Crystallographic raw data, education and refereeing

ABSTRACT

This article provides an overview of the preservation of raw diffraction data, then addresses the impact on future plans in the education and training of our community with respect to raw diffraction data and its potential reuse, and, thirdly presents the issue of referee access to the underpinning diffraction data and coordinates, as well as the Protein Data Bank Validation Report, in the review process of structural biology articles submitted for publication. Overall I pay tribute to the scientific achievements of Alex Wlodawer, who is also an ardent advocate of the importance of experimental data.

INTRODUCTION

I read with great interest Alex Wlodawer’s reminiscences of his pioneering work at the Stanford Synchrotron Radiation Laboratory in the 1970s that he published recently in Synchrotron Radiation News [1]. Likewise I have read his many decades perspective on the whole field of macromolecular crystallography that he published recently with his co-author Z Dauter in Protein and Peptide Letters [2]. It is a privilege also to read his educational guide on this theme in FEBS Journal [3].

We have clearly shared similar enthusiasms having both recognised, at the time when it was not obvious at all, the opportunities and potential importance of synchrotron radiation and neutron scattering in macromolecular crystallography [4,5]. A full description of my plans and ideas for the UK’s SRS I described in 1979 [6] and my summary of those early years at Daresbury, first on NINA and then the SRS (1976 to 1993) I have described at the SRS History website [7]. The spark for my interest was indeed that pioneering work [1] and which I drank thirstily from as it resonated with my developing research as a DPhil student in Oxford University 1974 to 1977. I provided a recent overview of the topic with Ed Mitchell at ESRF [8].

It was in those early years that as well as instrumentation and methods the data processing software had to be investigated in detail and I organised a classic style ‘round robin’ study as beamline station scientist at the SRS Daresbury Laboratory ([9] and figure 1). I presented this study and its results at the IUCr Ottawa Congress in 1981, my first lecture at an IUCr Congress; I have carefully preserved the overhead transparencies! Not least the results are still interesting and suggest a similar style of project would benefit the X-ray laser diffraction data processing methods. The education regarding crystallographic data extends to all of us, student to experienced investigators. I know myself the importance of my continual professional development (CPD).

The theme of this article is then my first efforts at analysing the changes needed due to the digital data storage revolution opportunities, which should allow us to preserve our raw diffraction data in digitised form and to make it readily accessible.

The need for archiving of the raw diffraction data, much of which has been championed by the IUCr Diffraction Data Deposition Working Group which I chair, has also been championed enthusiastically in ref. [2]. Access to the raw diffraction data allows the user/reader of crystallography results to see directly every calculation choice made by the original researchers, and take a different calculation route if they wish. The computer calculational revolution we have witnessed, Alex Wlodawer and I, is now extended to a digital storage revolution, making possible unimaginable new things in our research. I wish that we were at the start of our research careers again! I feel sure our research paths would again be similar, finding the important roads to follow.
RAW DIFFRACTION DATA PRESERVATION AND ACCESS TO IT

It has been envisaged for a long time (e.g. [10]) that the preservation of and access to raw diffraction data was important but technically and organisationally challenging; quoting from [10]: “Ideally, the full scientific record should provide access to the raw data……the IUCr is beginning to consider longer-term approaches to archiving the raw data”. Publication of raw diffraction data has one of its earliest exemplars in Lawrence Bragg’s 1913 publication on the crystal structures of the alkali halides with an extensive number of his own ‘Laue diffraction photographs’, measured in Cambridge, included in his article [11]. The IUCr global Diffraction Data Deposition Working Group (DDDWG) has for over four years now examined the issues and prospects for linking raw diffraction data sets to publications in the modern era. Considerable headway has been made. Our report for 2011 to 2014 can be found at http://forums.iucr.org/viewtopic.php?f=21&t=334. An important example from our DDDWG discussions, where the PDB is represented by John West-brook, is that the PDB now has a section of the PDB Deposition form available for depositors to log their raw diffraction data archive location (via a digital object identifier, DOI; see [12] and figure 2).

In terms of the practicalities of archiving raw diffraction data considerable headway has been made in the last year. Important strides have been made in the structural biology area (see below) and by the International Centre for Diffraction Data (ICDD) for retention of raw powder diffraction data. Long-time pioneers of raw diffraction data archiving are also at the National Crystallography Service at Southampton University, UK. At the neutron and synchrotron facilities major pioneering efforts include assigning digital object identifiers to all data sets by the facility, across all techniques, such as at ISIS, the Institut Laue Langevin and the ESRF.

Details of the benefits of access to raw diffraction data, such as the steady improvement of deposited protein structures, and the practicalities of use of data by others than those who measured them, are described in the articles in the October 2014 issue of Acta Crystallographica D [13-17]. Indeed, recently, a number of problems with structures of proteins, their ligands, nucleic acids, carbohydrates, bound metals, etc., have been identified and discussed in several publications (e.g. [18]). For storage of and access to such data sets a digital object identifier (DOI) for each raw data set is registered. In structural biology raw data archive examples are the NIH funded Big Data to Knowledge (BD2K, led by Wladek Minor at the University of Virginia), http://www.proteindiffraction.org/ (USA), http://zenodo.org (Europe), https://store.synchrotron.org.au/public_data/ (Australia; see ref. [15]), and Structural Biology Data Grid https://data.sbgrid.org/ (USA; see [19]). The BD2K initiative has ~2900 indexable and searchable diffraction experiments. In Poland (Mariusz Jaskolski personal communication) the Interdisciplinary Centre for Mathematical and Computational Modelling (ICM) of the University of Warsaw, has initiated in 2015 archiving of raw diffraction images with assigned DOI numbers in their RepOD (https://repod.pon.edu.pl/) open science repository; at present there are just a few deposits, but macromolecular crystallographers in Poland are encouraged to archive their raw data there. My own university of Manchester Library is an example of one university repository that now stores data sets from its researchers across all disciplines, and assigns DOIs to them.

A major development is that IUCr Journals (IUCrJ, J. Appl. Cryst., Acta Crystallographica D, F) have started linking their publications with primary data sets in such repositories. As an example, we have made the raw diffraction images for approximately thirty crystal structure studies on the binding of anti-cancer compounds (platins) to histidine in proteins (e.g. see doi:10.15127/1.215887, https://www.escholar.manchester.ac.uk/uk-ac-man-scw:215887) available at the University of Manchester Library Repository to accompany the corresponding publications at IUCr Journals, so that for these studies there is Gold Open Access to all the associated PDB files and all the raw diffraction image data.

Figure 1. An example of sharing data and improving methods in protein crystallography; abstract for the IUCr Ottawa Congress 1981; reproduced with the permission of the International Union of Crystallography. See [9].
In a wider context, the European Open Science Cloud (https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud) is providing tools and guidelines for ‘Open Science’ for a crowd sourcing of solutions to urgent societal problems including health challenges. A recent example of open science would be on the Zika virus. Hence funded laboratories in such a schema are encouraged to share their data as the work proceeds. Digital object identifiers for these data sets, thus representing a ‘data publication’, would be important to protect the individual researchers with their ‘IP’ as their research progresses. Crowd sourcing has already proven to be extremely successful in solving the crystal structure of an important drug target in antiretroviral therapy [20].

Overall, publication of raw data sets underpinning publications is now greatly facilitated by the data archives such as those above as well as those at the central facilities (synchrotron radiation, X-ray lasers and neutron sources) and linked to by journals, of which examples are also given above. Such a vision of the Open Science era will encourage digital object identifiers to be assigned for these data sets, which are an essential component of a data publication.

**RAW DATA AVAILABILITY ENCOURAGES AN ENHANCED EDUCATION AND TRAINING PROGRAMME**

I highlight just one area of education and training which I am focussing on currently, as I know Alex is very keen on education, e.g. as exemplified in his article in ref. [3]. At the European Crystallographic Meeting ‘ECM30’ upcoming in Basle I will speak in the Microsymposium on Education and Training focussing on the ‘raw data revolution’ and thereby its impacts in the future for our further education i.e. continual professional development. Firstly, researchers need to learn the new protocols associated with archived, open, raw diffraction data, as well as their processed diffraction data and derived coordinates being in the databases, with which they are quite familiar. Within the ‘archived, open, raw diffraction data’ approach the funding agencies are looking at an Open Science protocol for improvements to the speed of discovery, especially with respect to societal challenges, including right from the start of a funded research project the sharing of data. A second aspect of routine access to raw data that will need considering in breadth and depth is the impact on crystallographic teaching and our training courses for students entering crystallographic research such as our highly important European Crystallography School, an initiative begun by our Italian crystallography colleagues. These considerations represent the start of the new education initiatives needed for the Open Science era.

There is the pressing need also for the combined approach of mutual education in chemical and biological crystallography rather than the divergence which has dominated the last 30 years, e.g. as exemplified by the separate CSD and PDB databases, even though the PDB was launched by the CSD together with Brookhaven National Laboratory in 1971 ([21] and figure 3). As an example of our efforts at the University of Manchester in this direction of combining chemical and biological crystallography, we have provided a CPD Course covering both aspects, with training exercises; see http://www.iucr.org/news/newsletter/volume-19/number-1/manchester-school.
An example and very interesting aspect of this blending of chemical and biological crystallography was the enthusiastic joining in of Professor Durward Cruickshank, my long-time collaborator [22]. I have extended his approach of the ‘Cruickshank Diffraction Precision Index’ (DPI) with colleagues in Bangalore [23] including a webserver which provides the coordinate placement precision of each atom in an extended pdb coordinates file. DPI provides the overall precision due to the resolution, data completeness, R factor, average B factor as well as the individual atomic B factor. Articles reporting non-bonded distances, implying interactions, now can be easily, and properly, qualified by a standard uncertainty estimate on that distance. The community reaction to this so far seems to be to avoid giving more than one decimal place, even when they could, and even when integer precision is all that would be realistic (two atoms near to each other but with B factors of 80 Å², for example!).

Finally, as an educational benefit of raw diffraction images, we can reflect that a community that publishes detailed ontologies of all aspects of their data workflow (i.e. in our case the crystallographic information file ‘CIF’ dictionaries) makes quite transparent the concepts that are important in collecting, categorising and analysing the data. These provide very valuable material for scientists to deepen their understanding of all aspects of data analysis, and indeed to become more critical and careful in undertaking their analyses.

REFERENCES
7. Hellwell JR http://www.synchrotron.org.uk/ section on ‘Science’ and then ‘Biological Sciences’
Uwagi krystalografa o pierwotnych danych doświadczalnych, edukacji i recenzjach

John R. Helliwell

School of Chemistry, The University of Manchester, Manchester, UK

1/e-mail: John.helliwell@manchester.ac.uk

Słowa kluczowe: krystalografia makrocząsteczek z użyciem promieniowania synchrotronowego; krystalografia neutronowa makrocząsteczek; archiwizowanie surowych danych dyfrakcyjnych; inicjatywy edukacyjne w czasach otwartej nauki; recenzowanie artykułów wraz z danymi i współrzędnymi dyfrakcyjnymi; połączone szkolenie i edukacja w zakresie krystalografii biologicznej i chemicznej

STRESZCZENIE

W niniejszym artykule omówiono kwestię przechowywania surowych danych dyfrakcyjnych, plany dotyczące edukacji i szkolenia społeczności naukowej w odniesieniu do tych danych i ich potencjalnego ponownego użycia, a także problem dostępu recenzentów do źródłowych danych i współrzędnych dyfrakcyjnych oraz raportów walidacyjnych bazy Protein Data Bank podczas recenzowania prac z zakresu biologii strukturalnej. Przede wszystkim jednak artykuł ten jest wyrazem uznania dla naukowych osiągnięć Alexa Wlodawera, który sam niezmien