/esm/art%3A10.1038%2Fs41467-018-06957-w/MediaObjects/41467_2018_6957_MOESM1_ESM.pdf*

Nature Editorial (2018). *Referees should exercise their rights: Peer reviewers should not feel pressured to produce a report if key data are missin.g https://www.nature.com/articles/d41586-018-06006-y*

Science International Accord (2015). *http://www.science-international.org/#accord*

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

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). *The FAIR Guiding Principles for scientific data management and stewardship* Sci. Data, 3, 160018.

Wlodawer, A., Dauter, Z., Shabalin, I. G., Gilski, M., Brzezinski, D., Kowiel, M., Minor, W., Rupp, B., & Jaskolski, M. (2020). *Ligand-centered assessment of SARS-CoV-2 drug target models in the Protein Data Bank.* The FEBS journal, 10.1111/febs.15366. Advance online publication. *https://doi.org/10.1111/febs.15366*