HRVATSKA
UDRUGA
KRISTALOGRAFA
CROATIAN
ASSOCIATION of
CRYSTALLOGRAPHERS

Metadata for raw data from X-ray diffraction and other structural techniques

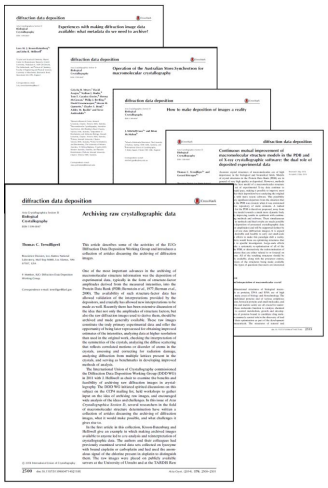
A Satellite Workshop to the 29th European Crystallographic Meeting

Programme and background materials

Organised by DDDWG (the IUCr Diffraction Data Deposition Working Group) and the Croatian Association of Crystallographers

This two-day Workshop is organised by the DDD Working Group (WG), appointed by the IUCr Executive Committee to define the need for and practicalities of routine deposition of primary experimental data in X-ray diffraction and related experiments. It will take the form of a two-day satellite of the 29th European Crystallographic Meeting with lectures from crystallographic practitioners, data management specialists and standards maintainers.

Objective: As part of the continuing activities of the IUCr Diffraction Data Deposition Working Group, this workshop will seek to define the necessary metadata that needs to be captured and deposited alongside experimental diffraction images in order that such raw data may be subsequently re-evaluated or re-used in more detailed scientific studies. The workshop will also explore the metadata requirements of other structural experimental techniques used by crystallographers.



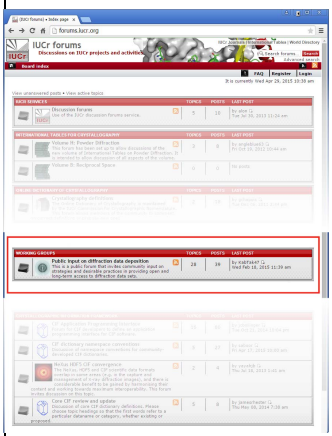
Included in this programme are the special articles commissioned by the DDDWG to analyse some of the issues involved in image deposition:

- Terwilliger, T. (2014). Archiving raw crystallographic data. *Acta Cryst.* D**70**, 2500–2501.
- Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). Experiences with making diffraction image data available: what metadata do we need to archive? *Acta Cryst.* D**70**, 2502–2509.
- Meyer, G. R., Aragão, D. *et al.* (2014). Operation of the Australian Store.Synchrotron for macromolecular crystallography. *Acta Cryst.* D**70**, 2510–2519.
- Guss, J. M. & McMahon, B. (2014). How to make deposition of images a reality. *Acta Cryst.* D**70**, 2520–2532.
- Terwilliger, T. & Bricogne, G.(2014). Continuous mutual improvement of macromolecular structure models in the PDB and of X-ray crystallographic software: the dual role of deposited experimental data. *Acta Cryst.* D**70**, 2533–2543.

Also included is the following experimental article:

Tanley, S. W. M., Schreurs, A. M. M., Helliwell, J. R. & Kroon-Batenburg, L. M. J. (2013). Experience with exchange and archiving of raw data: comparison of data from two diffractometers and four software packages on a series of lysozyme crystals. *J. Appl. Cryst.* **46**, 108–119.

There is a public forum for discussion of the issues covered in this workshop at http://forums.iucr.org



Welcome

This is the second full Workshop of the IUCr Diffraction Data Deposition Working Group (DDDWG). It follows a very successful meeting in Bergen in 2012 (programme and presentations are available at http://www.iucr.org/resources/data/dddwg/bergen-workshop). It is also a natural successor to the Crystallographic Information and Data Management Symposium at Warwick University in 2013, amplifying and building on many of the topics discussed there (http://www.iucr.org/resources/cif/comcifs/symposium-2013).

The Bergen Workshop surveyed the potential benefits of routine deposition of diffraction images, and explored some of the practical and cost implications of such a strategy. This led to a number of special articles published in *Acta Crystallographica Section D* that provided a detailed analysis of many of the issues involved. These articles are reproduced in this programme booklet.

A meeting of the Working Group at the IUCr Congress in Montreal in August 2014 concluded that there were promising movements towards widespread deposition of raw (otherwise known as 'primary') data, but that there were still a number of limiting factors. (1) Since there is no obvious single institution which will archive all crystallographic raw data, the initial strategy should be the encouragement of voluntary deposition in locations most convenient for authors (*e.g.* synchrotron and other instrument facilities, university and institutional repositories, domain repositories such as the Australian Synchrotron.Store). (2) Search and discovery functions across diverse locations would depend on common metadata identifying and describing data sets. The obvious candidate for an identifier is the Digital Object Identifier (DOI), because of the existing machinery to register and share DOI information. (3) Because molecular/atomic structural studies increasingly rely on a range of technologies and techniques, it would be desirable to harmonise metadata descriptions across as many such technologies as possible. Studying the 'arrangement of atoms' in its most general sense – as well as diffraction, spectroscopy and microscopy – has long been recognized as fitting within the remit of the IUCr.

While 'metadata' enters the discussion in the context of building distributed systems for search/discover, identification and retrieval of data sets, it rapidly becomes apparent that there is much more to metadata than that. 'Metadata' is variously defined, but the general sense is that it is the information that is needed to make sense of data, to allow its reuse, validation and critical analysis. Yet such 'information' is itself data – data that collectively open doors to further avenues of study, and even new scientific insight. Standard uncertainties on atomic positions modify the weights that should be given to structural models collected in databases, and so subtly affect our understanding of chemical bonding or biological function (*e.g.* in knowledge-based research using the Cambridge Structural Database or Protein Data Bank). The raw intensities ignored in models based solely on Bragg peaks (*i.e.* diffuse scattering) can now be reanalysed to provide insights into correlated disorder. Comparison of structural models derived from X-ray crystallography or from NMR can deepen understanding of protein structure and dynamics. Analysis of raw diffraction intensities from different experiments can yield examples of systematic bias (or, in extreme examples, dishonest practice).

Overall, the richer the metadata available to the scientist, the greater the potential for new discoveries. Crystallography is exceptional in the richness and granularity of metadata descriptors already available, mostly in diffraction-based research, and largely owing to the data dictionaries developed within the Crystallographic Information Framework (CIF), as so clearly shown in the Warwick Symposium. (That said, the achievements of other research communities in making available their data – such as the astronomers – should also be recognized. Our enthusiastic participation in organisations such as the International Council for Science (ICSU) and its Committee on Data (CODATA) is vital, both to represent crystallography, and to learn of best practice from other research communities.)

This two-day Workshop will survey the many uses already being made of crystallographic metadata, especially where associated with raw data capture, analysis and reuse. We will identify areas where better metadata descriptors are required, and we shall begin to look at the challenges of defining new metadata, especially in studies which do not have the clean, well-defined parameters of classical single-crystal or powder diffraction experiments. Some of the biggest challenges being faced are at the centralised synchrotron (and X-ray laser) and neutron facilities, where colossal quantities of diffraction, spectroscopy and especially microscopy raw data are being generated, and also in the databases which must organise and protect access to the fruits of all our researches in perpetuity.

We look forward to your active participation. We are grateful to our sponsors, who have made possible the web streaming and video recording of proceedings, so that we can reach a wider audience and provide a permanent record of the content of these two days. We shall enjoy the warm-hearted hospitality of our Croatian hosts in this beautiful location, and to whom we are indebted for their energetic and efficient logistical preparations. We welcome you to Rovinj, and to this latest IUCr DDDWG Workshop.

John Helliwell
Brian McMahon

Timetable

Saturday August 22

I. Introduction

10.00 am Introduction and welcome. John R. Helliwell and Brian McMahon

10.05 am Update on activities of the IUCr Diffraction Data Deposition Working Group (DDDWG)
John R. Helliwell
School of Chemistry, University of Manchester, M13 9PL, UK

II. Diffraction images – what can we get out?

10.20 am **Keynote:** The need for metadata in archiving raw diffraction image data
Loes M. J. Kroon-Batenburg
Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands

11.00 am Coffee break

11.20 am Crystallographic raw data: our plans and implementations within the NIH's Big Data to Knowledge resource
Wladek Minor
Department of Molecular Physiology and Biological Physics, University of Virginia, 1340 Jefferson Park Avenue, Jordan Hall, Room 4223, Charlottesville, VA 22908, USA

11.45 am Metadata needed for the full exploitation of diffuse scattering data from protein crystals
Michael E. Wall
Los Alamos National Laboratory, CCS-3 MS B256, Los Alamos, NM 87545, USA

12.10 pm X-ray Origins: Protection or Paranoia?
Natalie Johnson
School of Chemistry, Bedson Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

12.35 pm EIGER HDF5 data and NeXus format
Andreas Förster
Dectris Ltd, Neuenhoferstrasse 107, 5400 Baden, Switzerland

1.00 pm Lunch

III. Metadata for diffraction images and other experimental methods

2.00 pm Common diffraction image metadata specification in imgCIF, HDF5 and NeXus
Herbert J. Bernstein
School of Chemistry and Materials Science, Rochester Institute of Technology, Rochester, NY, USA

2.25 pm Towards a generalised approach for defining, organising and storing metadata from all experiments at the ESRF
Andrew Götz
European Synchrotron Radiation Facility, 71 Avenue des Martyrs, 38000 Grenoble, France

2:50 pm The PDB and experimental data
John Westbrook

RCSB PDB, Rutgers University, Piscataway, NJ, USA

3.15 pm Tea break

3:35 pm Realising the Living PDB and how raw diffraction data and its metadata can help
Tom Terwilliger
Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

4.00 pm Close of Day I

Sunday August 23

IV. Data in the Wider World – From Laboratory to Database

9.00 am Diffraction Data in Context: metadata approaches
Simon J. Coles
UK National Crystallography Service, Chemistry, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton SO17 1BJ, UK

9.25 am CCDC metadata initiatives
Suzanna Ward
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, UK

9.50 am Supporting Data Management Workflows at STFC
Brian Matthews
Scientific Computing, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell Oxford, Didcot OX11 0QX, UK

10.15 am Overview of metadata and raw data cataloguing at Diamond
Pierre Aller
Diamond Light Source, Division of Science, Didcot, Oxfordshire, UK

10.40 am CODATA and (meta)data characterisation in the wider world
Brian McMahon
IUCr, 5 Abbey Square, Chester CH1 2HU, UK

11.05 am Coffee break

V. What new metadata items are needed?

11.25 am What metadata is needed to make ESRF raw MX diffraction data intelligible for new users?
Gordon Leonard
Structural Biology Group, European Synchrotron Radiation Facility, CS40220, 38043 Grenoble Cedex 9, France

11.50 am What metadata is needed to make Institut Laue Langevin neutron diffraction raw data intelligible for new users?
Matthew Blakely
Institut Laue-Langevin, 71 Avenue des Martyrs, 38000 Grenoble, France

12.15 pm Metadata in high-pressure crystallography
Kamil Dziubek
LENS – European Laboratory for Non-Linear Spectroscopy, Sesto Fiorentino, Italy

12.40 pm Lunch

VI. Metadata schemas

1.40 pm Creating and manipulating universal metadata definitions
James Hester
Australian Nuclear Science and Technology Organisation, New Illawarra Road, Lucas Heights, NSW 2234, Australia

2.05 pm The Crystallographic Information Framework as a metadata library
Brian McMahon
IUCr, 5 Abbey Square, Chester CH1 2HU, UK

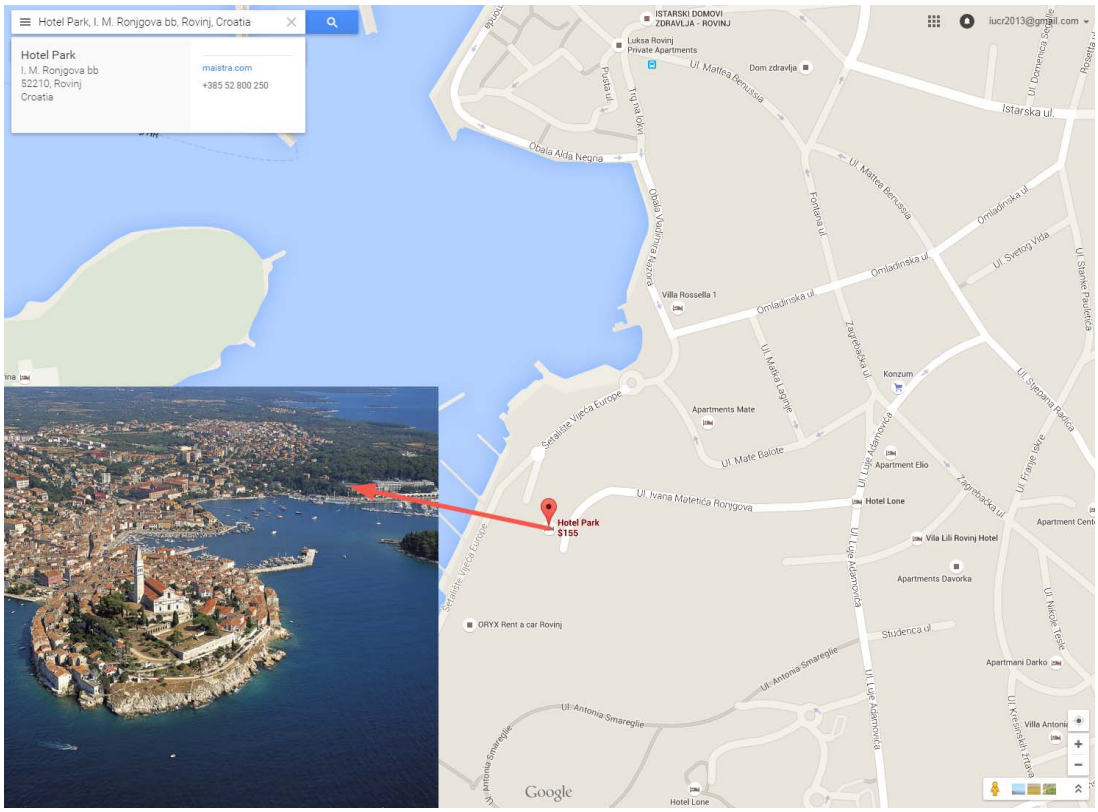
2.30 pm General discussion

2.40 pm Tea break

3.00 pm Practical session: building a metadata description

4.00 pm Close of Day II

6.00 pm ECM29 Opening Ceremony



Abstracts

Update on activities of the IUCr Diffraction Data Deposition Working Group (DDDWG)

John R. Helliwell

School of Chemistry, University of Manchester, M13 9PL, UK
Email: john.helliwell@manchester.ac.uk

John R. Helliwell[1] and Brian McMahon[2]
[1] School of Chemistry, University of Manchester, M13 9PL, UK
[2] IUCr, 5 Abbey Square, Chester CH1 2HU, UK

This workshop follows on from the 2012 Workshop on Diffraction Data Deposition at ECM27 in Bergen (http://www.iucr.org/resources/data/dddwg/bergen-workshop) and Working Group meetings at ECM28 (U. Warwick, UK; see http://forums.iucr.org/viewtopic.php?t=332) and the IUCr Congress (Montreal, Canada; http://forums.iucr.org/viewtopic.php?t=347). The Bergen Workshop identified the need for a thorough examination of current practice with metadata for raw diffraction data, and the possibility of using such a review to stimulate improved metadata characterization and handling in non-diffraction studies. This workshop will address both requirements.

In the wider scene 'Open Data' as a requirement of research publication is accelerating, whether it be derived, processed or raw data. Crystallography as a field compares well with other fields such as astronomy and particle physics in achieving 'open data' and each field finds raw data archiving challenging, especially the square kilometre array (SKA) in radio astronomy, since raw data is obviously the most voluminous. However, volume alone is not the greatest challenge. Stored raw data must be properly described so that its value and reliability can be assessed and understood, and individual data sets must be discoverable and reusable by other researchers, whether associated with formal publications or not. This is where metadata plays a key role.

We addressed technical options for achieving raw data archiving in Bergen and favoured flexibility in the physical location of the data sets, but with a key need for assigning DOIs to each raw data set. Interestingly, *Nature* magazine on 9 July 2015 highlighted 'the cloud' and commercial providers as being a preferred method for genomics data archiving. The change of attitude of the USA NIH, for example, where there were worries over the security of the commercial data store option, is significant.

John R. Helliwell *trained in physics and molecular biophysics and is now Emeritus Professor of Structural Chemistry at the University of Manchester. He is former Editor-in-Chief of the journals of the International Union of Crystallography and Past President of the European Crystallographic Association. His research involves crystallography methods developments applied to structural chemistry and biology. He is currently IUCr representative to CODATA (the ICSU Committee for Scientific Data) and chairs the IUCr Diffraction Data Deposition Working Group. He is also a member of the CODATA/VAMAS Working Group on the description of nanomaterials.*

The need for metadata in archiving raw diffraction image data

Loes Kroon-Batenburg

Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands
Email: l.m.j.kroon-batenburg@uu.nl

Loes M.J. Kroon-Batenburg[2] and John R. Helliwell[2]
[1]Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, The Netherlands
[2]School of Chemistry, Faculty of Engineering and Physical Sciences, University of Manchester, England.

Recently, the IUCr (International Union of Crystallography) initiated the formation of a Diffraction Data Deposition Working Group with the aim to develop standards for the representation of raw diffraction data associated with the publication of structural papers. Reports and minutes of DDDWG meetings can be found at forums.iucr.org. Archiving of raw data serves several goals: to improve the record of science, to verify the reproducibility and to allow detailed checks of scientific data, safeguarding against fraud or to allow reanalysis with future improved techniques. In a special series of papers on 'Archiving raw crystallographic data' [1], we reported on our experience of transferring and archiving raw diffraction data and on the problems encountered with acquiring and deciphering sufficient metadata [2].

To be able to process the raw data one needs information on the pixel geometry, information on pixel wise corrections applied, on beam polarization, wavelength and detector position amongst others, which are ideally contained in the image header. We will demonstrate that often one needs prior knowledge, evidently of how to read the (binary) detector format, but also on the set-up of goniometer geometries. This raises concerns with respect to long-term archiving of raw diffraction data. Care has to be taken that in the future unambiguous information is available *i.e.* one cannot simply 'deposit the raw data' without such metadata details.

We made available a local raw X-ray diffraction images data archive at the Utrecht University (raw-data.chem.uu.nl), subsequently mirrored at the Tardis Raw Diffraction Data Archive in Australia, and since March 2015 made available through digital object identifiers (doi) at the eScholar University of Manchester Library data archive. Since 2013 approximately 150 GB of data was retrieved from our archive and some of the data sets were reprocessed by other groups.

[1] Terwilliger, T. C. (2014). *Acta Cryst.* D**70**, 2500–2501.
[2] Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014) *Acta Cryst.* D**70**, 2502–2509.

Loes Kroon-Batenburg *heads a research group in the Department of Crystal and Structural Chemistry at the University of Utrecht. The research interests of her group focus on the development of methods for accurate integration of diffraction data. All methods are implemented in the software suite EVAL. Recently work started on data collection and the data processing of less orderedly packed crystals. Such crystals give rise to diffuse scattering. It is intended to develop measurement strategies and algorithms for data processing and interpretation of the diffuse scattering and for computing diffuse scattering from protein crystal structures. The diffuse scattering of crystals in between Bragg peaks is commonly ignored. However, these intensities are affected by so-called thermal diffuse scattering. Therefore, even the derived average structure is not fully accurate. The work is directed towards probing internal dynamics of macromolecules from the diffuse scattering.*

Crystallographic raw data: our plans and implementations within the NIH's *Big Data to Knowledge* resource

Wladek Minor

Department of Molecular Physiology and Biological Physics, University of Virginia, PO Box 800736, Charlottesville, VA 22908-073, USA
Email: wladek@iwonka.med.virginia.edu

The NIH pilot project 'Integrated resource for reproducibility in macromolecular crystallography' will create a web-based archive of diffraction images collected from macromolecular samples around the world. The resource will enhance and sustain the macromolecular diffraction data comprising the primary data sources for macromolecular atomic coordinates in the Protein Data Bank (PDB). The project will develop tools that will extract metadata from images alone, or from a combination of information obtained from a PDB deposit and diffraction images. All of the metadata needed for automatic determination and re-determination of macromolecular structures will be collected. Currently, the project has more than 1500 data sets and a preliminary system for extracting certain types of metadata. The data mining tools developed will allow for analysis of single experiments, as well as sets of experiments performed using various synchrotron and home based sources. Diffraction sets and metadata will be available from the project's website at http://www.proteindiffraction.org, or through a link on a PDB deposits page on the RCSB PDB website. This talk will present initial results of data mining performed on the archive.

Wladek Minor *is Professor of Molecular Physiology and Biological Physics at the University of Virginia. His laboratory studies macromolecular structure with the aim of in-depth understanding of structure–function relationships. X-ray diffraction analysis is the primary research tool, but other physical and biochemical methods of analysis are employed. The program emphasizes two broad themes; crystallographic studies on molecules of immediate interest, and methodology development. Most macromolecules under study relate to one or more of a few broad biological areas: cellular signal transduction and metalloproteins. The same systems have been chosen as subjects for methodology development. The methodology development includes the development of various crystallographic tools that create the HKL Package.*

Another research area is high-throughput crystallography and structural genomics. His lab is involved in a number of large, biomedically oriented projects that will revolutionize biomedical research in this decade. It is a member of the Midwest Center for Structural Genomics and the New York Structural Genomics Research Consortium (both centres of the NIH Protein Structure Initiative), and the Center for Structural Genomics of Infectious Disease (a project of the NIAID). It is also a part of the Enzyme Function Initiative (an NIH Glue Grant). It develops a methodology used in thousands of structural biology laboratories around the world. It collaborates with many synchrotron beamlines, in particular, with the Structural Biology Center at the Advanced Photon Source, and with many individual laboratories. The lab is well equipped to facilitate large scale protein purification, crystallization, biophysical characterization and detection of protein/protein or protein/small molecule interactions.

Metadata needed for the full exploitation of diffuse scattering data from protein crystals

Michael E. Wall

Los Alamos National Laboratory, CCS-3 MS B256, Los Alamos, NM 87545, USA
Email: mewall@lanl.gov

Technical release : LA-UR-15-23866

I will review efforts to model motions of crystalline proteins using diffuse X-ray scattering. This work requires analysis of raw diffraction images, which are mostly inaccessible in public databases. There is an abundance of potential metadata from these studies, including information about the analysis methods, measurements of

diffuse intensity, and results from the modeling. The time is now ripe for integrating diffuse scattering into traditional crystallography: modern beam lines and detectors are enabling higher quality data collection; computations which were previously inaccessible are now becoming feasible; and current protein crystal structure determination methods are approaching the limit of what is possible using the Bragg peaks alone. The deposition of raw images and associated metadata in public databases is a key step in enabling analysis of diffuse scattering for all protein crystallography studies.

Michael E. Wall *is a Scientist at Los Alamos National Laboratory in the Computer, Computational, and Statistical Sciences Division. He is generally interested in increasing the accuracy of models of molecular crystals obtained using X-ray diffraction. His interest in diffuse scattering began with his PhD dissertation study of three-dimensional diffuse features in X-ray diffraction from crystalline staphylococcal nuclease. Today his interests include diffuse data integration, quantum mechanical modeling of molecular crystals, and using diffuse scattering to develop models of crystalline protein conformational ensembles.*

X-ray Origins: Protection or Paranoia?

Natalie Johnson

School of Chemistry, Bedson Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK
Email: N.Johnson5@newcastle.ac.uk

Natalie Johnson and Michael R. Probert, School of Chemistry, Bedson Building, Newcastle University, Newcastle upon Tyne, NE1 7RU, UK

Deliberate fabrication of crystallographic data has previously led to falsified structures being published and then later retracted from respected scientific journals [1-3]. Identified perpetrators, in these cases, had made very simple modifications to structural files, such as manually changing unit cell sizes and atom types, to produce adjusted data. Fortunately they were found to be unable to produce raw experimental data to support their claims. Kroon-Batenburg and Helliwell [4] proposed that the requirement for the deposition of raw crystallographic data may be a potential method of preventing the submission of counterfeit structures. However we can show that the recreation of raw diffraction images is no longer difficult, opening the doors for those less scrupulous to take advantage, if this is not already occurring!

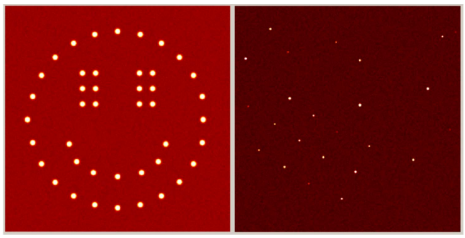


Figure 1. Two diffraction images - which is real?

Detector frame formats from many manufacturers are well documented and this information can be reverse-engineered to encode synthetic diffraction data. This process was brought to light as a product of research into optimising data collection parameters for charge density studies. The chosen method required us to produce an algorithm which takes data from integrated .raw files as a starting point to create replicas of experimental images. A simple misuse of this code could take structure factors calculated for an entirely fabricated compound and produce diffraction images that, when processed, return the artificial structure. The frames are not visually distinguishable from authentic, experimentally determined, ones and can be fully integrated using standard protocols. The authors find this situation potentially alarming and requiring immediate attention.

A structure refined from data processed from these artificial diffraction images could pass all IUCr checkCIF [5] protocols without raising alerts. We will present such a structure, full details of the algorithms employed and propose methodologies that may safeguard against this approach going undetected.

[1] T. Liu *et al.* (2010). *Acta Cryst.* E**66**, e13–e14.
[2] H. Zhong *et al.* (2010). *Acta Cryst.* E**66**, e11–e12.
[3] International Union of Crystallography (2010). *Acta Cryst.* D**66**, 222.
[4] L. M. Kroon-Batenburg & J. R. Helliwell (2014). *Acta Cryst.* D**70**, 2502–2509.
[5] A. L. Spek (2009). *Acta Cryst.* D**65**, 148–155.

Keywords: Data, Simulation, Software

Natalie Johnson *is a PhD Student in the School of Chemistry at Newcastle University. She works with Mike Probert, who developed a unique diffraction facility while at Durham for investigating crystalline materials under extreme conditions. This bespoke facility is now in operation at Newcastle.*

EIGER HDF5 data and NeXus format

Andreas Förster

Dectris Ltd, Neuenhoferstrasse 107, 5400 Baden, Switzerland
Email: andreas.foerster@dectris.com

Andreas Förster and Marcus Müller, Dectris Ltd, Neuenhoferstrasse 107, 5400 Baden, Switzerland

HDF5 is a container format designed for big data applications. In it, vast amounts of heterogeneous data can be stored in a small number of files that are easy to manage. Detectors of the EIGER series write datasets thousands of images big to HDF5 files and record most of the metadata that are required for data processing. The metadata are saved in a master file that is separate from the data but links to it. In this talk, I will present the HDF5 format and some of the metadata as written by EIGER detectors. I will also discuss metadata that are essential for processing but unknown to the detector and highlight blank fields that the EIGER HDF5 template provides for completion by beamline routines. A related talk by Herbert J. Bernstein [1] will explore ways of recording the geometry of the experimental setup. Software development, data processing, and effective archiving will all benefit from strict adherence to standards set by the NeXus committee.

[1] H. J. Bernstein 'Common diffraction image metadata specification in imgCIF, HDF5 and NeXus' in *Workshop on Metadata for raw data from X-ray diffraction and other structural techniques*, 22–23 Aug 2015, Rovinj, Croatia.

Andreas Förster *has been working as an Application Scientist MX at Dectris Ltd since March 2015. Previously he has worked at the Institut de Biologie Structurale in Grenoble, and at Imperial College London, where he was X-ray Facility Manager from 2013 to 2015. His 2005 PhD thesis was on the Mechanism of Proteasome Stimulation by 11S Activators.*

Common diffraction image metadata specification in imgCIF, HDF5 and NeXus

Herbert J. Bernstein

School of Chemistry and Materials Science, Rochester Institute of Technology, Rochester, NY, USA
Email: yayahjb@gmail.com

The introduction of a new generation of fast pixel-array detectors, such as the Dectris Eiger and the Cornell–SLAC Pixel Array Detector (CSPAD), has required us to revisit and extend approaches we have used in the past to represent the data (the diffraction images) and the metadata (the information needed to reconstruct the experimental environment within which the data were collected) [1] [2]. For example, the axis descriptions from the imgCIF (image-supporting-CIF) dictionary have proven effective in reliably preserving the information about the frame-by-frame relative positions of beams, crystals and detectors and have been mapped into the context of HDF5 and NeXus to support the new Eiger format. See Andreas Förster's talk [3] for a discussion of the Dectris Eiger-specific HDF5/NeXus format. We are introducing a new, extended templating scheme to allow each beam line to specify the unique characteristics that will allow the metadata for a beam-line to be specified as either an HDF5/NeXus file or as an equivalent CBF/imgCIF file from which a site-file to be merged with run-specific data and metadata will be generated. A central repository of site-templates will be offered for convenience. This approach will help both in ensuring ease of processing of original data and in facilitating reliable handling of archived data.

[1] H. J. Bernstein, J. M. Sloan, G. Winter, T. S. Richter, NIAC, COMCIFS, 'Coping with BIG DATA image formats: integration of CBF, NeXus and HDF5', *Computational Crystallography Newsletter*. 2014, **5**, 12–18
[2] A. S. Brewster, J. Hattne, J. M. Parkhurst, D. G. Waterman, H. J. Bernstein, G. Winter, N. K. Sauter, 'XFEL Detectors and ImageCIF', *Computational Crystallography Newsletter*, 2014, **5**, 19–25.
[3] A. Förster, M. Müller, 'EIGER HDF5 data and NeXus format', in *Workshop on Metadata for raw data from X-ray diffraction and other structural techniques*, 22–23 Aug 2015, Rovinj, Croatia

Work supported in part by Dectris and by NIGMS.

Herbert Bernstein *is a member of COMCIFS, Chair of the imgCIF dictionary working group, and lead developer of CIFtbx, a Fortran library for handling CIF data. He is also a member of the NeXus International Advisory Committee (NIAC).*

Towards a generalised approach for defining, organising and storing metadata from all experiments at the ESRF

Andrew Götz

European Synchrotron Radiation Facility, 71 Avenue des Martyrs, 38000 Grenoble, France
Email: andy.gotz@esrf.fr

After more than 20 years of operation the situation concerning metadata at the ESRF is still very disparate between beamlines. The way the metadata required to analyse the raw data are defined and collected depends largely on the beamline concerned. Approaches vary from fully automated solutions implemented on the MX beamlines to

a combination of automated and manual collection of metadata. This talk will not present the solution for MX (see talk by Gordon Leonard for more info) but will present a new approach for automating the collection and storing of well defined metadata for all experiments. The solution is based on a generic tool built at the ESRF which uses HDF5 for file format, Nexus for definitions (where possible) and ICAT for the metadata catalogue. The talk will present concrete examples of its use for nano tomography and fluorescence and radiation therapy. Ongoing work on how this will be extended to small angle scattering, coherent diffraction and eventually all other techniques will be presented. The talk will conclude with a discussion on the role of metadata in data policy and management.

Andy Götz *is Head of the Software Group at the European Synchrotron Research Facility.*

The PDB and experimental data

John Westbrook

RCSB PDB, Rutgers University, Piscataway, NJ, USA
Email: jwest@rcsb.rutgers.edu

John Westbrook *is a Project Team Leader of the RCSB Protein Data Bank, and is based at Rutgers University. He has played key roles in creating and maintaining the PDB database schema, in developing many of the software tools that underpin PDB operation, and in developing formal ontologies with other structural biology communities. He is a member of COMCIFS.*

Realising the Living PDB and how raw diffraction data and its metadata can help

Tom Terwilliger

Los Alamos National Laboratory, Mailstop M888, Los Alamos, NM 87545, USA
Email: terwilliger@lanl.gov

Thomas C. Terwilliger[1] and Gerard Bricogne[2]
[1]Los Alamos National Laboratory, Mailstop M888, Los Alamos, NM 87545, USA. Email: terwilliger@lanl.gov
[2]Global Phasing Ltd, Sheraton House, Castle Park, Cambridge, CB3 0AX, UK. Email: gb10@globalphasing.com

The Protein Data Bank (PDB) is the definitive repository of macromolecular structural information. The availability of structure factors for most entries in the PDB has made it possible to continuously improve the models in the PDB by reinterpreting the primary data for existing structures as new methods of analysis, new biological information, and new ways of describing structures become available. This continuous improvement will be even more powerful once the diffraction images associated with each entry become accessible.

The key factor is that depositing raw images will stimulate the improvement of integration and processing software in the same way as the deposition of merged X-ray data hugely stimulated progress in refinement software. Revisiting deposited images with that improved software will deliver more accurate data (especially, free from the currently inadequate treatment of contamination by multiple lattices) against which to re-refine the deposited structures themselves.

With the initial interpretation of a structure, the original structure factors and the raw images, it will become possible both to carry out extensive validation of structures and to apply new algorithms for structure determination and analysis as they become available, leading to structures of ever-increasing accuracy and completeness.

Keywords: Structure quality; validation; PDB; automation; structure determination; raw data deposition

Tom Terwilliger *obtained his doctorate at the University of California, Los Angeles. He was a Helen Hay Whitney Postdoctoral Fellow and a Presidential Young Investigator before joining Los Alamos National Laboratory in 1991. He developed the first completely automated procedure for finding the shapes of proteins by analyzing the diffraction of X-rays from crystals made of proteins. His SOLVE software converts the lengthy process used by macromolecular crystallographers to solve crystal structures into an optimization problem, and performs all the steps needed to go from diffraction spots to an electron density picture of a protein molecule.*

He was one of the founders of the field of structural genomics, in which the three-dimensional shapes of large numbers of proteins are determined in order to provide a foundation for understanding biology. He also founded the TB Structural Genomics Consortium to determine shapes of proteins from tuberculosis bacteria and provide a basis for drug discovery for treatment of the disease.

He is Chair of the IUCr Commission on Biological Macromolecules, Vice-President of the American Crystallographic Association, a Fellow of the American Association for the Advancement of Science and a LANL Fellow.

Diffraction Data in Context: metadata approaches

Simon J. Coles

UK National Crystallography Service, Chemistry, Faculty of Natural and Environmental Sciences, University of Southampton, Southampton, SO17 1BJ, UK
Email: s.j.coles@soton.ac.uk

Diffraction experiments and the results arising from them must often sit in a given scientific context – *e.g.* in chemical crystallography, they are often performed as part of a study concerned with synthesising and characterising new compounds. The context for an experiment, *i.e.* why it has been performed, is often lost – particularly in the case where data is published on its own.

I will present approaches not only to ascribing metadata to the results of crystallographic experiments, but also to the general chemistry leading up to them. The first stages of work to build a model to support this have been published – http://www.jcheminf.com/content/5/1/52. I will go on to discuss recent work in two projects: (1) a collaboration with five big pharma companies, instrument manufacturers, electronic lab notebook vendors and the Royal Society of Chemistry to derive metadata for capturing the 'process' of performing experiments; and (2) a project (https://blog.soton.ac.uk/cream/) aimed at using metadata actively in the process of performing research, as opposed to purely for archival purposes. I will conclude with insights as to how the approaches taken in assigning metadata in these projects are important to consider when archiving and disseminating raw crystallographic data.

Simon Coles *is Director of the UK National Crystallography Service and an Associate Professor at the Department of Chemistry at Southampton University.*

CCDC metadata initiatives

Suzanna Ward

Suzanna Ward Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK
Email: ward@ccdc.cam.ac.uk

Suzanna Ward, Ian J. Bruno, Colin R. Groom and Matthew Lightfoot
Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK

For half a century the Cambridge Crystallographic Data Centre (CCDC) has produced the Cambridge Structural Database (CSD) to allow scientists worldwide to share, search and reuse small molecule crystal structure data. An entry in the CSD is often seen as 'just' a set of coordinates, but the associated metadata (data that describes and gives information about other data), is essential to contextualise an entry. Data that describes the substance studied, the experiment performed and the dataset as a whole are all vital.

This presentation, timed to coincide with the 50th anniversary of the CSD, will look at how metadata is used from deposition to dissemination of the CSD. We will look at how recent developments surrounding metadata have been targeted to improve the discoverability, validation and reuse of crystal structure data before looking to see what the future may hold.

Suzanna Ward *has an MChem degree from the University of Southampton. During her Masters degree Suzanna got her first taste of crystallography through a project with Professor Mike Hursthouse and a placement at the pharmaceutical company Rhône-Poulenc. She then joined the CCDC in 2006 as a Scientific Editor, validating crystal structures into the CSD. Since then she has been involved with work to ensure data is released faster through WebCSD, changing the way data is curated into the database and the development of a new internal system, CSD-Xpedite, used in the creation of the database. These changes have transformed the way the team curate data into the CSD and have ensured the CCDC can keep up with increasing output of the crystallographic community. In 2013 Suzanna took on the role of Cambridge Structural Database Group manager and is now responsible for team that creates the CSD and manages all the transactions that go on behind the scenes with depositors, authors, publishers, referees and requests for data*

Supporting Data Management Workflows at STFC

Brian Matthews

Scientific Computing, Rutherford Appleton Laboratory, Science and Technology Facilities Council, Harwell Oxford, Didcot OX11 0QX, UK
Email: brian.matthews@stfc.ac.uk

STFC has developed a systematic approach for managing and archiving data generated from its large-scale analytic facilities, which is used with variations by the ISIS Neutron Source, the Diamond Light Source and the Central Laser Facility. This is centred around the ICAT experiment metadata catalogue. The ICAT acts as a core middleware component recording and guiding the storage of raw data and subsequent access and reuse of the

data; it has evolved into a suite of tools which can be used to build data management infrastructure. In this talk, I shall describe the current status of the ICAT.

The data rates and volumes generated from facilities are ever increasing and experimental science is becoming more complex. This is presenting challenges to the user community in accessing, handling and processing data. I shall describe some approaches to these problems and consider how we are exploring further support for data analysis and publication workflows within a large-scale facility. Finally, I shall consider how we might develop metadata to capture and share this information across communities.

Brian Matthews *has nearly 30 years of experience of work in computing science, mostly within STFC and its predecessor organisations, undertaking research and development in: formal methods of software engineering, data and metadata modelling, web-based systems, semantic web, distributed systems and trust. In particular, he has contributed to the development of metadata models for representing scientific data; development and deployment of data management infrastructure tools for facilities science data (the ICAT system). He led the work on Provenance in the EC sponsored PanData project, and is co-chair of the Research Data Alliance Interest Group on Data Needs of the Photon and Neutron science community.*

Brian currently leads STFC's Scientific Computing Department's programme of work in support of the Large-Scale Analytic facilities operated by STFC, in particular ISIS, Diamond and the Central Laser Facility. This includes innovative data management solutions to support the collection, storage, access and sharing of data, and access to high-performance computing clusters. The programme is also developing the direct support for experimental simulation and data analysis to enhance the range of capability offered to facilities users in an increasingly data intense scientific environment.

Overview of metadata and raw data cataloguing at Diamond

Pierre Aller

Diamond Light Source, Division of Science, Didcot, Oxfordshire, UK
Email: pierre.aller@diamond.ac.uk

Pierre Aller and Alun Ashton, Diamond Light Source, Division of Science, Didcot, Oxfordshire, UK

Diamond Light Source as a relatively new facility has been able to capture and catalogue all its raw data (now over 3.6 Petabytes). Additionally as much metadata and processed data as possible has always been collated with the raw data and captured into both databases (ISPyB) for query and quick access, and into the raw data files (imgCIF/CBF and NeXus). Progress and status on these developments will be presented.

Pierre Aller *is a Senior Support Scientist on MX beamlines. After a PhD on Molecular Dynamic Simulation, Pierre moved to the University of Vermont (Burlington, USA) for a post-doc. He studied protein crystallography for 6 years on DNA polymerase, before joining Diamond in August 2011.*

CODATA and data (meta)characterisation in the wider world

Brian McMahon

IUCr, 5 Abbey Square, Chester CH1 2HU, UK
Email: bm@iucr.org

This Workshop concentrates on scientific metadata and their importance in maximising the utility, trustworthiness and reuse of scientific data, especially to open doors to further avenues of study, and even new scientific insight. In a more general context, 'metadata' is a vehicle for categorising, classifying and collecting data sets. This presentation will review some of the organisations that take an interest in generic metadata and interoperability between metadata specifications from different disciplines or communities. The CODATA/VAMAS Working Group on Description of Nanomaterials provides a good example of collating different specialist metadata elements in a broad interdisciplinary framework. There will be a brief discussion of the granularity mismatch between generic and specialised metadata systems.

What metadata is needed to make ESRF raw MX diffraction data intelligible for new users?

Gordon Leonard

Structural Biology Group, European Synchrotron Radiation Facility, CS40220, 38043 Grenoble Cedex 9, France
Email: leonard@esrf.fr

The volume of diffraction data that can be collected during an experimental session on modern synchrotron-based Macromolecular Crystallography (MX) beamlines equipped with fast readout photon-counting pixel detectors and the rate at which it can be collected means it is currently difficult (or impossible) for users to manually process, during the experiment, all data sets collected. To help remedy this situation and to provide the at-beamline feedback that is sometimes necessary for a successful experiment 'autoprocessing' software [1,2] is often deployed with the results of automatic integration, scaling merging and reduction for individual data sets displayed in Laboratory Information Management Systems (LIMS) such as ISPyB [3] from where they can also be downloaded.

While the 'autoprocessing' approach works (*i.e.* provides data of sufficient quality for structure solution and refinement) in the vast majority of cases there is an increasing need for the post-experiment processing of raw diffraction images. In such cases the correct metadata for each dataset are essential to ensure the best results. For MX diffraction data collected at the ESRF this is stored in ISPyB, in the headers of the raw data images themselves and in automatically generated input files for the two main packages – *XDS* and *MOSFLM* – routinely used in the processing of ESRF-collected MX diffraction data. During my talk I will review the metadata currently logged during MX experiments at ESRF and look forward to what further metadata might be required when, either for validation purposes or the testing of new data processing and analysis protocols, raw data images are routinely made available to the wider scientific community.

[1] G. Winter *et al.* (2013). *Acta Cryst.* D**69**, 1260-1273. doi: 10.1107/S0907444913015308
[2] S. Monaco *et al.* (2013). *J. Appl. Cryst.* **46**, 804-810. doi:10.1107/S0021889813006195
[3] S. Delageniere *et al.* (2011). *Bioinformatics*, **27**, 3186-3192. doi:10.1093/bioinformatics/btr535

Gordon Leonard *is Structural Biology Group Leader at the European Synchrotron Radiation Facility.*

What metadata is needed to make Institut Laue Langevin neutron diffraction raw data intelligible for new users?

Matthew Blakeley

Institut Laue-Langevin, 71 Avenue des Martyrs, 38000 Grenoble, France
Email: blakeleym@ill.fr

Central facilities for neutron scattering and synchrotron X-ray sources in Europe are working together to develop and share infrastructure for the data collected there. Such co-operation should make it easier and more efficient for users to access and process their data, and provide more secure means of storage and retrieval. It should also increase the scientific value of the data by opening it up to a wider community for further analysis and fostering new collaborations between scientific groups. However, with these developments comes a need to define how raw data are stored and made accessible, and in particular, what metadata are included to allow diffraction data to be intelligible to new users. To this end, an ILL data policy (https://www.ill.eu/fr/users/ill-data-policy/) was established in 2012, and a number of tools (*e.g.* [i] https://data.ill.eu [ii] https://logs.ill.eu) are being developed. These currently allow experimental data (identified by a DOI) to be consulted and downloaded remotely and ultimately will allow for (re)processing and validation of experimental data.

Matthew Blakeley *is a Scientist at the Institut Laue-Langevin responsible for the quasi-Laue neutron diffractometer LADI-III.*

Metadata in high-pressure crystallography

Kamil Dziubek

LENS – European Laboratory for Non-Linear Spectroscopy, Sesto Fiorentino, Italy
Email: rumianek@amu.edu.pl

Kamil F. Dziubek[1] and Andrzej Katrusiak[2]
[1]LENS – European Laboratory for Non-Linear Spectroscopy, Sesto Fiorentino, Italy
[2]Faculty of Chemistry, Adam Mickiewicz University, Poznan, Poland

The deposition of metadata related to specific techniques used for crystallographic experiments can be simplified by formulating guidelines for their preparation. One of the experimental techniques quickly gaining ground in the field of crystallographic research is high-pressure diffraction studies. They involve additional equipment for pressure generation, pressure calibration, etc. The high pressure cell can interfere with the primary or diffracted beam, which can contaminate the diffraction patterns and introduce errors in reflection intensities. The experimental details are vital for the evaluation and analysis of the data, and therefore the metadata are needed to be stored along with the raw diffraction images.

The most essential descriptors concern: (1) orientation of the pressure cell with respect to the incident beam and the detector; (2) the sample preparation and shape, both for powder and single crystals; (3) a reference

to the high-pressure vessel, and for unique equipment the dimensions of its relevant components, such as the anvil design, gasket thickness, chamber diameter, backing-plate type; (4) chemical composition of the cell parts, *e.g.* anvils, gasket and backing plates, pressure-transmitting medium; (5) the method of fixing the sample in the high-pressure chamber, if used; (6) the method of positioning the pressure cell during the data acquisition; (7) the pressure-measurement method. This information is indispensable for reproducing the results of structural refinements from the raw data or for attempting other methods of refinement. The pressure transmitting medium can dramatically change the sample compression, due to its possible interaction with the sample (such as penetration into the pores) or hydrostatic limit of the medium. The sample history can also affect the results. If the sample was recrystallized *in situ* in isothermal or isochoric conditions, from solution or melt, the details of the crystallization protocol should be provided. Simple edition rules and a checklist can considerably simplify the deposition of metadata and increase their informative value.

The authors are representing the IUCr Commission on High Pressure, AK is the chair of the Commission. KFD gratefully acknowledges the Polish Ministry of Science and Higher Education for financial support through the 'Mobilność Plus' program.

Kamil Dziubek *was born in Poznan, Poland and studied chemistry with Prof. Andrzej Katrusiak at Adam Mickiewicz University. During PhD work on 'High Pressure Crystallization of Liquids' at Poznan University, he interned at the DESY synchrotron facility in Hamburg and at the Geophysical Laboratory, Carnegie Institution, Washington DC. He has served as a consultant to the commission on High Pressure of the International Union of Crystallography and now works at the European Laboratory for Non-Linear Spectroscopy.*

Creating and manipulating universal metadata definitions

James Hester

Australian Nuclear Science and Technology Organisation, New Illawarra Road, Lucas Heights, NSW 2234, Australia
Email: jamesrhester@gmail.com

Metadata discussions are often closely linked to particular formats. However, facts about the natural world cannot depend on the medium used to transmit those facts. It follows that we are able to completely describe our metadata without reliance on any particular file format, and that we can distil metadata definitions from pre-existing data transfer frameworks regardless of the particular format used. This promising, if obvious, general conclusion does not specify what information needs to be provided in our format-free metadata definition. Following Spivak and Kent [1] I suggest that it is sufficient that the metadata definitions can be expressed as functions mapping some domain to some range.

This talk will explore some of the implications of this approach, including the independence of file format and metadata specification, specification of algorithms for interconversion between data files in differing formats, unification of disparate metadata projects, and simple steps to produce a complete metadata description.

[1] Spivak D.I., Kent R.E. (2012) 'Ologs: A Categorical Framework for Knowledge Representation.' *PLoS ONE*, **7**(1):e24274. doi:10.1371/journal.pone.0024274

James Hester *is an Instrument Scientist at the Bragg Institute of the Australian Nuclear Science and Technology Organisation (ANSTO), a leading facility in the use of neutron scattering and X-ray techniques to solve complex research and industrial problems in many important fields. He is Chairman of COMCIFS.*

The Crystallographic Information Framework as a metadata library

Brian McMahon

IUCr, 5 Abbey Square, Chester CH1 2HU, UK
Email: bm@iucr.org

The Crystallographic Information File (CIF) was introduced as a data exchange standard in crystallography in 1991 [1] and has become embedded in the practice of single-crystal and powder diffractometry. Versions of CIF exist that describe macromolecular structure and diffraction images [2], so that CIF may be used anywhere in a data pipeline from image capture to publication in what has been called a 'coherent information flow' (also CIF!) [3].

In practice, slight differences in data model and file format have led to 'dialects' of CIF, which can coexist quite happily. However, there is a consequent barrier to full interoperability. A new version of the CIF format [3] will allow the development of a new generation of CIF 'dictionaries' (the formal data description schemas or 'ontologies'). This will allow fully automatic interconversion between existing CIF data files in either formalism, but has the added bonus of providing a descriptive framework for any type of crystallographic information. Formally, the CIF approach makes no distinction between 'data' and 'metadata', and so is arbitrarily adaptable and extensible to any domain of structural science or further afield. Since the CIF format has a very simple

syntactic structure which makes the contents very easy to read, CIF dictionaries can provide a simple template for developing new metadata schemas by working scientists who are not experts in informatics.

[1] Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography. *Acta Cryst.* A**47**, 655-685.
[2] Hall, S. R. & McMahon B. (eds) (2005). *International Tables for Crystallography, Volume G: Definition and exchange of crystallographic data.* Dordrecht: Springer. Corrected reprint (2010). Chichester: Wiley.
[3] McMahon, B. (2013). A coherent information flow in crystallography. Presentation at ECM28 Satellite Symposium on Crystallographic Information and Data Management, U. Warwick, UK. See also https://youtu.be/BiYETNUbfVo

Brian McMahon *is the Research and Development Officer at the International Union of Crystallography's offices in Chester, UK, and a former IUCr Representative to CODATA, the ICSU Committee on Scientific Data. He is Coordinating Secretary of COMCIFS and a Co-editor of International Tables for Crystallography Volume G: Definition and exchange of crystallographic data.*

Minutes of DDDWG meeting at IUCr Congress, Montreal

The following were posted on the DDDWG public discussion forum on September 30, 2014. They are reproduced here to provide more background to the activities iof the Working Group and the impetus for the current Workshop.

Diffraction Data Deposition Working Group Meeting Montreal 11 August 2014 7:30p.m.

Present: John R. Helliwell (Chair), Brian McMahon, Marian Szebenyi, Frances C. Bernstein, Herbert J. Bernstein, Loes Kroon-Batenburg, Patrick Mercier, Matt Zimmerman, Kamil F. Dziubek, Marcin Kowiel, Saulius Grazulis, George Phillips, Jim Kaduk, Andrew Allen.

A. Summary of DDDWG recommendations

John Helliwell (JRH) reviewed the recommendations of the DDDWG in the triennial report presented to the IUCr Executive Committee. The main issues that were identified as relevant for the IUCr were:

- With advancing technology and the consequent surge in volume of generated experimental data, it might be necessary to consider subsets of data for deposition/retention, or to retain for limited time periods.
- There is a need to address the question of rights of access to publicly funded but unpublished research data after some appropriate time lapse.
- In addition to the volume of data, there is a possible need for active triage of data at source owing to the rate of generation from the latest instruments or experiments (*e.g.* X-ray lasers, Eiger detectors).

Upcoming recommendations

- IUCr Commissions should be charged by the IUCr Executive to conclude their work projects to define their experimental raw data metadata.
- *Journal of Applied Crystallography* should consider introducing a 'difficult raw data' category of allowed articles.
- A centralised crystallographic repository of raw dataset metadata should be scoped and piloted.
- With such a repository in place, we should revisit the proposal that authors **shall** provide a permanent and prominent link from an article to the associated raw datasets.

B. Articles in press on raw data archiving and use in *Acta Crystallographica Section D*

A special set of four articles and an introduction by Tom Terwilliger were in proof for *Acta Cryst. D*, and publication in the October issue was anticipated.

C. 'Difficult data articles' category in *JAC*

Following a suggestion of Loes Kroon-Batenburg (LKB), one of the Main Editors of *JAC* (Anke Pyzalla) had been approached and was receptive to the idea of a new category of research articles in which authors would describe the nature of problematic or otherwise challenging data sets. These articles would invite the community to work on these data where there were potentially interesting results *e.g.* relating to multiple lattices, diffuse scattering, incommensurate structures *etc.* The author must explain in detail what analysis had already been done; the article would be peer-reviewed; and there would be a link to a repository where the data would be available for a reasonable timescale and should have a DOI (or other robust persistent identifier).

Andrew Allen (AA), another Main Editor of *JAC*, said that the idea was interesting, but required discussion and consultation among the full set of Main Editors, two of whom were relatively recent appointees. There might be concerns that such articles should not be understood as a dumping ground for low-quality work, and there were worries that they would be poorly (or not) cited, or seem to have an adverse effect on the journal's reputation or performance benchmarking by bibliometric criteria. In general discussion, it was emphasised that the actual number of such articles might be small, and that their acceptance criteria should be demanding, so that only articles of genuine scientific interest would be published. They might also be very well cited as exemplars for difficulty and also hopefully for successful community research action ('crowd research').

There was a general discussion of some of the technical issues that could be problematic if such a category were instituted. LKB anticipated that the most suitable contributions in such a category would be studies where the total accumulated datasets were of the order of GBytes in size, because of the time/bandwidth constraints in network transfer. [This did not rule out studies with large data generation rates, *e.g.* in XFELS experiments, where subsetting of the collected data was already routinely performed.] Herbert Bernstein (HJB) suggested that lossy compression (*e.g.* using techniques described by James Holton at the ECM Bergen Workshop) could be used for data transfer – this could make it easier for fellow-scientists to sample a difficult data set before deciding whether to invest time or effort in subsequent solution attempts.

AA commented that the proposed new category fell outside the traditional domain of interest of *JAC*. If it were introduced, it would be necessary to develop suitable guidelines for both Editors and authors. George Phillips suggested that the process would help to educate the community not only in technical matters but in terms of the policy and ethics associated with this new mode of working; he drew parallels with the ethics of structure-factor deposition in the macromolecular community.

AA confirmed the approach described by LKB that the data should remain with the original authors, and that the article and

network of DOI links to associated data sets formed an extension of the original research effort. If not done that way this could make the assignment and tracking of intellectual property rights complicated if other groups made use of the original data. JRH indicated that the emergence of data-centric licensing protocols such as CC0 was intended to help to address these concerns, but agreed that further work might be needed to define the possible IP issues more clearly.

Action: the DDDWG to send a formal request to the JAC Editors to consider this proposal. AA would act as point of contact. [LKB]

D. Review of DDDWG interactions with ResearchGate *etc.*

JRH reported on the current state of efforts in Manchester University to provide satisfactory archiving of some of his data sets. The data were now safely retained in the University data store, but DOIs had not yet been assigned. [Post meeting note by JRH: The Manchester University 'Data Librarian' has confirmed to JRH that weblink identifiers have been assigned for his datasets and a licence from Datacite sought with a view to commencing their DOI attributions in early 2015.] There seemed to be a general sense that Universities (and some research facilities) were happy to provide 'safe retention' policies, but were still reluctant to take on a fully fledged archiving role. Tom Terwilliger had approached ResearchGate, a growing social network provider for academic researchers, who are interested in retaining raw data.

LKB gave a comprehensive review of possible repositories for experimental data sets. Among possible solutions that the DDDWG had already given some thought to were arXiv.org (currently restricts supplementary data to a few Mbytes); Dryad; Figshare; ResearchGate; the PaNData project covering large facilities; TARDIS and the Store.Synchrotron initiative. Additional possibilities included Zenoob and Dataverse Network – the latter is implemented in Holland as EASY.

Saulius Grazulis (SG) outlined a possible approach to robust distributed data repositories built on a 'least-authority filesystem'. This allowed the configuration of multiple depositories sharing encrypted data, set up in such a way that any 3 from a pool of up to 255 nodes could retrieve any data set. Individual nodes could be of the order of 30 TB (probably affordable for a University), allowing an aggregate storage of several petabytes. Because the filesystem is encrypted, authors would need to provide authorised access to their holdings (but the security keys could be held in escrow during any embargo period). The Crystallography Open Database (COD) had plans to start a pilot project along these lines and would be interested in working with DDDWG to explore this approach. Nature Publishing Group (publishers of *Nature Scientific Data*) already listed COD as an approved repository for supplementary data. It was noted that data would only be recoverable from a least-authority filesystem if the encryption keys were not lost – their maintenance would need to be an important aspect of the maintenance metadata required of such a depository.

Action: SG to define the collaboration with DDDWG proposal in detail.

HJB reminded the meeting that Google still offered cost-effective large-scale storage at costs in the region of $120/TB/year. SG remarked that storage space rented from commercial suppliers such as Google could indeed be utilized within a least-authority filesystem solution. George Phillips favoured the idea of an early triage that identified high-value data sets that required particular effort to archive and maintain robustly. In his view, a trusted resource such as the Protein Data Bank was still the preferred option for such data (assuming there was sufficient community support to lead to any necessary increase in funding); but the possibility of retaining the lower-value residue also, in a lower-cost distributed repository, was appealing.

E. Other issues in archiving large data sets with journals, in University repositories or in funded public archives

There was some general discussion about the appropriate 'horizon' after which unpublished data should be released into the public domain, Marian Szebenyi thinking that 3 or even 5 years could be considered too short for data of high potential value. Frances Bernstein remarked that PDB release embargo periods used to be longer than is now the case, and had been lowered in accordance with community wishes. JRH emphasised that the DDDWG was not seeking to mandate a specific horizon, but that active discussion within the community (or communities) should take place to establish a timescale that represented a reasonable consensus. It was also pointed out that (unlike in space science, where a rocket launch marked a specific 'time zero') there might be difficulties in establishing the starting point from which an embargo period should properly be reckoned.

HJB revisited the idea of data 'triage' to reduce the volume of accumulated data. For some detectors there was natural triage, in that data frames were already discarded at source or stored using lossy compression. While he did approve of the general idea of retaining as much data as possible that was associated with published structures, he did argue for a strategy that tried to retain some lossy version of other data that would otherwise be completely discarded.

F Commission Reports

While the DDDWG was keen to encourage progress by Commissions in defining experimental raw data metadata, there was as yet no coordinated way to achieve this. Patrick Mercier (PM) noted that the Commission on Inorganic and Mineral Structures was aware of the requirement and wished to proceed, but needed help in starting. JRH suggested that the XAFS and SAS articles published in recent years in *Acta*, together with the forthcoming *Acta D* special articles, would form a useful reference and starting point. SG emphasised that the Crystallography Open Database would be happy to help where possible.

Action: PM to consult with his Commission colleagues to confirm if they now have enough details to proceed. [Post meeting note: Dr Simon Coles of Southampton University, UK could be approached by PM as a Commission consultant on such matters.]

G Next triennium

A complete work plan for the next triennium would depend on feedback from the Executive Committee to the Working Group report presented at this Congress. In anticipation of a renewed mandate, an application for IUCr funding had been submitted, to support a one-day workshop at the ECM meeting in Rovinj in August 2015. The DDDWG itself has presented the IUCr Executive Committee with a set of summary points for why it should continue (see Appendix below).

John R. Helliwell
Brian McMahon
September 2014

Appendix

To the IUCr Executive Secretary August 10th 2014

Dear Mike,

The DDDWG wishes to commend to the EC that the activities of the DDDWG should continue in the next Triennium and we unanimously offer our reasons below.

Yours sincerely
John

Prof John R Helliwell DSc

We offer the following good reasons to continue:-

1. *The raw data archiving and correct metadata logging may take longer than we initially imagined to get ingrained as community best practice and thereby we would be on hand as a DDDWG to push further where and when needed.*
2. *Increasingly the DDDWG is recognised as an established and experienced group for crystallographers, as well as central X-ray and neutron facilities, and the IUCr Executive Committee and its Commissions to consult on raw data archiving policy as well as technical matters.*
3. *Technical opportunities for raw data archiving are still evolving and we are better placed as a group rather than as individuals to follow what is happening.*

The above said we may need to add new specialists to the DDDWG. The most obvious is to have someone on board that represents those scientists that take many years to prepare experiments notably of difficult samples and who would be properly alert to mandates like "privileged access to measured data can be for a limited period before open access to those data by others would be required".

There may be other specialists that the EC would like to suggest.

About our sponsors

We acknowledge the generosity of many corporate sponsors who have made possible this Workshop. In particular, their financial contributions have allowed us to webcast the proceedings live and to retain a video archive for posterity, which will be made available through the IUCr website.

In addition to the companies and organisations listed below, we are grateful to the International Union of Crystallography, which provided the base funding for the Workshop, and to the Croatian Association of Crystallographers, who have worked tirelessly to provide logistical support.

Dectris is the technology leader in X-ray detection. The DECTRIS photon counting detectors have transformed basic research at synchrotron light sources, as well as in the laboratory and with industrial X-ray applications. DECTRIS aims to continuously improve measurement quality, thereby enabling new scientific findings. This pioneering technology is the basis of a broad range of products, all scaled to meet the needs of various applications. DECTRIS also provides solutions for customer developments in scientific and industrial X-ray detection.

DECTRIS was awarded the 2010 Swiss Economic Award in the High-Tech Biotech category, the most prestigious prize for start-up companies in Switzerland.

IUCr Journals are published by the International Union of Crystallography, an International Scientific Union whose objectives are to promote international cooperation in crystallography and to contribute to the advancement of crystallography in all its aspects.

The IUCr contributes to these objectives by publishing high-quality crystallographic research in nine primary scientific journals: *Acta Crystallographica Section A: Foundations and Advances*; *Acta Crystallographica Section B: Structural Science, Crystal Engineering and Materials*; *Acta Crystallographica Section C: Structural Chemistry*; *Acta Crystallographica Section D: Biological Crystallography*; *Acta Crystallographica Section E: Crystallographic Communications*; *Acta Crystallographica Section F: Structural Biology Communications*; *Journal of Applied Crystallography*; *Journal of Synchrotron Radiation*; and, launched for the International Year of Crystallography, *IUCrJ*, a gold open-access title publishing articles in all of the sciences and technologies supported by the IUCr.

CODATA, the Committee on Data for Science and Technology, is an interdisciplinary Scientific Committee of the International Council for Science (ICSU), established in 1966 to promote and encourage, on a world-wide basis, the compilation, evaluation and dissemination of reliable numerical data of importance to science and technology.

The mission of CODATA is to strengthen international science for the benefit of society by promoting improved scientific and technical data management and use.

It works to improve the quality, reliability, management and accessibility of data of importance to all fields of science and technology. CODATA provides scientists and engineers with access to international data activities for increased awareness, direct cooperation and new knowledge. It is concerned with all types of data resulting from experimental measurements, observations and calculations in every field of science and technology, including the physical sciences, biology, geology, astronomy, engineering, environmental science, ecology and others. Particular emphasis is given to data management problems common to different disciplines and to data used outside the field in which they were generated.

The **Cambridge Crystallographic Data Centre** (CCDC) is dedicated to the advancement of chemistry and crystallography for the public benefit through providing high quality information, software and services.

Chemists in academic institutions and commercial operations around the world rely on the CCDC to deliver the most comprehensive and rigorous molecular structure information and powerful insights into their research.

The CCDC is a non-profit organisation and a registered charity, supported entirely by software subscriptions from its many users. The CCDC compiles and distributes the Cambridge Structural Database (CSD), the world's repository of experimentally determined organic and metal–organic crystal structures. It also develops knowledge bases and applications which enable users quickly and efficiently to derive huge value from this unique resource.

Bruker Corporation has been driven by the idea to always provide the best technological solution for each analytical task for more than 50 years.

Today, worldwide more than 6,000 employees are working on this permanent challenge at over 90 locations on all continents. Bruker systems cover a broad spectrum of applications in all fields of research and development and are used in all industrial production processes for the purpose of ensuring quality and process reliability.

Bruker continues to build upon its extensive range of products and solutions, its broad base of installed systems and a strong reputation among its customers. Being one of the world's leading analytical instrumentation companies, Bruker is strongly committed to further fully meet its customers' needs as well as to continue to develop state-of-the-art technologies and innovative solutions for today's analytical questions.

FIZ Karlsruhe is a leading international provider of scientific information and services. Our mission is to supply scientists and companies with professional research and patent information as well as to develop innovative information services. As a key player in the information infrastructure we pursue our own research program and also cooperate with leading universities and research associations.

The Inorganic Crystal Structure Database (ICSD) is the worlds biggest database of fully evaluated and published crystal structure data. Science and industry are offered high-quality records that will provide a basis for studies in materials science, *e.g.* for identifying unknown substances. ICSD contains more than 165,000 crystal structures of inorganic substances published since 1913. Metals were included in ICSD several years. The metal structures were recorded retroactively in cooperation with FIZ Karlsruhe's partner, NIST (National Institute for Science and Technology, Washington, DC, USA).

FIZ Karlsruhe is a non-profit corporation and the largest non-university institution for information infrastructure in Germany. FIZ Karlsruhe is a member of the Leibniz Association, which comprises almost 90 institutions involved in research activities and/or the development of scientific infrastructure.

Oxford Cryosystems is a market-leading manufacturer of specialist scientific instrumentation and software. The origins of the company lie in the design and manufacture of the original Cryostream Cooler in 1985, which immediately became the system of choice for cooling samples in X-ray diffraction experiments. The range of products for use in sample cooling has expanded over the last twenty-five years to include liquid-free systems, helium coolers and specially adapted systems for use with powder samples. Today the company is considered to be the global market leader in X-ray diffraction sample cooling.

WILEY

Wiley's Scientific, Technical, Medical, and Scholarly (STMS) business, also known as Wiley-Blackwell, serves the world's research and scholarly communities, and is the largest publisher for professional and scholarly societies. Wiley-Blackwell's programs encompass journals, books, major reference works, databases, and laboratory manuals, offered in print and electronically. Through Wiley Online Library, we provide online access to a broad range of STMS content: over 4 million articles from 1,500 journals, 9,000+ books, and many reference works and databases. Access to abstracts and searching is free, full content is accessible through licensing agreements, and large portions of the content are provided free or at nominal cost to nations in the developing world through partnerships with organizations such as HINARI, AGORA, and OARE.
