Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Archiving raw crystallographic data

Thomas C. Terwilliger‡

Bioscience Division, Los Alamos National Laboratory, Mail Stop M888, Los Alamos, NM 87507, USA

‡ Member, IUCr Diffraction Data Deposition Working Group.

Correspondence e-mail: terwilliger@lanl.gov

This article describes some of the activities of the IUCr Diffraction Data Deposition Working Group and introduces a collection of articles discussing the archiving of diffraction images.

One of the most important advances in the archiving of macromolecular structure information was the deposition of experimental data, typically in the form of structure-factor amplitudes derived from the measured intensities, into the Protein Data Bank (PDB; Bernstein *et al.*, 1977; Berman *et al.*, 2000). The availability of such structure-factor data has allowed validation of the interpretations provided by the depositors, and crucially has allowed new interpretations to be made as well. Recently there has been extensive discussion of the idea that not only the amplitudes of structure factors, but also the raw diffraction images used to derive them, should be archived and made generally available. These raw images constitute the truly primary experimental data and offer the opportunity of being later reprocessed for obtaining improved estimates of the intensities, analyzing data at higher resolution than used in the original work, checking the interpretation of the symmetries of the crystals, analyzing the diffuse scattering that reflects correlated motions or disorder of atoms in the crystals, assessing and correcting for radiation damage, analyzing diffraction from multiple lattices present in the crystals, and serving as benchmarks in developing improved methods of analysis.

The International Union of Crystallography commissioned the Diffraction Data Deposition Working Group (DDD WG) in 2011 with J. Helliwell as chair to examine the benefits and feasibility of archiving raw diffraction images in crystallography. The DDD WG initiated spirited discussions on this subject on the CCP4 mailing list, held workshops to gather input on the idea of archiving raw images, and encouraged wide analysis of the ideas and challenges. In this issue of *Acta Crystallographica Section D*, several researchers in the field of macromolecular structure determination have written a collection of articles discussing the archiving of diffraction images, what it would make possible, and what challenges it gives rise to.

In the first article in this collection, Kroon-Batenburg and Helliwell give an example in which making archived images available to anyone led to a re-analysis and reinterpretation of crystallographic data. The authors and their colleagues had previously examined several data sets collected on lysozyme with bound cisplatin or carboplatin and had used the anomalous signal of the chlorine present in cisplatin to distinguish them. The raw images were placed on publicly available servers at the University of Utrecht and at the TARDIS Raw

Diffraction Data Archive in Australia. K. Diederichs, not associated with the original work, took the opportunity to re-analyze the raw images, to re-evaluate the chlorine occupancies, and also to re-evaluate the presence of high-resolution data in the images (Tanley *et al*., 2013). The re-analysis showed the scientific benefit of archiving of images but also highlighted challenges in reprocessing image data and the need to use available methods to capture and archive the metadata associated with the raw image data.

Next, Meyer *et al.* describe how archiving images and making them publicly available has been accomplished at the TARDIS Raw Diffraction Data Archive. At the Australian Synchrotron images are automatically processed and at the same time they are archived along with the metadata needed to process them. Researchers can use a cloud computing environment to analyze their diffraction data and web servers to access data and results. Further, users can make their data publicly accessible and enhance its utility with additional metadata about the crystals and the experiment. Importantly data sets can be permanently identified and retrieved with universally recognized digital object identifiers (DOIs). These DOIs can be included in depositions in the PDB and in publications so that interpretations can be permanently linked with the data. The TARDIS example shows that it is not only possible but also practical to make diffraction images available for future research.

In the third article in this collection, Guss and McMahon extend the discussion of Meyer *et al.* to describe how archiving of raw crystallographic data could be accomplished in general. They describe the characteristics of an ideal archive, which include persistence, identifiability, discoverability, linkage to publications and deposits, and verification. They go on to describe how this might be achieved with centralized archives (for example the PDB) or with distributed archives, where there are obvious benefits from having centralized archives but significant costs in maintaining them. Guss and McMahon emphasise the importance of metadata, the information about the experiment that makes it possible to interpret that experiment. The article describes current efforts to carry out archiving of raw crystallographic data and presents a vision of how broad efforts to archive raw data could be promoted by the research community. It is important to note that efforts could begin immediately with registration of raw data sets with DOIs and that there is a need to bring the community together to develop standards for the metadata that should accompany raw crystallographic data.

In the last article in this collection, Bricogne and I discuss how the availability of crystallographic data makes a cycle of continuous improvement of interpretation of crystallographic data possible. As new or improved methods are developed, having the crystallographic data available allows re-interpretation and improvement of the structures in the PDB. In addition, the availability of crystallographic data accelerates the development of new methods and algorithms, making ever improved interpretations possible. Making raw diffraction images available will extend the scope of this iterative process of mutual improvements of methods and results by allowing the extraction of more information and the derivation of more accurate estimates of diffraction intensities as methods for data processing are improved while also stimulating those improvements themselves. We conclude that as methods for interpretation and representation of crystallographic data improve, the resulting new models should be made available in the PDB so that the wide audience of PDB users have access to the most accurate and complete models available.

We all hope that you will enjoy reading these articles and that they will stimulate extensive debate and action in the crystallographic community.

References

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Guss, J. M. & McMahon, B. (2014). *Acta Cryst.* D**70**, 2520–2532.

Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). *Acta Cryst.* D**70**, 2502–2509.

Meyer, G. R., Aragão, D., Mudie, N. J., Caradoc-Davies, T. T., McGowan, S., Bertling, P. J., Groenewegen, D., Quenette, S. M., Bond, C. S., Buckle, A. M. & Androulakis, S. (2014). *Acta Cryst.* D**70**, 2510–2519.

Tanley, S. W. M., Diederichs, K., Kroon-Batenburg, L. M. J., Schreurs, A. M. M. & Helliwell, J. R. (2013). *J. Synchrotron Rad.* **20**, 880–883.

Terwilliger, T. C. & Bricogne, G. (2014). *Acta Cryst.* D**70**, 2533–2543.