Triennial report to the IUCr Executive Committee from the Diffraction Data Deposition Working Group (DDDWG)

IUCr DDDWG Executive Summary for Triennium 2011 to 2014

* The Working Group has been established with a global coverage, an excellent range of specialist consultants and good links to the IUCr Commissions. It has been very active as measured for example by organising the following events:- a launch meeting in the Madrid IUCr Congress, a meeting at the ACA Boston in 2012, an international Workshop at ECM27 Bergen, and a lunchtime Forum held at ECM28.

* We have provided a stimulus to the IUCr Commissions to define their technique's raw data and associated metadata definitions. Publications that have directly or indirectly arisen on these topics so far span:- X-ray Absorption Scattering 'XAS' (2012 in JSR), Small Angle Scattering 'SAS' (2012 in Acta Cryst D) and Macromolecular Crystallography 'MX' (2013 in J Appl Cryst). We also wish to draw attention to the report of a very recent meeting held at the Advanced Light Source 'ALS' Berkeley on diffuse scattering (in Structure) and which will, we expect, surely stimulate further interest in raw diffraction data archiving.

* We have ourselves directly or indirectly gained practical experience of raw diffraction data storage and or archiving at:- neutron facilities, SR facilities (including a survey of ten facilities worldwide), within a UK university repository and via a personal weblink at a Dutch university. We have also held discussions directly or by email with the organisers of commercial or semi-commercial data or research repositories (Dryad and Research Gate respectively).

* We have identified examples of leading good practice of raw data long term archiving eg at the ISIS neutron facility and at the Australian synchrotron and of long term data storage eg at Diamond Light Source.

* We have identified the developing global plans for raw data archiving by particle physicists (using 'the cloud') as presented at CODATA 2012 held in Taipei. At this and other meetings we have made personal contact with the directorate and staff of the World Data System based in Japan.

* We have obtained agreement from the Acta Cryst D Main Editors for several articles to be commissioned focused on these DDD topics and two articles have been accepted and four more are promised by their authors imminently.

* We have made two Recommendations of principle to the IUCr Executive Committee meeting held in Adelaide in December 2012, and which were previously endorsed by the ECM 27 Bergen

Workshop participants, namely to encourage an increase in raw data archiving and availability.

* We have drafted four further Recommendations, within this document, of practical steps forward, to help inform further discussion by the IUCr Executive Committee in the Montreal Congress.

Introduction

Period covered: 2011 to 2014

This Working Group was set up in 2011 and has as its **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 title is 'Diffraction Data Deposition Working Group of the IUCr'.

A membership for the Working Group was established:-

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; Director of the Synchrotron Radiation Source Daresbury Laboratory 2002* * Loes Kroon-Batenburg *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)* * Heinz-Josef Weyer *Swiss Light Source and SwissFEL, 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 activities

An excellent launch meeting was held at the IUCr Madrid Congress. As well as the WG Members (excepting two who sent apologies) and its consultants, the attendance included many IUCr Commission Chairs.

The minutes of this inaugural meeting were posted at what has become a lively **Discussion forum** (http://forums.iucr.org) **for 'Public input on diffraction data deposition'**. The DDD Forum has been harnessed in collating comments at three different levels of input: especially the public at large tier, IUCr Officers including its Commissions, and the DDDWG itself. Documents from ICSU, CODATA and so on also feature. The DDDWG email group has been set up and proved very valuable comprising approximately 50 names spanning various IUCr roles.

In the Madrid Congress the **Commissions were charged at the DDD inaugural meeting to define the metadata** that should accompany their raw data linked to one or more exemplar publications. **A number of publications have now appeared, as listed below:-**

(1) A macromolecular X-ray 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 each IUCr Commission needs to specify '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 could be a sub-group of the IUCr DDDWG.

Other categories of metadata exist ie for a particular research academic subject discipline. For biology for example our attention has been drawn to biological ontologies by Chris Morris, Chair of the Research Data Alliance, Structural Biology Working Group:-

Ontology for Biological Investigations http://obi-ontology.org/page/Main_Page
PROV-O http://www.w3.org/TR/prov-o/

An **exemplar of good practice** demonstrating access to raw data **is at the ISIS UK neutron source** (see https://data.isis.stfc.ac.uk/doi/INVESTIGATION/24079627/) 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.*

We would also wish to highlight the Australian synchrotron "Store.Synchrotron" data archiving, including its support of publications with raw data sets and their doi registrations and finally release of data sets for public analysis; for details see:- https://www.rdsi.edu.au/rdsi-story-storesynchrotron

At the ACA 2012 held in Boston, MA, USA July 29, 2012 (Present: Frances Bernstein, Herbert Bernstein, Patrick Mercier, Paul Langan, John Westbrook, Marian Szebenyi, James Holton. Chair: Tom Terwilliger) a wide variety of points were discussed; these provided an excellent and fertile lead up to the ECM27 Workshop in Bergen the following month (described in the next section). In Boston the Summary discussion stated: "It will be highly useful to have demonstration projects that identify both the extent of use and bottlenecks in storing raw images."

As a satellite meeting to ECM27 held in Bergen in August 2012 the DDDWG organised a Workshop whose 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 DDDWG 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/dddw ... n-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 related to user data handling: CRISP, PaNdata, NMI3, Biostruct X, HDRI and CALIPSO; strong endorsement of close cooperation between users and facility responsibles *Heinz-Josef 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 was stated to have 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 *i.e.* the chosen space group, the diffraction resolution limit etc; try the raw 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 data storage at a synchrotron or neutron or X-ray laser (or other large-scale experimental) facility
- 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 registration of DOIs (digital object identifiers: unique identifiers stored by a federation of registration agencies and able to be resolved to a specific location) in cases where the raw data could be hosted on a reliable public site.]
* So we suggested that we encourage all three practical options and **recommended 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.*

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 above.

It was again emphasised that 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 cogently 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]'.

In the ECM27 Workshop, discussion of the concept of 'The Living Publication' (an *ICSTI Insight* article written by JRH, BMcM and TT; see http://www.icsti.org/IMG/pdf/Living_publication_Final-2.pdf) led to the perceived need for a new category name for publications stemming from a 'starter publication and data set', and which may well be in different journals. Specifically, the **CrossMark** scheme of CrossRef, an organisation defining such standards across publishers, is an initiative among

publishers that tracks versions of a publication, and that might have some relevance to this requirement. Examples of the role of raw data in future, within a 'Living Publication' framework, could include for example better ways of estimating the diffraction resolution limit and also likely increased uses of diffuse X-ray scattering data, both of which are in macromolecular X-ray crystallography (for a recent meeting summary on planning an increased usage of diffuse scattering, held at ALS Berkeley, see:- http://www.cell.com/structure/fulltext/S0969-2126%2814%2900010-0).

At the **28th ECM in Warwick**, August 2013, an Open Forum meeting was held, following a series of presentations in the Symposium Crystallographic Information and Data Management. A detailed report of these activities is presented as an Appendix to this report.

In the last year a short series of articles has been commissioned by DDDWG Member T. Terwilliger to appear in *Acta Crystallographica Section D: Biological Crystallography* to bring some of the relevant issues to a wider community.

Accepted and/or submitted articles for this are:-

- Loes Kroon-Batenburg and John R Helliwell - Experience with making image data available. What metadata do we need to archive?
- Mitchell Guss and Brian McMahon - How to make deposition of images a reality

Further authors and titles for this series are expected to be 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
- Tom Terwilliger and Gerard Bricogne - Continuous improvement of macromolecular structures

Publication is expected in mid-2014.

Summary

In the run up to the IUCr Congress 2014 we firstly 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).

The practical challenges for raw data availability are very significant however and can be summarised as follows:-

(i) A preliminary analysis of the feasibility of storing diffraction data images on the IUCr journals platform concluded that this would be challenging because of network bandwidth limitations and the heavy investment in infrastructure that would be needed to scale up to host large data sets, even though actual storage costs per terabyte might be affordable on a per article basis (*i.e.* they could be added as a simple one-off fee to the publication costs).

(ii) The SR facilities, where a large fraction of macromolecular crystallography is measured (~90%), for example, have been surveyed and do not wish to be regarded as offering a 'data archive' service, although there are examples like Diamond Light Source that has so far retained all its measured data. The ESRF has published a splendid summary on their view of 'Big Data' at SR facilities in general and with the challenges involved today as exemplified by ESRF itself. See: http://mag.digitalpc.co.uk/fvx/iop/esrf/1312 . A leading exemplar of the SR facilities is the Australian synchrotron, referred to above, which operates its "Store.Synchrotron" with diffraction images data archiving, its support of publications with raw data sets via doi registrations and finally release of datasets for public analysis.

(iii) What might universities provide for their researchers? For example, discussions have taken place between JRH (with BMcM) and data archive and repository staff of the University of Manchester, UK, to explore what is required and/or possible to archive raw data through institutional repositories. This University launched a data archive in September 2013 as an extension of its staff 'eScholar' Green Open Access article repository. This initiative is driven by the University feeling the burden of responsibility in effect entrusted to them by the funders of research projects, being of the view that PIs are retaining their raw data (although precise definitions of raw data seem to vary). It is now becoming clear that (i) there is a clear policy of its research staff taking on the responsibility for raw data retention; (ii) the University makes available long term (*i.e.* effectively in perpetuity) disc storage on request for raw data; (iii) raw data sets can be lodged at the University institutional repository up to a certain file size limit, which seems to be around 300 Mbytes. Thus, a typical raw diffraction images dataset totalling

approximately 1 Gbyte has to be separated into approximately 3 x 300 Mbyte sub-folders. At the time of writing this report it is not yet clear how DOIs will be assigned, but the University data archive and repository staff continue to signal that this should be possible.

(iv) What might centralised repositories like Research Gate or arXiv offer? Research Gate already allows linking to data sets and informal discussions (TT) suggest that the file size limit could be expanded as a pilot project. arXiv also already allows linking to data sets but the limit set is a 'few Mbytes'.

(v) The use of a researcher's personal web link has been shown to be effective, at least over a short term (actual period undefined but presumably not 'in perpetuity'), in an example linked to the *J. Appl Cryst* paper (Tanley *et al*. 2013) cited above. This study compared data sets relating to a group of similar protein anti-cancer metal ligand complexes, from different home laboratory X-ray diffractometers, and using different software packages. Thus a reader can access the associated eleven 'raw datasets' (comprising ~35 Gbytes of X-ray diffraction data images) at Utrecht University in the Netherlands. Subsequent to this publication the TARDIS data archive in Australia set up a mirror facility for access to these datasets.

(vi) The imgCIF dictionaries continue to be developed in a way that will facilitate interoperability with NeXus/HDF5 workflows at Synchrotron Radiation facilities, and imgCIF/CBF is now supported as an image format by all the major vendors.

(vii) The next generation of pixel area detectors (PADs) such as the Eiger, whilst splendid in themselves, are perceived as exacerbating the raw data retention practical challenges. Mitigating factors however may be that there will be only a few centres that will have one; and/or there will be some degree of raw data pre-processing or data compression.

The X-ray lasers and the 'ultimate synchrotron emittance' upgrades and the new SR facilities of that type are now upon us and/or soon to arrive. This has the potential to put us right in the middle of the challenge identified by ICSU in its report http://www.icsu.org/publications/reports-and-reviews/strategic-coordinating-committee-on-information-and-data-report/SCCID_Report_April_2011.pdf :

"The explosion in the quantity of data and information available to science continues apace. Whilst the absolute size of this explosion varies across disciplines, the general trend is for rapid growth in all disciplines from the social sciences to seismology, from the humanities to high energy physics. By the end of 2011 it is estimated that 30,000 human genome sequences will have been completed, creating information about billions of bases and requiring petabytes of data storage. A study by the International Data Corporation (IDC) in 2010 estimated that by the year 2020 there will be 35 zetabytes (ZB) of digital data created per annum, which is 44 times the amount of digital data produced in the year 2009. The IDC estimate of the total digital storage in the world to be available in 2020 is 15 ZB, less than half the amount of digital data produced by then. When the Square Kilometre Array radio telescope in astronomy is fully functional in 2024 it will produce more digital data than is capable of being processed in all the world's computers put together (for a recent description see www.alexstjohn.com/WP/2014/02/26/square-kilometer-array-telescope-ska) .

Many data sets that exist only in printed form, including those located in LEDCs, still need to be converted into digital form so that they become more widely and easily accessible."

Issues

A. The IUCr, with its brief to oversee all of its techniques and disciplines cognate to crystallography, must have within its purview certainly its own version of 'Big data', and even perhaps 'Massive data' towards the level of the anticipated data-deluge headache of the Square Kilometre Array radio telescope in astronomy.

B. An issue for IUCr that will likely come to the fore in the next triennium, and which has already commenced, is that of rights of access to publicly funded, but unpublished, research data after an elapsed period *e.g.* 3 to 5 years. Our 'allies' in the practice of data linked to publication, the astronomy community, have a relatively easier policy problem here, inasmuch as the night sky is 'open access' to any observational technique. However, there can be a considerable investment of effort required to create a sample such as a well ordered membrane protein crystal, perhaps after decades of research, on which diffraction, spectroscopic or imaging measurements are finally made. These considerations should surely mean that those measured data remain in the sole ownership of the research team for as long as it wishes. Thus in such specific cases the data would not be made 'open access after an elapsed period' until and unless a publication is made. Then of course the usual benefits of open access to the reader would apply, also to the benefit of the authors. But how would such 'special cases' for not finally releasing data be policed? In the case of ISIS data, a user of the ISIS facility must make a written a case to the ISIS Director.

C. If in the Montreal IUCr Congress the IUCr Executive Committee casts doubt on the feasibility of routine deposition of the 'data deluge' cases that we have detailed, other strategies might need to be considered. These could include 'triage', where individual full raw data sets are retained for only limited periods (eg 3 to 5 years) or retention is made of only a subset of data frames large enough to allow confirmation of key steps en route to the usual 'processed structure factor data' and thereby the 'derived coordinates data'; the disadvantage of this would be that new data processing methods eg for extracting diffuse scattering would be automatically excluded. An extreme option, although outside our terms of reference, would be the establishment of an archive of measured (ie identical or very similar) samples to allow later structure redetermination; we would observe though that 'sample damage' and 'sample shelf life' issues would not apply to the original measured raw data, whose archiving we firmly believe is the best option even in data deluge situations.

Recommendations from the DDDWG for the upcoming Triennium

A. Various of the IUCr Commissions still need to define their metadata needs. One way to encourage satisfactory clarification of metadata technical definitions and standards is for the IUCr Executive Committee to require all its Commissions to provide metadata recommendations as soon as possible.

B. A particular category of raw data set publication that would be of immediate benefit to the crystallographic research community is that of data sets which defy interpretation or from which no molecular structure determination can be made, in full or in part. These 'difficult raw data sets' would be sufficiently small in number that the data network transfer overhead and disc store overhead at Chester would be manageable. It is proposed (LK-B), and endorsed by the DDDWG, that a new category of article could be introduced, in *J. Appl. Cryst.*, for example, where the authors would describe the nature of the dataset and in effect invite the community at large to work with these data. Preliminary discussions (JRH) with the *J. Appl. Cryst.* Main Editor Prof. Dr Anke Pyzallah in 2013 suggest that this proposal from the DDDWG would be well received. We note that *Nature* has announced its own data journal (see: http://www.nature.com/news/announcement-launch-of-an-online-data-journal-1.13906?WT.ec_id=NATURE-20131010) .

C. Specifications for a centralised crystallographic repository of metadata describing and locating experimental data sets should be scoped, along with functional requirements for a useful search interface, and the possibility of providing a pilot service investigated. Such an exercise should involve existing DOI registration agencies to ensure compatibility with evolving practice in other fields of science.

D. Subject to the establishment of a viable pilot metadata registry, we again recommend 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.

A meeting of the DDDWG of similar composition to that held at the IUCr Madrid Congress will be held in the IUCr Congress at Montreal.

John R Helliwell and Brian McMahon, Chair and Co-Chair;
Loes Kroon-Batenburg, Tom Terwilliger, Steve Androulakis, Sol Gruner, Heinz-Josef Weyer and John Westbrook

Members, DDDWG

Appendix: a detailed report on the most recent DDDWG event (held at ECM28)

A report on DDDWG activities and a number of other presentations relevant to the remit of the Working Group were presented at the COMCIFS Satellite Symposium to the 28th European Crystallography Meeting (University of Warwick, 25-29 August 2013). An Open Forum meeting was held on August 29 2013 to continue the process of public consultation that informs the activities of the DDDWG. 15 people were present, including representatives of databases, commercial instrument manufacturers, software developers and practising structural scientists. The IUCr Forum posting can be seen here:- http://forums.iucr.org/viewtopic.php?f=21&t=333 and the talks can be viewed here: http://www.youtube.com/playlist?list=PL6UK2yPUlpxq-VgI-CMMt4dnfEKV0hKSQ

John Helliwell gave a necessarily brief report on the progress of the Working Group so far. In changing the initial wording of the DDDWG proposal to the IUCr Executive Committee (EC) from "**Authors should provide raw data with a publication...**" to "**Authors *may* provide raw data with a publication ...**" the EC had signalled a yet more finely measured progress towards any eventual community-based requirements to attach raw data to publication, but endorsed the allocation of resources to allow this to be achieved, *e.g.* within its own journals. There was also a second proposal by the DDDWG for IUCr Commissions to define the metadata best suited in their field to characterising their experimental data, which was straightforwardly endorsed by the EC.

A set of potentially difficult discussion points was proposed by John Helliwell on behalf of the DDDWG, and any new input solicited from the audience. None were added. Thus the main topics for discussion were:

- Do people actually request or air a view wishing to have access to raw data, whether published or unpublished?
- How long should the raw data be available? As much as in perpetuity in the case of publication?
- After a time period without a publication should raw data derived from public funding be mandated for release? Some research fields operate such a mandate after 3 years (*e.g.* space research)?
- Local data archivists, rather than those at a specialised centralised repository, may be inexperienced at checking that depositors give all necessary metadata, thus rendering the raw data of limited future use by other researchers.

In the discussion, it was suggested that scientists would be receptive to the provision of a lightweight interface allowing easy annotation of data sets during the process of uploading to a storage service. There seemed to be a strong sense that a community-driven central storage facility would be attractive, but a central metadata registry coupled to distributed storage (possibly using commercial suppliers) was another potentially workable solution. Also we should point out that funding alone is not the sole barrier of accomplishing raw data archiving. The decoupling of metadata management

from bulk raw data storage might be helpful; on the other hand, the analysis from IUCr journals of bandwidth limitations involved in physical transmission of raw data sets suggested these were a more substantive problem associated with such an approach, rather than the cost of the data storage device after network transmission.

An invitation to argue against the retention of raw data - in principle in perpetuity - raised no dissenting voices. Some discussion touched on an optimum target for realistic retention periods; it was pointed out that in practice this could vary according to national policies (*e.g.* 10 years after the last access was UK's EPSRC's policy, we were informed). There was also general approval for the idea of collating raw data sets into a repository for structures that had proved impossible to solve, as a future resource to be exploited; this could be pursued in parallel, it was felt, with the main objective of securing more raw data set examples linked with publications.

Perhaps most surprisingly no one spoke against the posited statement regarding release of publicly funded data after a given time period, *i.e.* where no publication had resulted. This is a practice adopted by *e.g.* space science and astronomy, who use three years as the time period. Most interestingly this procedure is already adopted by the UK National Crystallography Service for the chemical samples that are submitted by its user-customers.

Report by Brian McMahon and John R Helliwell, 2 September 2013