Activities of the IUCr Diffraction Data Deposition Working Group 2011–2013

Terms of reference

It is becoming increasingly important to deposit the raw data from scattering experiments; a lot of valuable information gets lost when only structure factors are deposited. A number of research centres, *e.g.* synchrotron and neutron facilities, are fully aware of the need and have established detector working groups addressing this issue.

The IUCr is the natural organization to lead the development of standards for the representation of data and associated metadata that can lead to the routine deposition of raw data. A Working Group on these matters has thereby been launched by the IUCr Executive Committee, to which the Working Group will report, to be Chaired by Professor John R. Helliwell. Its provisional title is 'Diffraction Data Deposition Working Group of the IUCr'.

Discussion forum

Discussion forums were established on the IUCr web site (//forums.iucr.org), including one for 'Public input on diffraction data deposition' to which all interested parties are invited to contribute (http://forums.iucr.org/viewforum.php?f=7). The DDD Forum has been very actively harnessed in collating comments at three different levels of input: the public at large, IUCr Officers including its Commissions, and the DDD WG itself. Documents from ICSU, CODATA and so on also feature.

Working group members: Steve Androulakis *Representative of TARDIS (Australian repositories for diffraction images)* ● Sol Gruner *Diffuse scattering specialist and Synchrotron Radiation Facility Director* ● John R. Helliwell, Chair. *IUCr Representative to CODATA and to ICSTI; Chair, IUCr Commission on Journals 1996-2005* ● Loes Kroon-Battenburg *Data processing software developer and user* ● Brian McMahon *Coordinating Secretary, COMCIFS* ● Tom Terwilliger *Representative of IUCr Commission on Biological Crystallography* ● John Westbrook *Representative of wwPDB (Worldwide Protein Data Bank)* ● Hans-Josef Weyer *Synchrotron Radiation and Neutron Facility user* ● **Consultants:** ● Alun Ashton *Diamond Light Source* ● Herbert Bernstein *Head, imgCIF Dictionary Maintenance Group and member of COMCIFS* ● Frances Bernstein *Observer on data deposition policies* ● Gerard Bricogne *Active software and methods developer* ● Bernhard Rupp *Macromolecular crystallographer*

Report on the Bergen Workshop

This is an edited version of the full report by John Helliwell, Tom Terwilliger and Brian McMahon that appears at http://forums.iucr.org/viewtopic.php?f=21t=102

A full-day Workshop was organised by the DDD WG as a Satellite of the ECM27 meeting in Bergen, Norway, on August 6th 2012. Its purpose was to review progress during the Working Group's first year of activity, and to help frame a policy to be drafted by the IUCr DDD WG on raw diffraction data deposition for final approval by the IUCr Executive Committee.

Presentations on the following topics are available at http://www.iucr.org/resources/data/dddwg/bergen-workshop

- The IUCr Diffraction Data Deposition Working Group Activities since IUCr Madrid *J. R. Helliwell and Brian McMahon*
- Motivations, challenges, horror stories and opportunities: Experiences of diffraction data management, archival and publication at the UK National Crystallography Service. *Simon J. Coles*
- Report on several important EU projects: CRISP, PaNdata, NMI3, Biostruct X, HDRI and CALIPSO *Heinz J. Weyer*
- Linking raw experimental data with scientific workflow and software repository: some early experience in the PanData-ODI project *Erica Yang and Brian Matthews*
- Ten years and change: the MX data archive at ALS 8.3.1 *James Holton*
- Continuous improvement of macromolecular crystal structures *Thomas C. Terwilliger*
- Towards policy for archiving raw data for macromolecular crystallography: Recent experience *Loes M. J. Kroon-Batenburg, Antoine M. M. Schreurs, Simon W. M. Tanley and John R. Helliwell*
- Some Economic Considerations for Managing a Centralized Archive of Raw Diffraction Data *John Westbrook*
- A vision involving raw data archiving via local archives as a supplement to the existing processed data archives (PDB, CSD, ICDD etc) *John R. Helliwell, Brian McMahon and Thomas C. Terwilliger*

The need to have clarity on DDD issues has two main aspects. First, crystallographers have obligations to securely and properly retain the raw data that they measure ('loss of data is viewed as research malpractice'). Second, the reader of a published article involving crystallography can and should have access to the raw data on which the article is based ('don't take my word for it; try the data yourself and see directly the research results').

The actions and recommendations arising from the Bergen Workshop can be summarised as follows:

- The Workshop noted that there is an enthusiasm and encouragement to archive more than derived or processed data in many areas of science besides our own.
- The crystallographic community prides itself in making its processed data accompany its publications; indeed, this has been obligatory in IUCr journals for over 10 years.
- We, the crystallographic community, basically have three practical options in the near future to extend these principles to our raw data;
- via a local Data Archive
- via synchrotron or neutron or X-ray laser (or other large-scale experimental facility) data storage
- or via the corresponding author setting up a personal link to datasets underpinning publications on their personal websites. [At the Workshop the Protein Data Bank (John Westbrook) offered that the PDB would help to coordinate DOI registration in cases where the raw data could be hosted on a reliable public site.]
- So we suggest that we encourage all three practical options and recommend to the IUCr Executive Committee that:

– *Authors should provide a permanent and prominent link from an article to the raw data sets underpinning a journal publication*, with a view to making this a formal requirement on authors at such time as the community has adopted raw data deposition as a routine procedure.

Post meeting note; The IUCr Executive Committee endorsed this proposal from the DDD WG but replaced the word 'should' by 'may'. This is indeed still a positive step forward as it endorses the commitment of resources, for example by IUCr Journals, in assisting authors with this. See, for example, the article of Tanley et al. (2013) cited below.

There is an urgent need to be clear about the metadata required for the types of experiment and their raw data. John Westbrook of the RCSB stressed the importance of this: 'if the metadata details required are not standardised then there will be datasets which are nothing more than a mess and which would not be effectively usable by someone retrieving [them]'.

Some recent publications attempt to describe in detail the technical metadata required. (1) A protein crystallography article entitled 'Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals' by Simon W. M. Tanley, Antoine M. M. Schreurs, John R. Helliwell and Loes M. J. Kroon-Batenburg. *J. Appl Cryst.* (2013), **46**, 108–119. (2) An article defining data formats in X-ray absorption spectroscopy entitled 'Towards data format standardization for X-ray absorption spectroscopy' by B. Ravel, J. R. Hester, V. A. Solé and M. Newville. *J. Synchrotron Rad.* (2012), **19**, 869–874. (3) Data and metadata definitions have been published also for SAXS and SANS: 'Publication guidelines for structural modelling of small-angle scattering data from biomolecules in solution' by D. A. Jacques, J. M. Guss, D. I. Svergun and J. Trewhella. *Acta Cryst.* (2012), D**68**, 620–626.

While IUCr Commissions need to specificy 'technical' metadata – *i.e.* those specific to their experimental raw data – there is also a need to review 'generic' metadata – *e.g.* who 'owns' a data set, details of research grants, embargo periods etc. A higher-level classification of the domain of study may be needed. *E.g.*, a synchrotron facility might need to define different data storage policies for, say, X-ray diffraction images versus X-ray tomography images. Such policies could be automatically implemented if data sets had characteristics identifying what sort of scientific study they represent. We feel that it would be beneficial to form a specialist group analysing these requirements. Members of this sub-group would be specialists able to represent different subject areas and experimental facilities. It would probably be a sub-group of the IUCr DDD WG.

One way to encourage satisfactory clarification of metadata technical definitions and standards is for the IUCr Executive Committee to require its Commissions to provide metadata recommendations as soon as possible.

An exemplar of good practice demonstrating access to raw data is at the ISIS UK neutron source. In the workshop Dr Erica Yang showed an example STFC DOI landing page for a particular data set and discussed the ISIS data management policy, from which we highlight a couple of points: • *[5.4] PIs and researchers who carry out analyses of raw data and metadata are encouraged to link the results . . . with the raw data / metadata using the facilities provided by the on-line catalogue. Furthermore, they are encouraged to make such results publicly accessible.* • *[3.3.3] Access to raw data and the associated metadata . . . from an experiment is restricted to the experimental team for a period of three years after the end of the experiment. Thereafter, it will become publicly accessible . . . [unless a PI can] make a special case to the Director of ISIS.*

In the Workshop, discussion of the concept of 'The Living Publication' led to the suggestion that journals need a new category name for articles stemming from a 'starter article and data set'. '*Ad extensum*' was offered as a suggested article type name. Subsequent investigation revealed that a mechanism already exists in the electronic publishing world to let the reader know that an article is related to other articles and data, and this can be accompanied by metadata explaining the relationships. So perhaps there is no need for an '*Ad extensum*' designation; but for derivative articles there could instead be an agreed set of metadata across publishers. Specifically, the **CrossMark** scheme of CrossRef, an organisation defining such standards across publishers, is an example of an attempt to get publishers to collaborate in this way.

Next step: special articles

A short series of articles has been commissioned to appear in *Acta Crystallographica Section D: Biological Crystallography* to bring some of the relevant issues to a wider community.

Tentative authors and titles for this series are as follows:

- Gerard Bricogne – Why deposition of diffraction data is important
- James Holton – What data should be deposited for macromolecular crystallography?
- John Westbrook – Practicalities of storage and deposition of image data
- John Helliwell and Loes Kroon-Batenburg – Experience with making image data available. What metadata do we need to archive?
- Tom Terwilliger and Gerard Bricogne – Continuous improvement of macromolecular structures
- Mitchell Guss and Brian McMahon – How to make deposition of images a reality

Publication is expected in late 2013 or early 2014.

Run up to IUCr Congress, Montreal

In the run up to the IUCr Congress 2014 we anticipate further progress by IUCr Commissions in clarifying their metadata needs to accompany raw data relevant to them. Secondly the proactive efforts of authors at 'grass roots' level and the IUCr Executive at 'top down' level should help contribute to making available raw data in general (and diffraction data images in particular). Initiatives of this type are likely to be increasingly appropriate in the 'open access' era, which extends beyond the written word to the data that form the firm platform on which science is based. Raw data availability will be a natural extension to our existing practice, over several decades, of making available in an organised way processed data (structure factors) and derived data (molecular coordinates).