Research data management

A Satellite Workshop to the 67th Annual Meeting of the American Crystallographic Association

Programme and background materials

Organized by DDDWG (the IUCr Diffraction Data Deposition Working Group)

This is the third and final workshop organized by the Diffraction Data Deposition Working Group (DDDWG), 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 takes the form of a full-day workshop at the 2017 American Crystallographic Association Meeting with lectures from crystallographic practitioners, data management specialists and standards maintainers.

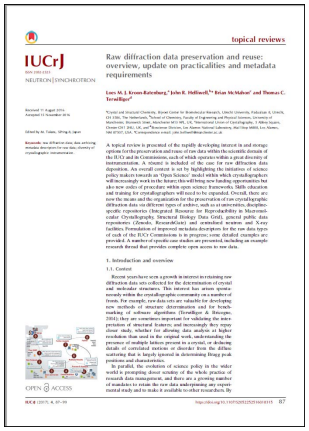
Objective: This workshop has two plenary sessions:

What every experimentalist needs to know about recording essential metadata of raw diffraction data
This will include sample preparation and characterization; correct recording of instrument axes, correction factors, calibration – instrument manufacturers; attention to diffuse scattering or other interesting 'metadata'.

Research data management policy mandates and requirements on Principal Investigators (PIs)
This will include metadata standardization; data repositories; primary data linking to publications.

There will also be an optional technical session:

High-data-rate/high-performance-computing issues in macromolecular crystallography
For synchrotron- and XFEL-based macromolecular crystallography (MX), high source brightness and the new generation of pixel array detectors raise big-data, high-performance-computing and high-performance-networking issues in research data management. There will be an optional early evening sub-session of the Research Data Management workshop to discuss the high-data-rate/high-performance-computing issues of research data management for MX that will include discussion of appropriate hardware choices and programming techniques that are useful in this context. All registrants for the Research Data Management workshop are welcome to attend this workshop sub-session.



This booklet contains the following background material:

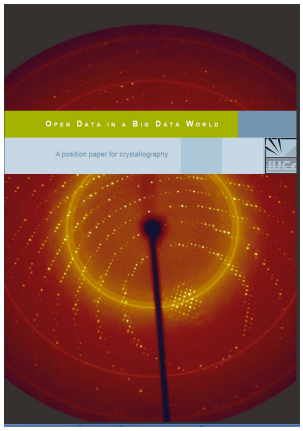
A recent *Topical Review* that surveys metadata requirements following the Rovinj Workshop:

- Kroon-Batenburg, L. M. J., Helliwell, J. R., McMahon, B. & Terwilliger, T. C. (2017). Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements. *IUCrJ*, **4**, 87–99.

The IUCr official response to the 2015 Science International Accord:

- Hackert, M. L., Van Meervelt, L, Helliwell, J. R. & McMahon, B. *Open Data in a Big Data World. A position paper for crystallography*. Chester, UK: International Union of Crystallography.

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



Welcome

This is the third and final workshop of the IUCr Diffraction Data Deposition Working Group (DDDWG). The Working Group was established at the 2011 Madrid Congress of the IUCr with the following brief:

> *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.*

A one-day Workshop at the Bergen European Crystallographic Meeting (6 August 2012) provided an overview of the needs, benefits and challenges of routine deposition of diffraction images, and existing or projected mechanisms for achieving this. The abstracts and presentations from this Workshop are available at http://www.iucr.org/resources/data/dddwg/bergen-workshop

Following the Workshop, a number of special articles were commissioned by the DDDWG to analyse some of the issues identified:

Terwilliger, T. C. (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. C. & 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.

A specific outcome of this work was the recognition that extensive and high-quality metadata were a prerequisite for effective data archiving, and that there was much need for metadata standardization in other crystallographic and structural-science experimental techniques. Metadata in the context of crystallography is generally handled within the Crystallographic Information Framework (CIF) managed by the IUCr Committee for the Maintenance of the CIF Standard (COMCIFS), and members of the DDDWG were closely involved with a COMCIFS Workshop and Symposium at the Warwick European Crystallographic Meeting (23–25 August 2013). The Workshop explored technical aspects of CIF dictionary design and the development of a methods dictionary definition language to facilitate data validation and extraction, while the symposium covered a wide range of information management topics. Materials from these events (including full video recordings from the symposium) are available at http://www.iucr.org/resources/cif/comcifs/workshop-2013 and http://www.iucr.org/resources/cif/comcifs/symposium-2013

The second full DDDWG Workshop was a two-day event at the Rovinj European Crystallographic Meeting (22–23 August 2015) that focussed on *Metadata for raw data from X-ray diffraction and other structural techniques*, although it also presented the many initiatives now springing up to handle large volumes of raw diffraction data. Abstracts and presentations (including full video recordings) from this event are available at http://www.iucr.org/resources/data/dddwg/rovinj-workshop

Many of the considerations discussed at this Workshop were reviewed in a recent publication:

Kroon-Batenburg, L. M. J., Helliwell, J. R., McMahon, B. & Terwilliger, T. C. (2017). Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements. *IUCrJ*, **4**, 87–99.

The DDDWG will present its final report to the IUCr Executive Committee at the Hyderabad Congress in 2017. Future data-related activites will be overseen by a new Committee on Data of the IUCr, which will carry forward many of the ideas, policies and practical recommendations advanced by the DDDWG through its activities over the past six years. This final formal Workshop will review the overall state of research data management in crystallography, with particular emphasis on what experimentalists need to know about recording essential metadata for the archiving and re-use of raw data.

We hope that you enjoy the day and learn a great deal from it, and we welcome your contributions in the General Discussion sessions that have been scheduled throughout the Workshop.

John Helliwell
Brian McMahon
Tom Terwilliger

Timetable

Friday May 26

8.30 am Introduction to the DDDWG 2017 Workshop on Research Data Management.
John R. Helliwell and Brian McMahon

Session I: What every experimentalist needs to know about recording essential metadata of primary (raw) diffraction data

8.40 am The Science International Accord on *Open Data in a Big Data World* and the IUCr's response
Marvin L. Hackert, Luc Van Meervelt, John R. Helliwell and Brian McMahon

9.00 am What every experimentalist needs to know about recording essential metadata of primary (*i.e.* raw) diffraction data
Herbert J. Bernstein

9.30 am Correct recording of metadata: towards archiving and re-use of raw diffraction images in crystallography
Loes M. J. Kroon-Batenburg

10.00 am Coffee break

10.30 am Research data management at CHESS
D. Marian Szebenyi, Devin Bougie, Aaron Finke, Richard Gillilan, Jesse Hopkins, David Schuller and Werner Sun

11.00 am Metadata for small-angle scattering measurements
Andrew Allen, Fan Zhang, Jan Ilavsky and Pete Jemian

11.30 am General Discussion

12.00 noon Lunch

Session II: Research data management policy mandates and requirements on Principal Investigators (PIs)

1.00 pm Open Science and research data policy mandates and requirements on Principal Investigators (PIs)
Marshall Ma and Simon Hodson

1.30 pm Research data management: structure factors and atomic coordinates
Stephen Burley

2.00 pm The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC)
Wladek Minor

2.30 pm Research data mangement: administration, raw diffraction data, structure factors and coordinates at the UK's National Crystallographic Service (NCS)
Simon Coles

3.00 pm SBGrid Databank
Peter Meyer, Stephanie Socias, Jason Key, Mercè Crosas and Piotr Sliz

3.30 pm General Discussion

4.00 pm Tea break

Session III: High-data-rate/high-performance-computing issues of research data management in macromolecular crystallography

4.15 pm Dealing with the avalanche of data generated in high-data-rate macromolecular crystallography
Jean Jakoncic, Herbert J. Bernstein, Alexei Soares, Wuxian Shi, Martin Fuchs, Robert Petkus, Robert Sweet and Sean McSweeney

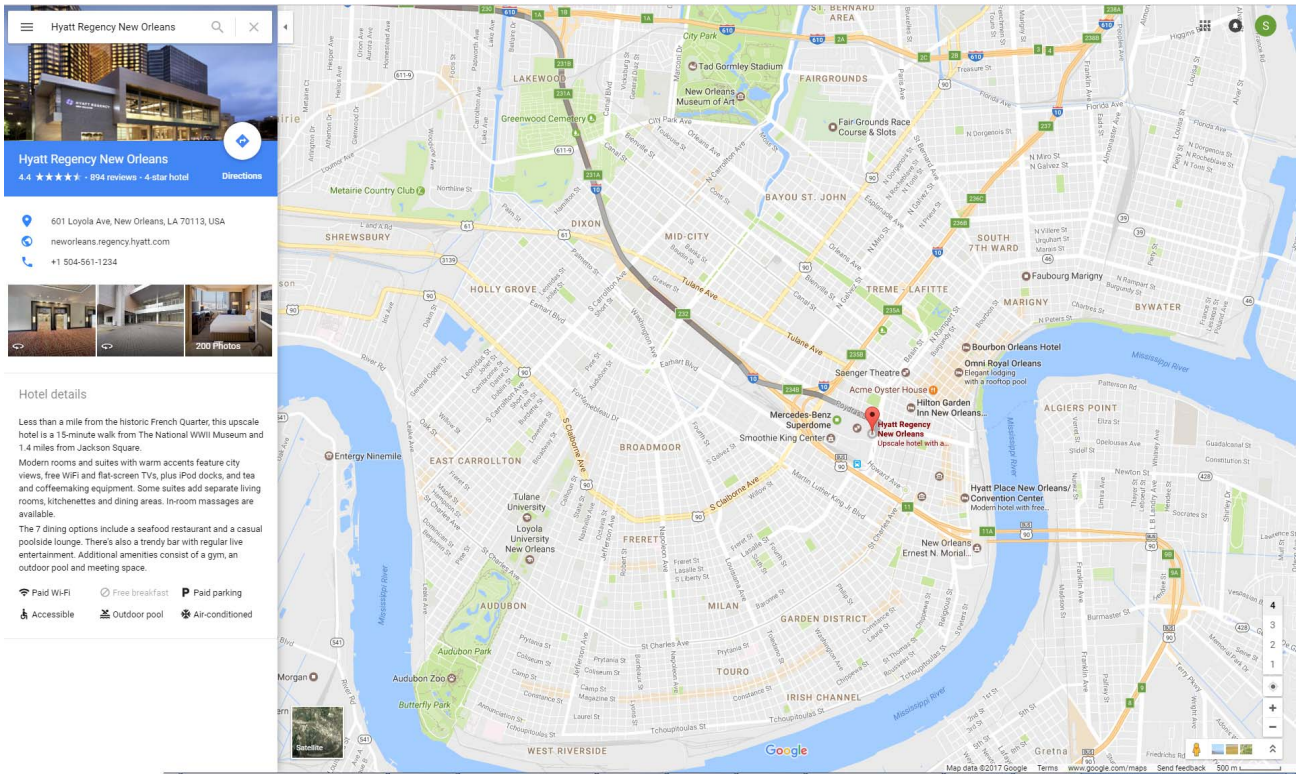
4.45 pm Intel Scalable System Framework
Henry Gabb

5.15 pm Intel software and programming tools ecosystem for HPC
Henry Gabb

5.45 pm General Discussion

6.15 pm Close

6.30 pm ACA2017 Opening Ceremony (Celestin A & B)



Abstracts

Introduction to the DDDWG 2017 Workshop on Research Data Management

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

The IUCr Executive Committee established a Diffraction Data Deposition Working Group (DDDWG) to define the need for and practicalities of routine deposition of primary experimental data in X-ray diffraction and related experiments. Since the Working Group's first Workshop in Bergen, Norway (August 2012), important strides have been taken to make routine deposition of raw data a reality. The major facilitator for this has been the establishment of digital data storage repositories registered to issue persistent unique Digital Object Identifiers (DOIs) for a raw dataset. Such repositories include universities (*e.g.* University of Manchester), the EU's Zenodo initiative, and several centralized neutron, synchrotron and X-ray laser facilities. As stressed by John Westbrook of the PDB (http://www.iucr.org/resources/data/dddwg/bergen-workshop), metadata descriptors for raw data are vital for its effective re-use. The PDB has extensive experience of specifying metadata descriptors for structure factors, coordinates and *B* factors, as well as for cryoEM and bioNMR data depositions. Kroon-Batenburg and Helliwell [1] provided an example of appropriate metadata, critically including a picture of their diffractometer, for their local raw diffraction data archive. This archive has seen successful examples of raw data re-use such as by Wladek Minor and collaborators [2]. A second DDDWG workshop on 'Metadata for Raw Data' (Rovinj, Croatia, August 2015) brought together another wide range of global experts (http://www.iucr.org/resources/data/dddwg/rovinj-workshop), including the Chair of the IUCr Committee for the Maintenance of the CIF Standard (James Hester), who has vast experience of metadata descriptors for processed and derived data. An outcome of the second Workshop was '*checkCIF* for raw diffraction data', a notional service akin to the existing IUCr *checkCIF* for processed structure factors and derived atomic coordinates data (http://checkcif.iucr.org). This third Workshop at ACA 2017, New Orleans, broadly titled 'Research Data Management', includes the charge to Workshop participants to focus on metadata (including their experiences with processed structure factors and derived atomic coordinates data), and help to define as closely as possible the optimum metadata for raw diffraction data to guide the raw data archives listed above. Re-use of raw data leveraged upon metadata descriptions has already been shown to be viable [1,2]. This should now be built on more energetically by the single-crystal diffraction community (including chemical crystallography), as well as by the various scattering, diffraction, imaging and spectroscopy techniques represented in the various IUCr Commissions. Excellent headway has been made in defining SAXS and EXAFS metadata, for example. For an overview of raw diffraction data preservation and re-use including an update on practicalities and metadata requirements see the very recent publication by Kroon-Batenburg *et al.* 2017 [3].

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

[2] Shabalin, I., Dauter, Z., Jaskolski, M., Minor, W. & Wlodawer, A. (2015). *Acta Cryst.* D**71**, 1965–1979.

[3] Kroon-Batenburg, L. M. J., Helliwell, J. R., McMahon, B. & Terwilliger, T. C. (2017). *IUCrJ*, **4**, 87–99.

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.*

Session I: What every experimentalist needs to know about recording essential metadata of primary (raw) diffraction data

The Science International Accord on *Open Data in a Big Data World* and the IUCr's response

Marvin L. Hackert

Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA
Email: m.hackert@austin.utexas.edu

Marvin L. Hackert[1], Luc Van Meervelt[2], John R. Helliwell[3] and Brian McMahon[4]
[1] Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA
[2] Chemistry Department, Universiteit Leuven, Celestijnenlaan 200F, B-3001 Leuven, Belgium
[3] School of Chemistry, University of Manchester, M13 9PL, UK. Email: john.helliwell@manchester.ac.uk
[4] IUCr, 5 Abbey Square, Chester CH1 2HU, UK. Email: bm@iucr.org

Science is best served when access barriers to data (and publications) are low. *Open Data in a Big Data World* [1] is a response by the IUCr to an international Accord [2] by ICSU, IAP, TWAS and ISSC in an emerging scientific culture of big data on the values of open data that are discoverable, accessible, intelligible, assessable and usable. Technological advances in scientific instrumentation and computer technology have dramatically increased the quantities of data involved in scientific inquiry. The Accord expresses the dependence of scientific assertions on supporting data and asserts that 'openness and transparency are the bedrock of modern science.' The IUCr supports this assertion, but argues that such data should also be subject to scrutiny through peer review and automated validation where possible to look for systematic bias or error. An overlooked challenge in handling ever-growing volumes of data is the need to apply the same level of critical evaluation as has historically been applied to smaller data sets. Any software implementations used to scrutinize such data should employ open algorithms where results could be cross-checked by independent implementations.

A major barrier to access is cost. Evaluating, storing and curating quality data is an expensive component of the scientific process, and care must be taken to understand how to obtain the maximum benefit from public funding of science.

[1] http://www.iucr.org/__data/assets/pdf_file/0011/125687/OpenData_crystallography_web.pdf

[2] http://www.icsu.org/science-international/accord

Marvin L. Hackert *is President of the International Union of Crystallography (2014–2017) and is William Shive Centennial Professor of Biochemistry and Chairman, Department of Chemistry and Biochemistry, University of Texas at Austin. His research interests are in structural molecular biology. A primary research focus has been structure/function relationships of pyruvoyl- and PLP-dependent enzymes using biochemical and protein crystallographic techniques.*

What every experimentalist needs to know about recording essential metadata of primary (*i.e.* raw) diffraction data

Herbert J. Bernstein

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

As the rate of production of diffraction images rises to several hundred datasets per day per beamline, it is becoming increasingly important to record essential metadata in an efficiently retrievable form. It is impractical to expect to refer to laboratory notebooks and do manual metadata entry in such an environment. Indeed, as data rates increase further it will become impractical to handle the same images multiple times in order to transform metadata from one convention to another. The last time our community faced a similar speed-constrained transition was with the Dectris Pilatus pixel-array detectors which strained computers and networks of that time by producing ten images per second, leading to the adoption of the imgCIF/CBF and miniCBF metadata conventions. Now, with data arriving one to three orders of magnitude faster and the introduction of NeXus/HDF5 images, and adoption of new experimental techniques including serial synchrotron crystallography, adoption of consistent, well-documented crystallographic-image metadata handling is essential to conserve processing resources and maximize beamline structure production. To this end, the necessary concordances of imgCIF/CBF - miniCBF - NeXus NXmx metadata specifications [1] [2] [3] are being maintained on a common web site. In this talk we review compromises between a common minimal set of metadata to allow for processing of simple rotation data and richer sets of metadata needed for more demanding experiments. We also consider the implications of these choices for future reprocessing of archived datasets.

[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] M. Mueller, '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.

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).*

Correct recording of metadata: towards archiving and re-use of raw diffraction images in crystallography

Loes Kroon-Batenburg

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

In recent years scientists and policy makers have made major steps toward Open Science. The incentive is to allow validation and falsification of the research based on the data and to allow its re-use, as the aquisition of the data is mostly funded by the tax payer. New methods and technologies can be developed with the availability of large data bases covering diverse types of experiments. In this framework the IUCr established a Diffraction Data Deposition Working Group (DDDWG) with the aim of developing standards for the representation of raw diffraction data in crystallography. Two key issues play a role: the importance of persistent identifiers and the full recording of metadata. Whilst discussions are vividly going on about what data to archive, only those related to published papers or also of incomplete or unsuccessful research that could be particularly interesting for the development of new science, the field should prepare itself for depositing fully self-contained data. A recent review [1] summarizes the ongoing developments. Ideally, metadata should comprise the following: identification of the image format, number of pixels, pixel sizes, byte-storage architecture, baseline offset and handling of overflows, information on the corrections that are applied (dark current, distortion correction, non-uniformity correction), detector gain, goniometer axes orientations and rotation directions, and information on the experiment such as exposure time, number of repeats, oscillation axis and range, wavelength used, beam polarization, detector position (or beam position) and offsets. Details and the importance of such information will be discussed. The necessity to use a structured language (DDL) that defines data names (tags) in data formats like CIF or Nexus [2] to ensure unambiguous interpretation, will be demonstrated. Awareness of detector manufacturers and experimentalists of recording sufficient metadata is essential, and guidelines for these are under way.

[1] Kroon-Batenburg, L.M.J., Helliwell, J.R., McMahon, B. & Terwilliger, T.C. (2017). *IUCrJ*, **4**, 87–99.

[2] Bernstein, H.J., DDDWG Workshop (2015). http://www.iucr.org/resources/data/dddwg/rovinj-workshop

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.*

Research data management at CHESS

D. Marian Szebenyi

MacCHESS and CHESS, Cornell University, Ithaca, NY 14853, USA
Email: dms35@cornell.edu

D. Marian Szebenyi, Devin Bougie, Aaron Finke, Richard Gillilan, Jesse Hopkins, David Schuller and Werner Sun, MacCHESS and CHESS, Cornell University, Ithaca, NY 14853, USA

Historically, the Cornell High Energy Synchrotron Source, CHESS, with its relatively small number of beamlines, has relied on users to manage their own data. The facility has provided adequate RAID storage at each station for a 6–8 week run, with some longer term backup. The advent of increasing numbers of experiments involving massive amounts of data has strained this system. Accordingly, we have recently implemented a large, centralized, more organized, system ('CHESS DAQ'), with separate storage for raw data, metadata, and general user data. Nightly incremental backups and full archiving at the end of each run protect against data loss. This system is used for most experiments at CHESS, with individual variations to suit the needs of users and staff. Our primary goal has been, and remains, to facilitate research by our users, by providing them the means to collect, process, and store the most useful data possible, while avoiding excessive bureaucracy.

BioSAXS raw data from SAXS and WAXS detectors (the two detectors record images simultaneously), as well as metadata, are written directly to CHESS DAQ. Processing is carried out locally on copies of the raw data, and processed data are backed up locally as well as on the DAQ and to user-supplied media.

Raw crystallographic data from the dedicated MX station, *i.e.* diffraction images, are stored locally, with users responsible for processing data and transferring raw and processed data to their home labs. Raw data are kept on-line for a few weeks and off-line for several years. Limited metadata are stored in image headers. Implementation of a new database, to facilitate organization of raw data and metadata, is under development in parallel with adoption of a new user interface for data collection, based on JBluIce.

Marian Szebenyi *is a staff scientist, and (since 2008) serving as Director, at MacCHESS, the NIH-supported resource for macromolecular diffraction at the Cornell High Energy Synchrotron Source. She has been heavily involved in the development of data collection software and hardware for macromolecular crystallography at CHESS, and in the establishment of policies regarding users' data handling.*

Metadata for small-angle scattering measurements

Andrew J. Allen

Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
Email: andrew.allen@nist.gov

Andrew J. Allen[1], Fan Zhang[1], Jan Ilavsky[2] and Pete R. Jemian[2]
[1]Material Measurement Laboratory, National Institute of Standards and Technology, 100 Bureau Drive, Gaithersburg, MD 20899, USA
[2]X-ray Science Division, Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA

Measurements based on small-angle scattering (SAS) of X-rays or neutrons (SAXS or SANS) differ critically in several ways from those based on X-ray or neutron Bragg diffraction (XRD or ND) or on X-ray or neutron spectroscopic methods. XRD or ND measurements yield diffraction peaks at discrete scattering angles or scattering vectors, Q, from which a pattern may be identified, and from there the underlying crystal structure. Similarly, spectroscopic measurements frequently yield information directly relatable to bond energies or energies of transition within the sample material. In contrast, SAXS or SANS measurements yield data that usually comprise a smooth curve of SAS intensity as a function of scattering angle or Q. This requires interpretation in terms of the likely scattering features (inhomogeneities) that underlie the sample microstructure before a quantifiable data analysis can be carried out in any meaningful way. Thus, in archiving SAXS or SANS data, very significant emphasis is required on the metadata to accompany the measured data – both metadata providing detailed qualitative information on sample microstructures, and metadata providing detailed instrumental parameters and other information on the measurements, themselves.

Metadata requirements for SAS are inextricably linked to aspects that may be more-or-less closely related to the measurements, themselves. Examples might include the measurement configuration (SAXS versus SANS, transmission versus grazing-incidence geometry, 1D Bonse-Hart versus 2D pinhole camera, angular-dispersive SAXS or SANS versus time-of-flight SANS, *etc.*), the nature of the sample (*e.g.* precipitates in metallic alloys, pores in ceramics, polymer structures, nanoparticles in suspension, protein complexes, expected polydispersity in feature size and shape, *etc.*), absolute calibration and correction issues (e.g. for scattering geometry and Q-values, scattering intensity, effective sample volume), the effective spatial and Q-resolution, background subtraction issues, and even the requirements for common data formats and publication standards. This paper will discuss these issues and current ongoing international efforts within the SAS community to address them.

Andrew Allen *is a Physical Scientist in the Materials Structure and Data Group of the Materials Measurement Science Division at NIST. He is a Main Editor of Journal of Applied Crystallography.*

Session II: Research data management policy mandates and requirements on Principal Investigators (PIs)

Open Science and research data policy mandates and requirements on Principal Investigators (PIs)

Marshall Ma

Department of Computer Science, University of Idaho, 875 Perimeter Drive MS 1010, Moscow, ID 83844-1010, USA
Email: max@uidaho.edu

Marshall Ma[1] and Simon Hodson[2]
[1]Department of Computer Science, University of Idaho, Moscow, ID 83844-1010, USA
[2]Executive Director, CODATA, 5 rue Auguste Vacquerie, 75016 Paris, France

This invited presentation will explore the policy landscape relating to research data. It aims to cast light on the latest developments in funder, institutional and journal policies and to clarify a number of issues relating to Open Science, Open Data, FAIR Data, Research Data Management *etc.* A simplified but useful and well-tested typology describes three categories of publicly-funded research data: (1) data resulting from large data creation/collection exercises that are often cumulative (*e.g.* EO/remote sensing, statistical data, meteorological data, refined crystallographic data); (2) full datasets created by funded research projects; (3) data that directly underpins research publications as the evidence [often a subset of (2)]. The presentation will analyse the data policies that exist in relation to these 'types' of data and the requirements they impose upon Principal Investigators and other parties.

The presentation will examine the benefits and challenges of Open Science and FAIR data in relation to the following issues and developments:

• the major transformations and opportunities described in the Science International Accord on *Open Data in a Big Data World*;

• the implications for peer-review, for the scrutiny and validation of data, and for the way in which scientific contribution is assessed and recognised;

• the need and opportunities for international development and coordination of standards and vocabularies within and across established disciplines;

• the funding, governance and economic challenges for data resources being addressed by the CODATA–OECD Global Science Forum Project on Business Models for Sustainable Research Data Repositories.

Xiaogang (Marshall) Ma *is an assistant professor of computer science at the University of Idaho. He received his PhD degree of Earth Systems Science and GIScience from University of Twente, Netherlands in 2011, and then completed postdoctoral training in Data Science and Semantic eScience at Rensselaer Polytechnic Institute. His research focuses on deploying data science in the Semantic Web to support cross-disciplinary collaboration and scientific discovery, with broad interests in participatory knowledge engineering, data interoperability and provenance, and visualized exploratory analysis of Big and Small Data. Ma was one of the four invited early-career panelists at the International Data Forum 2016. He is active in international societies of data science and geoinformatics, including CODATA, ESIP, RDA, AGU and IAMG.*

He is currently co-Chair of the CODATA Task Group looking at 'Coordinating Data Standards amongst Scientific Unions' (http://www.codata.org/task-groups/coordinating-data-standards) which stands at the heart of the CODATA/ICSU-sponsored initiative and meeting entitled 'Inter-Union Workshop on 21st Century Scientific and Technical Data: Developing a roadmap for data integration'.

Ma received the IAMG Vistelius Research Award in 2015 and the inaugural ICSU-WDS Data Stewardship Award in 2014. He won the ESIP Funding Friday Competition Award twice in 2013 and 2012.

Research data management: structure factors and atomic coordinates

Stephen K. Burley

Director, RCSB Protein Data Bank, Institute for Quantitative Biomedicine, Rutgers, The State University of New Jersey, 174 Frelinghuysen Road, Piscataway, NJ 08854, USA
Email: stephen.burley@rcsb.org

The Protein Data Bank (PDB; pdb.org) was established in 1971 as the first open-access digital data resource in biology. Today, the PDB archive serves as the single global repository for more than 125,000 experimentally determined atomic-level structures of biological macromolecules (protein, DNA, RNA) and their complexes.

The worldwide PDB (wwPDB; wwpdb.org) partnership, the international collaboration that manages the PDB archive, supports Deposition, Biocuration, Validation, and Distribution of PDB data. The mission of the wwPDB organization is to ensure that the PDB archive will continue in perpetuity as a high-quality, open-access digital data resource with no limitations on usage.

Through its global collaboration, the wwPDB has developed OneDep, a unified platform for Deposition, Biocuration, and Validation of 3D biological macromolecules experimentally determined by X-ray crystallography, NMR spectroscopy, and 3D Electron Microscopy. Data are submitted to the PDB archive via this OneDep system. OneDep is designed to help the wwPDB and the global structural biology research community meet the challenges of rapidly changing technologies and keep pace with evolving data archiving needs over the coming decades. The PDB archive and the OneDep system are underpinned by an extensible data architecture based on the PDBx/mmCIF dictionary (mmcif.wwpdb.org). Community involvement in the development of this data dictionary is coordinated by the wwPDB PDBx/mmCIF Working Group (Chaired by Paul Adams, LBL/UC Berkeley).

At present, $\sim$ 90% of PDB holdings were derived from diffraction methods. The earliest PDB entry in the PDB archive for which structure factors are available was deposited in 1976. Deposition of structure factors became mandatory in 2008, and $\sim$ 90% of all crystallographic entries now include these data.

Management of structure factors and atomic coordinates within the PDB archive will be discussed, with emphasis on current efforts to extend the range and the complexity of the diffraction data and metadata items that can be deposited.

Acknowledgements: The RCSB Protein Data Bank (RCSB PDB; rcsb.org) is a founding member of the Worldwide Protein Data Bank organization (wwPDB; wwpdb.org). Additional members of the wwPDB include the Protein Data Bank in Europe (PDBe; pdbe.org), Protein Data Bank Japan (PDBj; pdbj.org), and BioMagResBank (BMRB; bmrb.org). Core RCSB PDB operations are funded by a grant to SKB (NSF DBI-1338415) from the National Science Foundation, the National Institutes of Health, and the US Department of Energy.

Stephen Burley *MD, DPhil, is an expert in structural biology, proteomics, bioinformatics, structure/fragment based drug discovery, and clinical medicine/oncology. Burley currently serves as a Distinguished Professor in the Department of Chemistry and Chemical Biology, Director of the Center for Integrative Proteomics Research, and Director of the RCSB Protein Data Bank at Rutgers, The State University of New Jersey. He is also the Founding Director of the Institute for Quantitative Biomedicine at Rutgers and a Member of the Rutgers Cancer Institute of New Jersey.*

From 2008 to 2012, Burley was a Distinguished Lilly Research Scholar in Lilly Research Laboratories. Prior to joining Lilly, Burley served as the Chief Scientific Officer and Senior Vice President of SGX Pharmaceuticals, Inc., a publicly traded biotechnology company that was acquired by Lilly in 2008. Until 2002, Burley was the Richard M. and Isabel P. Furlaud Professor at The Rockefeller University, and an Investigator in the Howard Hughes Medical Institute.

He has authored/coauthored more than 250 scholarly scientific articles. He is a Fellow of the Royal Society of Canada and of the New York Academy of Sciences. Burley received an MD degree from Harvard Medical School in the joint Harvard-MIT Health Sciences and Technology program and, as a Rhodes Scholar, received a DPhil in Molecular Biophysics (structural biology) from Oxford University. He trained in internal medicine at the Brigham and Women's Hospital, and did post-doctoral work with Gregory A. Petsko at the Massachusetts Institute of Technology and William N. Lipscomb at Harvard University. With William J. Rutter and others at the University of California, San Francisco and Rockefeller, Burley co-founded Prospect Genomics, Inc., which was acquired by SGX in 2001.

The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC)

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 Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) has been developed as part of the BD2K (Big Data to Knowledge) NIH project to archive raw data from diffraction experiments and, more importantly, to extract metadata from diffraction images alone, or from a combination of information obtained from a PDB deposit and diffraction images. As of February 2017, the IRRMC resource contained indexed data from 3235 macromolecular diffraction experiments (6189 data sets), accounting for around 3% of all structures in the Protein Data Bank (PDB). The IRRMC utilizes a distributed storage system implemented with a federated architecture of many independent storage servers, which provides both scalability and sustainability. The resource, which is accessible via the web portal at https://www.proteindiffraction.org, can be searched using various criteria. All data are available for unrestricted access and download. The resource serves as a proof of concept and demonstrates the feasibility of archiving raw diffraction data and associated metadata from X-ray crystallographic studies of biological macromolecules. The goal is to expand this resource to include data sets that have failed to yield X-ray structures in order to facilitate collaborative efforts that will improve protein structure-determination methods and to ensure the availability of 'orphan' data left behind for various reasons by individual investigators and/or extinct structural genomics projects. Every dataset in the IRRMC resource is

assigned a DOI (Digital Object Identifier), which should provide a reliable mechanism of data location, even if the URL or the maintainer of the data changes.

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.

Research data mangement: administration, raw diffraction data, structure factors and coordinates at the UK's National Crystallographic Service (NCS)

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

The need to manage, curate and disseminate data has become paramount in the modern era of academic research. The data explosion that has occurred at the same time has prompted an increased requirement for transparency about its generation and a greater responsibility and accountability for facilities to provide accurate and long term mechanisms for archival and curation.

The NCS has led the way for chemical crystallography for around 15 years in developing approaches to addressing this problem [1, 2]. The eCrystals project (http://ecrystals.chem.soton.ac.uk/) developed an institutional repository approach to curating and disseminating coordinates, structure factors and a range of other information relating to the 'derived' data from a crystallographic experiment. However, raw diffraction data, although being rigorously archived and in the last 15 years highly curated, is only available on request directly to the NCS. eCrystals has been designed to act as a discipline specific data repository, which has resulted in a pragmatic metadata scheme for the description of its contents and this promotes discovery and reuse of the material it makes available.

There is also a necessary administrative function in running a facility that provides a service and this is intrinsically related to the data itself. Over 30 years of operation the NCS has accumulated a range of databases, spreadsheets and forms to meet ever-changing requirements for administration, tracking and reporting. For the last 15 years a range of NCS projects has been researching and addressing this problem. However, becoming an EPSRC mid-range facility in 2010 prompted a review of requirements and with additional demands for enhanced user interaction and reporting, alongside the Web becoming a more prevalent and mature technology that people readily engage with, 'Portal' was conceived. Portal aimed to bring together all the elements described above into a single unified and coherent system.

We have learnt a lot from this work and Portal has largely achieved its goals, however there are significant aspects of the data repository yet to be incorporated and also the need to maintain a modern codebase. We have therefore embarked on an 18-month project to address these matters. 'Portal - The Next Generation' will be a combination of a laboratory information system and a data repository with specific functions and plug-ins tailored for the operation of a crystallographic facility and its resulting data. The design objectives of this system, development progress and the potential for its availability to the community will be discussed.

[1] S. J. Coles *et al.* (2005), *J. Appl. Cryst.* **38**, 819–826.

[2] M. B. Hursthouse & S. J. Coles (2014), *Crystallogr. Rev.* **20**:2, 117–154.

Simon Coles *is Professor of Structural Chemistry and Director, UK National Crystallography Service within Chemistry at the University of Southampton. His 1997 PhD in structural systematics and molecular modelling was supervised by Professor Mike Hursthouse, with whom he moved to Southampton to establish a new laboratory and develop the National Crystallography Service (NCS). Simon took over the role of NCS Director in 2009. He has diverse research interests. Structural Chemistry research interests include the study of solid-state reactions and transformations, structural systematics, the determination of charge density distributions and their application to reactivity and*

solid-state behaviour, discovering and investigating structure-property relationships and crystal growth. Further structural science research involves collaborations with the areas of macromolecular crystallography, second-harmonic-generation laser spectroscopy, CT Imaging and 3D Inkjet Printing. Over the last decade Simon has been awarded a number of grants in the areas of Information Management, eResearch and eLearning.

SBGrid Databank

Peter Meyer

BCMP, Harvard Medical School and SBGrid Consortium, USA
Email: meyer@hkl.hms.harvard.edu

Peter A. Meyer[1], Stephanie Socias[1], Jason Key[1], Mercè Crosas[2] and Piotr Sliz[1,3]
[1]BCMP, Harvard Medical School and SBGrid Consortium, USA
[2]IQSS, Harvard University and Dataverse Project, USA
[3]Boston Children's Hospital, Dept of Pediatrics, USA

Access to experimental X-ray diffraction image data is fundamental for validation and reproduction of macromolecular models, and indispensable for development of improved data processing algorithms. We have established a diffraction data publication and dissemination system, the SBGrid Databank, to preserve diffraction datasets supporting published crystal structures. Published datasets are openly available through direct download and through Data Access Alliance (DAA) sites. Data deposition is open to all structural biologists, and datasets for unpublished structures can be held for later publication. Existing databases (such as the PDB and PubMed) are used to reduce the amount of additional information depositors need to provide. Reprocessing of published datasets is used to provide a baseline for ensuring that the datasets will be useful to other researchers. A set of REST APIs supports reprocessing pipelines, and allows users to access information about published datasets programmatically.

Pete Meyer *is the chief curator for the SBGrid Data Bank. He moderates data uptake, develops automated data validation tools, and works on implementation of the Data Locality module that will propagate a subset of SBDB datasets to various supercomputing sites. Dr Meyer joined the SBGrid team in 2014 after completing his postdoctoral training at the Medical College of Wisconsin. He holds a PhD degree from Cornell University. During his graduate and postdoctoral work, he studied a variety of RNA Polymerase II Transcription Factor Complexes, using and improving methods for studying large complexes with low-resolution X-ray crystallography. He is also an experienced computing programmer and has 10 years of Linux administration experience.*

Session III: High-data-rate/high-performance-computing issues of research data management in macromolecular crystallography

Dealing with the avalanche of data generated in high-data-rate macromolecular crystallography

Jean Jakoncic

Brookhaven National Laboratory, Upton, NY, USA
Email: jjakoncic@bnl.gov

Jean Jakoncic[1], Herbert J. Bernstein[2], Alexei Soares[1], Wuxian Shi[3], Martin Fuchs[1], Robert Petkus[1], Robert M. Sweet[1] and Sean McSweeney[1]
[1]Brookhaven National Laboratory, Upton, NY, USA
[2]Rochester Institute of Technology, Rochester, NY, USA
[3]Case Western Reserve University, Cleveland, OH, USA

Newly commissioned state of the art MX beamlines fitted with current advanced hybrid pixel detectors are now in operation. At the NSLS-II, AMX and FMX, two high-brightness microfocusing beamlines ($> 10^{11}$ and $> 5x10^{12}$ph/s/μm^2 respectively) are fitted with Dectris Eiger detectors and are equipped with advanced automation that will ultimately allow screening of up to 1000 crystals per day. We have seen throughput greater than 1 GB/s per beamline during demanding experiments and are expecting this to increase in the upcoming months. With this level of throughput, near real time data analysis feedback is a necessity. This requires infrastructure with a high bandwidth network, fast-I/O large storage and significant computational capacity. Optimized data processing software and pipelines are being developed to help in coping with the throughput. We will present the state of current problems that the community is facing and some of the solutions that are currently deployed at various facilities.

Jean Jakoncic *is a Scientist at the National Synchrotron Light Source II at Brookhaven National Laboratory.*

Intel Scalable System Framework

Henry Gabb

Sr. Principal Engineer, Intel Corporation, 1300 S. Mopac Expwy, Austin, TX 78746, USA
Email: henry.a.gabb@intel.com

The world depends on high-performance computing (HPC) to solve ever larger scientific, industrial, and societal problems, but we face growing technical and architectural challenges as HPC systems get larger. In traditional HPC, computation, memory/storage, and network performance are becoming more unbalanced so an integrated, holistic approach is needed for future systems. Also, different workloads (*e.g.* modeling and simulation, scientific visualization, big data analytics, machine learning) stress different parts of the system (compute, memory, I/O). This can lead to divergent, specialized system infrastructures that are dedicated to a particular type of workload. Specialized systems are more expensive to design, build, and manage because they do not benefit from economies of scale. They often require proprietary solutions that can limit software reusability. The solution to this problem requires innovative technologies that are tightly integrated. Intel Scalable System Framework (SSF) provides breakthrough compute, memory/storage, and network performance; a common infrastructure that supports a variety of workloads; standards-based programmability; and broad vendor availability. This is made possible by Intel's broad portfolio of innovative compute, memory and storage, network fabric, and software technologies, which allows unprecedented co-design and system integration. Tighter component integration improves compute density, I/O bandwidth, and network latency while lowering power consumption and overall cost. Intel SSF creates a stable system target for software vendors to help reduce development and maintenance costs. HPC users benefit from a common infrastructure. Reference designs based on Intel SSF help lower entry barriers for equipment manufacturers while still allowing them to innovate. The technical details of each of these high-level Intel SSF features will be discussed.

Henry Gabb *is a senior principal engineer in the Developer Products Division of the Intel Software and Services Group. He first joined Intel in 2000 to help drive parallel computing in various application domains. Prior to joining Intel, Henry did basic research in computational structural biology, mainly studying large-scale conformational transitions in biomolecules and molecular docking. Henry holds a BS in biochemistry from Louisiana State University, an MS in medical informatics from the Northwestern Feinberg School of Medicine, and a PhD in molecular genetics from the University of Alabama at Birmingham School of Medicine. He has published extensively in computational life science and high-performance computing. Henry recently rejoined Intel after spending four years working on a second PhD in information science at the University of Illinois at Urbana-Champaign, where he applied informatics and machine learning to problems in healthcare and environmental chemical exposure.*

Intel software and programming tools ecosystem for HPC

Henry Gabb

Sr. Principal Engineer, Intel Corporation, 1300 S. Mopac Expwy, Austin, TX 78746, USA
Email: henry.a.gabb@intel.com

High-performance computing (HPC) users are no strangers to code optimization and performance tuning. However, future HPC systems are likely to be even more heterogeneous than they are now. The mix of CPU, GPU, FPGA, and ASIC architectures could be quite diverse. Software will have to be modernized to take advantage of this heterogeneity, and Intel has an extensive ecosystem of programming tools to help. The Intel Fortran and C/C++ compilers have extensive auto-vectorization capability to deliver maximum performance on Intel processors. Support for the most popular productivity language is provided through the Intel Distribution for Python. Intel also provides a wide range of performance libraries, *e.g.* the Intel Math Kernel Library (FFT and numerical linear algebra), the Intel Integrated Performance Primitives (compression/decompression, image, vision, and signal processing), and the Intel Data Analytics Acceleration Library (big data analytics and machine learning). Many Python modules, machine learning frameworks, and third-party libraries already take advantage of the Intel performance libraries. Finally, Intel offers programming tools to support parallel debugging and tuning at the vector, thread, and process level. The key features of each tool will be discussed.

About our sponsors

We acknowledge the generosity of the corporate sponsors who have made possible this Workshop.

In addition to the companies and organizations listed below, we are grateful to the International Union of Crystallography, which provided the base funding for the Workshop.

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: Structural Biology*; *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 in 2014, *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.

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.

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, online access is provided 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.

IUCrJ
ISSN 2052-2525
NEUTRON | SYNCHROTRON

Raw diffraction data preservation and reuse: overview, update on practicalities and metadata requirements

Loes M. J. Kroon-Batenburg,[a] John R. Helliwell,[b]* Brian McMahon[c] and Thomas C. Terwilliger[d]

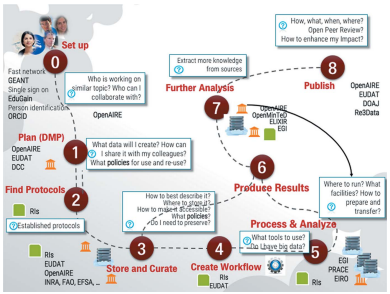
[a]Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, Utrecht, CH 3584, The Netherlands, [b]School of Chemistry, Faculty of Engineering and Physical Sciences, University of Manchester, Brunswick Street, Manchester M13 9PL, UK, [c]International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK, and [d]Bioscience Division, Los Alamos National Laboratory, Mail Stop M888, Los Alamos, NM 87507, USA. *Correspondence e-mail: john.helliwell@manchester.ac.uk

Received 11 August 2016
Accepted 15 November 2016

Edited by M. Takata, SPring-8, Japan

Keywords: raw diffraction data; data archiving; metadata descriptors for raw data; diversity of crystallographic instrumentation.

A topical review is presented of the rapidly developing interest in and storage options for the preservation and reuse of raw data within the scientific domain of the IUCr and its Commissions, each of which operates within a great diversity of instrumentation. A résumé is included of the case for raw diffraction data deposition. An overall context is set by highlighting the initiatives of science policy makers towards an 'Open Science' model within which crystallographers will increasingly work in the future; this will bring new funding opportunities but also new codes of procedure within open science frameworks. Skills education and training for crystallographers will need to be expanded. Overall, there are now the means and the organization for the preservation of raw crystallographic diffraction data *via* different types of archive, such as at universities, discipline-specific repositories (Integrated Resource for Reproducibility in Macromolecular Crystallography, Structural Biology Data Grid), general public data repositories (Zenodo, ResearchGate) and centralized neutron and X-ray facilities. Formulation of improved metadata descriptors for the raw data types of each of the IUCr Commissions is in progress; some detailed examples are provided. A number of specific case studies are presented, including an example research thread that provides complete open access to raw data.



OPEN ACCESS

1. Introduction and overview

1.1. Context

Recent years have seen a growth in interest in retaining raw diffraction data sets collected for the determination of crystal and molecular structures. This interest has arisen spontaneously within the crystallographic community on a number of fronts. For example, raw data sets are valuable for developing new methods of structure determination and for benchmarking of software algorithms (Terwilliger & Bricogne, 2014); they are sometimes important for validating the interpretation of structural features; and increasingly they repay closer study, whether for allowing data analysis at higher resolution than used in the original work, understanding the presence of multiple lattices present in a crystal, or deducing details of correlated motions or disorder from the diffuse scattering that is largely ignored in determining Bragg peak positions and characteristics.

In parallel, the evolution of science policy in the wider world is prompting closer scrutiny of the whole practice of research data management, and there are a growing number of mandates to retain the raw data underpinning any experimental study and to make it available to other researchers. By

early 2016, all UK scientific research councils had stated positions on data management, access and long-term curation (Digital Curation Centre, 2016; Research Councils UK, 2015). A useful summary of US Federal Funding Agency requirements for scientific data management is hosted by Northwestern University Library (2016). A noteworthy recent proposal calls for a European Open Science Cloud for Research (Jones, 2015).

Different communities have different ideas of what data they value most – and, indeed, of what constitutes 'data'. The USA's National Science Foundation (NSF) makes this explicit in its published 'Frequently Asked Questions' (National Science Foundation, 2010):

1. What constitutes 'data' covered by a Data Management Plan?

What constitutes such data will be determined by the community of interest through the process of peer review and program management. This may include, but is not limited to: data, publications, samples, physical collections, software and models.

In consequence, there is great variety amongst different scientific disciplines in their approaches to data management and retention, and therefore in the availability of public repositories and in the software tools to manage deposition, access and reuse. Nevertheless, two themes recur in the various published mandates and best-practice guidelines: the importance of persistent identifiers for data sets, and the vital need to characterize them as fully as possible by appropriate metadata.

Crystallography is generally regarded as a science that has its house in good order regarding data management, validation, access and reuse. This is largely true so far as 'derived' data (by which we mean atomic positional coordinates and displacement parameters resulting from structure determinations) and associated publications are concerned. It is more debatable where processed diffraction data are concerned – the post-experiment processed data (typically structure factors) that form the basis of the atomic and molecular structure determination and subsequent refinement leading to a structural model. Some journals require deposition of structure factors in support of any publication, and the Protein Data Bank (PDB; Berman *et al.*, 2000) requires structure factors to be deposited along with the atomic coordinates. However, these are usually the final set of structure factors used in refinement, and may lack information discarded when merging symmetry-related diffraction peaks, or excluded for other reasons from early cycles of refinement. The PDB *will* accept unmerged processed intensity data, and there are community recommendations encouraging their deposition (International Structural Genomics Organization, 2001), but the practice is not yet universal in macromolecular crystallography. For small-unit-cell crystal structures, even journals that accept structure factors have not hitherto required unmerged intensities. However, there is growing recognition that they are important, both for further development of the *checkCIF* validation carried out during the peer review process, and indeed to encourage future researchers to revisit and re-evaluate the published results, perhaps when new ideas or tools become available (A. Linden, personal communication).

However, historically there has not been a tradition of retaining the raw X-ray diffraction images collected by electronic detectors, although centralized neutron facilities have long-standing traditions of raw data preservation. In recent years the practices nurtured by the neutron facilities have been spreading; each type of large-scale centralized instrumental facility (synchrotrons and latterly free-electron lasers, as well as neutron reactors) has begun to move towards raw data preservation. This trend has been encouraged by rapidly improving electronic data-handling procedures.

In 2011, the International Union of Crystallography (IUCr) established a Working Group to explore the merits and challenges of retaining the initial experimental data. This group, the Diffraction Data Deposition Working Group (DDDWG), has conducted a number of consultations, discussion meetings and workshops to explore the topic. A set of papers published in *Acta Crystallographica Section D* (Terwilliger, 2014) provided an overview of the reasons for archiving raw data in the field of macromolecular crystallography, models for doing so on a routine or large-scale basis, current practical initiatives, and the potential benefits for improving macromolecular structure models.

These papers also highlighted the importance of assigning persistent identifiers to data sets to facilitate their management and long-term curation, and to ensure that each data set was characterized by rich metadata, both to facilitate discovery and to allow effective scientific reuse (Guss & McMahon, 2014; Kroon-Batenburg & Helliwell, 2014).

In the remainder of this *Introduction*, we introduce a recent workshop that concentrated on metadata in crystallographic and related experiments; we review the arguments for depositing raw data as a routine practice; and we place these activities in the context of global science policy initiatives. The paper then looks in more detail at the current and evolving mechanisms for the deposition of raw experimental data (especially X-ray diffraction images); at detailed requirements for metadata that describe archived data sets, in order to ensure the reproducibility of the derived scientific results; and at the next steps forward.

1.2. Improving the metadata

To focus on the metadata issues, the DDDWG conducted a two-day workshop at Rovinj, Croatia, in August 2015. A complete record of the workshop is maintained online at http://www.iucr.org/resources/data/dddwg/rovinj-workshop and a number of articles arising from the meeting are in preparation. We detail here some specific outcomes from the workshop.

1.2.1. Efforts of the IUCr Commissions. The IUCr manages its scientific mission through a number of Commissions, each responsible for a particular topic area within crystallography. The DDDWG has requested each Commission to consider its own needs for defining metadata for raw experimental data

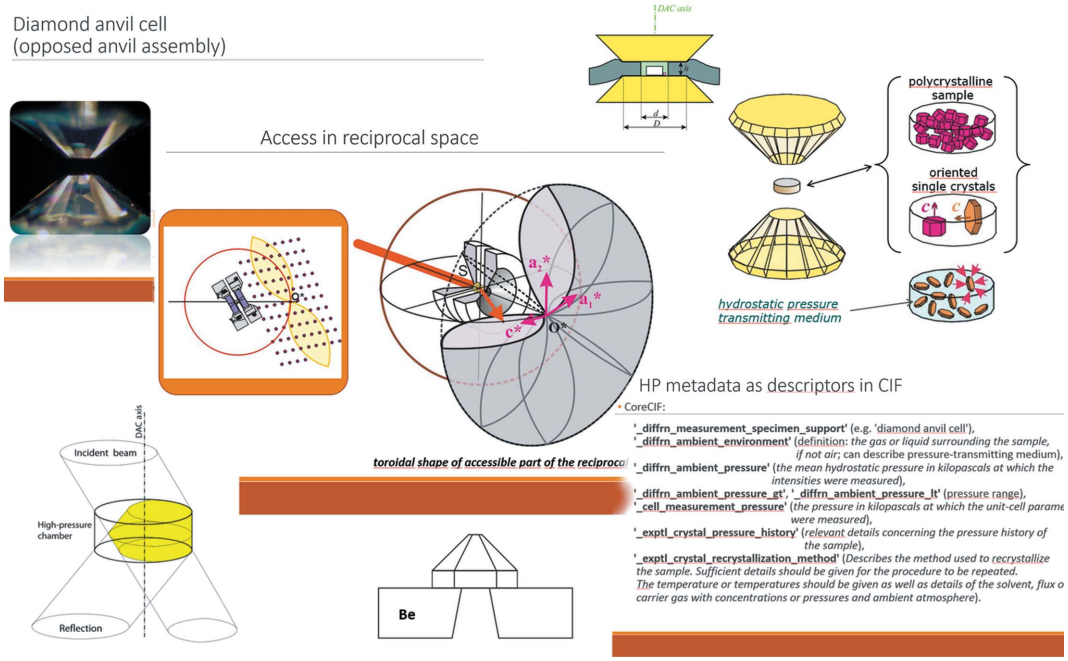


Figure 1
Montage of slides from Kamil Dziubek's presentation at the Rovinj workshop, illustrating aspects of diffraction experiments under high pressure and other non-ambient conditions that need to be well characterized and recorded. (Graphics courtesy of Ronald Miletich-Pawliczek, University of Vienna.)

within its field. Among those that have been most active in responding to this request are the Commission on XAFS (Ravel *et al.*, 2012); the Commission on Small-Angle Scattering (Jacques *et al.*, 2012); the Commission on High Pressure (Fig. 1); and the Commission on Biological Macromolecules (*e.g.* Gutmanas *et al.*, 2013).

The International Centre for Diffraction Data (ICDD, Pennsylvania, USA; http://www.icdd.com) has been active in the harnessing of raw powder diffraction data sets for some time and reported to us at ECM29 in Rovinj (August 2015) that they have now incorporated over 10 000 raw powder diffraction data sets into the Powder Diffraction File. They note that one-dimensional data sets are generally reasonably well characterized in terms of the experimental metadata catalogued in the powder CIF (pdCIF) dictionary (Toby, 2005), but that interpretation of two-dimensional diffraction images is hampered by a lack of consistency in reporting such characteristics as goniometer axes, detector dark current, distortion and other corrections (T. Fawcett, personal communication; see also Section 1.2.2). The Commission on Powder Diffraction is planning further work on neutron powder diffraction raw data and will liaise with the Commission on Neutron Scattering as appropriate. The Commission on Structural Chemistry has had enthusiastic participants in events convened by the DDDWG in Madrid, Bergen and Rovinj.

1.2.2. Characterizing X-ray diffraction images. The class of experimental data sets that most closely fits the original remit of the DDDWG is X-ray diffraction images collected from CCD or pixel detectors. A good catalogue of the metadata needed, in general, to interpret a raw image data file was given by Kroon-Batenburg & Helliwell (2014). Many of the individual items required are defined in the imgCIF dictionary (Bernstein, 2005), and there have been partial implementations of some of them in so-called 'mini-CBF' headers of image files written by a number of commercial detector systems. However, this has not been done in a consistent way between vendors nor even across the entire product range of individual vendors. (CBF, the crystallographic binary file, and imgCIF, its pure ASCII counterpart, are equivalent implementations of the CIF ontology for diffraction images.)

Increasingly, images are being stored using the HDF5/NeXus data format (Könnecke *et al.*, 2015), and although the physical format of the data file should not affect its ability to store specific structured information (Hester, 2016), some effort will be needed to ensure that the CIF and NeXus data representations are equally capable of storing the appropriate experimental metadata. Significant effort to achieve this at the technical level has already been invested following participation in an earlier workshop by representatives of COMCIFS (Committee for the Maintenance of the CIF Standard) and NIAC (NeXus International Advisory Committee), the bodies responsible for managing the CIF and NeXus data formats, respectively (Bernstein *et al.*, 2013). Nevertheless, presentations at the Rovinj Workshop by Kroon-Batenburg (https://youtu.be/XXFDlNn21SY) and by Minor (https://youtu.be/eQbs9sB_pOM) emphasized that there is still a long way to go before the myriad different formats generated by commercial electronic position-sensitive detectors do contain the necessary common metadata to allow for easy interpretation and management (see further discussion in Section 3.2).

The arrival of the new Dectris Eiger pixel detector, with its colossal increase in diffraction image data rates, has high-

lighted the importance of efficient data format and metadata recording, not only for diffraction data processing on a synchrotron or X-ray laser beamline, but also for subsequent processing outside the facility, and ultimately for reprocessing/reanalysis from a raw data archive as may be needed. The various issues have been highlighted in detail in a discussion thread on the CCP4bb mailing list in early March 2016 (involving, amongst others, G. Winter, A. Förster, H. J. Bernstein, C. Vonrhein and G. Bricogne).

1.3. The case for raw data deposition

We summarize the case for routine storage and retrieval of raw data to emphasize its potential value to the community. At the same time we acknowledge the cost and other practical constraints of storing all collected data sets indefinitely, and we are unable to give a definitive indication of where the balance might lie between archiving and discarding raw data. However, we show in Section 1.4 that there are discernible trends towards storing more data sets than we might have expected in the early work of the DDDWG.

There is a broad philosophical view of the importance of access to raw diffraction data, namely that science requires the ability to conduct a comprehensive analysis through one's own eyes and not the lens of someone else. Raw diffraction images offer several opportunities for improved or novel science. They permit the analysis of data at higher resolution than used in the original work [allowing comparisons not only among data processing software (Tanley *et al.*, 2013), but also in the effectiveness of structure determination and refinement with ever weaker data beyond normal limits]. Raw data sets can serve as benchmarks in developing improved methods of analysis. They allow checking of the interpretation of the symmetries of the crystals, and detailed analysis of diffraction from multiple lattices present in the crystals. More generally, they promote the study of the diffuse scattering that reflects correlated motions or disorder of atoms in the crystals, namely the 'structural dynamics'.

The retention of raw data can be seen as complementing the extensive archives of derived data (*i.e.* cell parameters, molecular coordinates, anisotropic displacement parameters) and processed data (structure factors, Rietveld refinement profiles) in the crystallographic databases. The contributions of the former are very well understood: they form part of the scientific record, they lead to database-driven discovery, *e.g.* in understanding protein–ligand interactions, they lead to new pathways to synthesis, improvements in manufacturing and better understanding of energetics, and they have use in identification and indexing applications (*e.g.* in forensic science).

Until the advent of CIF and the automated structure validation checks with the *checkCIF* suite (Strickland *et al.*, 2005) that it enabled, many structures were published which required subsequent correction. Often, the interpretation of the results produced molecular structures that were broadly correct, but overlooked higher lattice symmetries. Such examples were best detected and corrected through access to the deposited structure factors (well illustrated by Marsh *et al.*, 2002).

So, broadly speaking, structure validation (the credibility of a structural model, both in its adherence to norms of geometric configuration and its derivation from X-ray diffraction images) can be carried out with reference to the derived data sets (the structural coordinates) and the structure factors alone, and this has been the practice in various crystallography journals for a considerable length of time. However, the availability of the raw data (*i.e.* original diffraction images) can enhance structure validation in the following ways:

(i) The structure can be re-refined, perhaps making use of diffraction peaks that were excluded because the processed diffraction data were truncated at an arbitrary resolution limit. Retention of the original data also permits re-evaluation of the space-group symmetry, which is normally settled upon during an early stage of conventional refinement.

(ii) Data reduction is often performed according to established protocols, but retention of the original images allows the opportunity to test those protocols, especially if there is any suspicion of systematic bias. Indeed, statistical analysis of a collection of stored raw images may allow the detection of systematic biases that are not at all apparent in individual experiments. Further, the availability of large collections of raw data sets allows periodic recalibration of solution methods and the development of new methods to tackle data sets that have previously been resistant to conventional solution.

(iii) Attention to diffuse scattering between the diffraction spots allows insight into correlated motions or disorder of atoms in crystals. This might involve quasicrystalline behaviour, determination of incommensurate modulation or multi-phase representation, macromolecular motions or conformational changes *etc.*

Note that these benefits may not be apparent for every structure, and the cost–benefit calculus informing policies of routine deposition has still to be determined by the community and funding bodies (Guss & McMahon, 2014). It may be that there are different entry points where the potential benefits can be most readily realised, *e.g.* by making available the experimental data for 'difficult structures' that have proved impossible to refine satisfactorily.

However, more-or-less routine deposition of primary data would help to improve the quality and reliability of the scientific record (Minor *et al.*, 2016). It would allow closer scrutiny of scientific deductions by peer reviewers prior to publication; it would allow for revisiting and revising structural models already in the databases, as new techniques are developed – *e.g.* the notion of 'continuous improvement of macromolecular structure models' (Terwilliger, 2012); it allows reanalysis of a structure or series of structures independent of an author's interpretational bias (B. D. Bax, personal communication); and it provides the experimental evidence needed to support any claims made by the publishing author. In this last role, it helps to guard against the use of the wrong data set, either through error or deliberate intention.

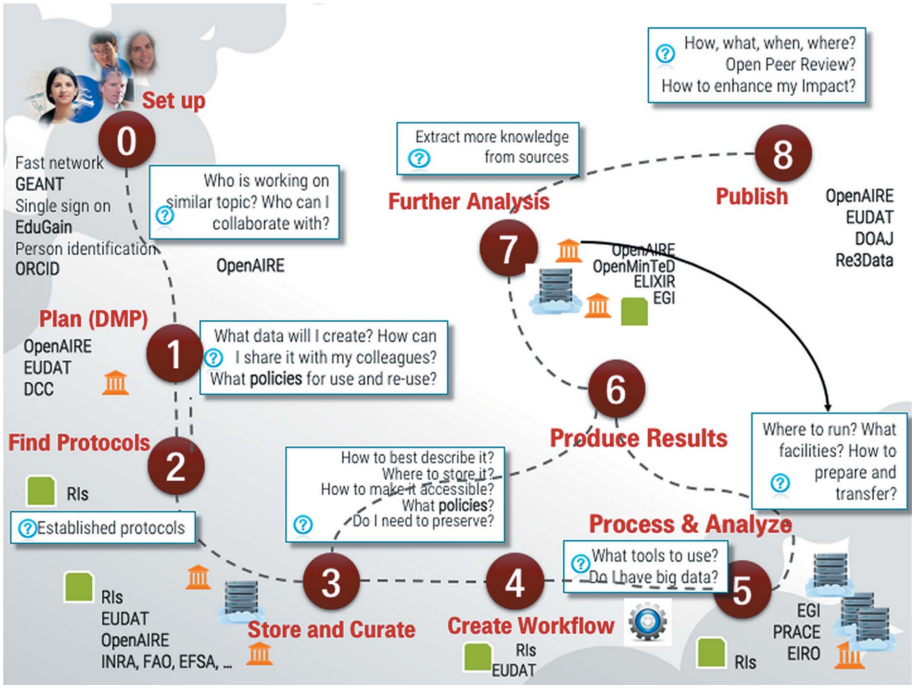


Figure 2
A graphic linking data publishing and management workflow to EU research infrastructural components. Part of a presentation introducing the European Open Science Cloud for Research (illustration courtesy of Natalia Manova for the European OpenAIRE project).

1.4. Deposition imperatives and opportunities

As previously mentioned, there have been developments since the DDDWG was established in the climate for data deposition and sharing, both in the wider scientific world and in the field of crystallography and related structural sciences. The benefits of open data (*i.e.* collecting research data arising from publicly funded scientific research and making it available for reuse without charge to the end user) have been reiterated in recent years in international, governmental and scientific policy discussions and practical initiatives. Among a few portal web sites of note are the United Nations data portal (UNdata: http://data.un.org), the US Government open data site (https://www.data.gov) and the federated 'Global Science Gateway' http://worldwidescience.org. Calls for implementation include 'The Good Growth Plan', a collaboration for agricultural development involving the UK Open Data Institute (ODI; https://theodi.org) and Syngenta; the European Open Science Cloud (EOSC), a European Union strategy for linking research networks, data storage facilities and computing resources across the continent (Jones, 2015; Fig. 2); and an Open Data Accord (Science International, 2015) launched by the International Council for Science (ICSU), the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC).

Although these various initiatives are very diverse in their objectives, collectively they are raising the perceived importance of data repositories to research funders, to researchers who are encouraged or in some cases mandated to deposit their data in robust and durable repositories, and to other researchers who are becoming increasingly aware of the availability of other data sets and their potential usefulness to their own work. A gradual change in cultural attitudes to research data is taking place.

Since the DDDWG was established in 2011, there have been a number of developments, some catalysed by these high-level initiatives, that have increased the options for deposition of diffraction images:

(i) The number and scope of university data repositories has expanded.

(ii) The European Synchrotron Radiation Facility (ESRF; Grenoble, France) has launched a Data Archive, in which every raw data set measured can be associated with a registered DOI.

(iii) The Zenodo science data archive, hosted on the extremely high capacity CERN storage system, has gathered momentum.

(iv) A repository for diffraction experiments used to determine protein structures has been established as part of the US National Institute of Health's BD2K (Big Data to Knowledge) programme (Grabowski *et al.*, 2016); it is run by Wladek Minor's group at the University of Virginia, USA (http://www.proteindiffraction.org/).

(v) The Structural Biology Data Grid (SBDG) has been established as a diffraction data publication and dissemination system for structural biology (Meyer *et al.*, 2016).

(vi) The Protein Data Bank (PDB) now requests the DOI (digital object identifier) for raw data and metadata for raw data during a deposition (Fig. 3).

(vii) *IUCrData* (an IUCr data service, initially handling derived data sets) has been launched.

Some of these are described in more detail in Section 2.2.

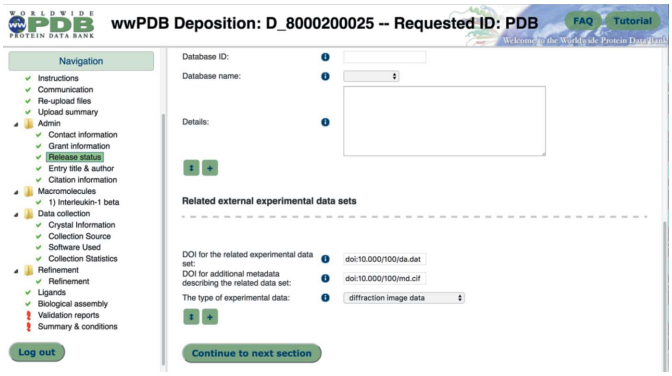


Figure 3
Online form allowing PDB depositors to link experimental data sets and their associated metadata with a deposited macromolecular structure.

2. Mechanisms for raw diffraction data preservation

We review some of the *de facto* repositories that are currently hosting, and in many cases providing access to, experimental data sets in our domain.

2.1. Institutional data repositories. Case study: University of Manchester

The meticulous approach of the University of Manchester makes one of us (JRH) feel very fortunate to be working in this research environment. In researching the binding of the anti-cancer agent cisplatin to histidine [which has received intense interest; see, for example, Messori & Merlino (2016)], JRH's research group made the raw diffraction data open access at the University of Manchester institutional data repository. Fig. 4 illustrates the data access record within the Library system, while Fig. 5 illustrates the classification-level metadata required by such a repository. This type of institutional cataloguing and archive is increasingly characteristic of modern data archive initiatives. In addition, we have followed the standard community data deposition requirements of depositing coordinates and processed diffraction data at the Protein Data Bank. To permit the widest possible access to our work, we have also been able, *via* the EPSRC funding we have had, to publish the bulk of our articles reporting our results as 'gold' open access (*i.e.* the full peer-reviewed articles of record can be accessed without a journal subscription) in *Acta Crystallographica Sections D* and *F*.

In becoming pioneers of making both our raw diffraction data and our data and model interpretations fully open (Table 1), thus achieving a rare breadth and depth of openness within a focused research theme, our research has received a gratifying amount of detailed interest. There have been many downloads of these raw data, both from their original web location at Utrecht University and subsequently from the University of Manchester. The download totals for each year from Utrecht were: 2012 17 GB, 2013 47 GB, 2014 57.69 GB and 2015 31.47 GB; equivalent download information is not available from the University of Manchester. One such raw data download featured in a new publication (Shabalin *et al.*, 2015), a wide-ranging critique of the whole field of cisplatin binding to various proteins. This article suggested improvements to three of our cisplatin–lysozyme models in the PDB *via* three of their own alternative interpretations; two of these involved use of our processed diffraction data held at the PDB (4xan and 4mwk) and one of our raw data (4g4a in Table 1 and Fig. 4). We have accepted some of their recommendations and rejected others (Tanley *et al.*, 2016). Some of these points of 'data debate' also suggest a lack of mature community standards, even within one journal (Tanley *et al.*, 2015), but they also show a way forward for discussions to be conducted, *e.g.* within IUCr journals. In other aspects, it shows the benefits of the continuing pursuit of improved methods of analysis and a better understanding of the role of weak data in improving protein model refinements (Diederichs & Karplus, 2013), which we harnessed in detail in Tanley *et al.* (2016). Such improvements have arisen even in just the last few years, and illustrate the 'young age' of macromolecular crystallography, a field that is still clearly maturing as a technique.

2.2. General data repositories for structural biology

The importance of data capture and archiving has been widely recognized around the world and several repositories are now available where nearly any researcher can, or will

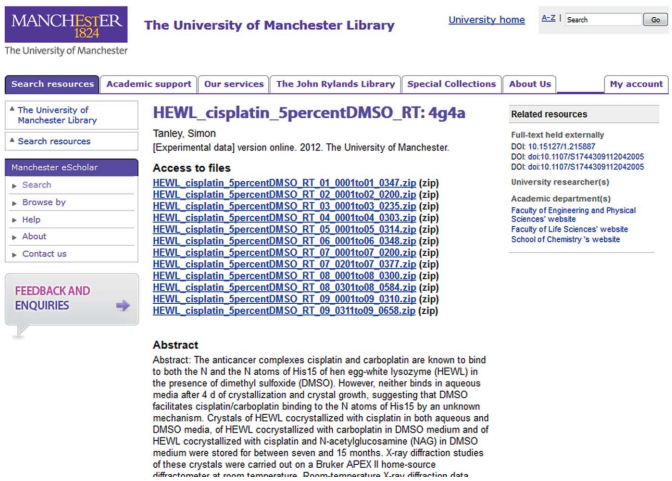


Figure 4
Manchester University Library access record for experimental data sets associated with a published research article. Links are provided to the published article in the 'Related resources' column.

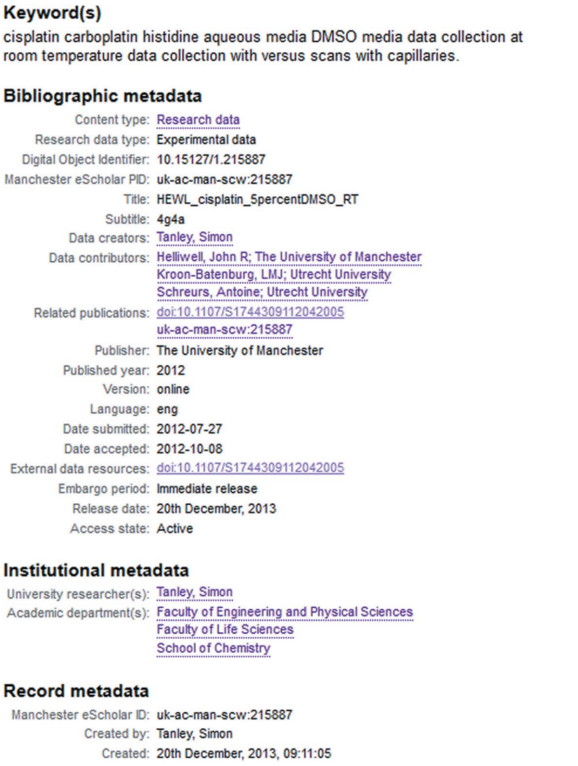


Figure 5
Classification-level metadata associated with experimental data sets archived at the University of Manchester Data Library. These identify the archived data sets and provide links to related resources.

Table 1
A thematic raw data collection as an example: the suite of research studies, relating to platins binding to histidine, held at the University of Manchester Data Library.

Entry No.	Raw diffraction data DOI	PDB code	Article DOI
1	10.15127/1.215887	4g4a (now 5hll)	10.1107/S1744309112042005 and 10.1107/S2053230X16000856
2	10.15127/1.219240	4dd2	10.1107/S0021889812044172 and 10.1107/S0907444912006907
3	10.15127/1.219241	4dd3	10.1107/S0021889812044172 and 10.1107/S0907444912006907
4	10.15127/1.219257	4dd9	10.1107/S0021889812044172 and 10.1107/S0907444912006907
5	10.15127/1.219267	4g4h	10.1107/S1744309112042005
6	10.15127/1.219263	4g4c	10.1107/S1744309112042005
7	10.15127/1.219242	4dd7	10.1107/S0021889812044172 and 10.1107/S0907444912006907
8	10.15127/1.219318	4gcb	10.1107/S090744491204423X
9	10.15127/1.219319	4gcc	10.1107/S090744491204423X
10	10.15127/1.219320	4gcd	10.1107/S090744491204423X
11	10.15127/1.219321	4gce	10.1107/S090744491204423X
12	10.15127/1.219322	4gcf	10.1107/S090744491204423X
13	10.15127/1.219260	4ddc	10.1107/S0021889812044172 and 10.1107/S0907444912006907
14	10.15127/1.219238	4ddb	10.1107/S0021889812044172 and 10.1107/S0907444912006907
15	10.15127/1.219230	4dd0	10.1107/S0021889812044172 and 10.1107/S0907444912006907
16	10.15127/1.219233	4dd4 (now 5l3h)	10.1107/S0021889812044172, 10.1107/S0907444912006907 and arXiv:1606.01372
17	10.15127/1.219236	4dd6 (now 5l3i)	10.1107/S0021889812044172, 10.1107/S0907444912006907 and arXiv:1606.01372
18	10.15127/1.219264	4g4b	10.1107/S1744309112042005
19	10.15127/1.219259	4dda	10.1107/S0021889812044172 and 10.1107/S0907444912006907
20	10.15127/1.219266	4g49	10.1107/S1744309112042005
21	10.15127/1.215883	4dd1	10.1107/S0021889812044172 and 10.1107/S0907444912006907
22	10.15127/1.266911	4nsj	10.1107/s2053230x14016161
23	10.15127/1.266910	4nsi	10.1107/s2053230x14016161
24	10.15127/1.266909	4nsh	10.1107/s2053230x14016161
25	10.15127/1.266908	4lt3	10.1107/s2053230x14016161
26	10.15127/1.266907	4lt0	10.1107/s2053230x14016161
27	10.15127/1.266906	4nsf (then 4xan now 5hmj)	10.1107/s2053230x14016161 and 10.1107/S2053230X16000777
28	10.15127/1.266905	4owb	10.1107/s2053230x14013995
29	10.15127/1.266904	4owa	10.1107/s2053230x14013995
30	10.15127/1.266903	4ow9	10.1107/s2053230x14013995
31	10.15127/1.266899	4mwk (now 5hmv)	10.1063/1.4883975 and 10.1063/1.4948613
32	10.15127/1.266900	4mwm (now 5hq1)	10.1063/1.4883975 and 10.1063/1.4948613
33	10.15127/1.266901	4mwn (now 5i5q)	10.1063/1.4883975 and 10.1063/1.4948613
34	10.15127/1.266902	4oxe (now 5idd)	10.1063/1.4883975 and 10.1063/1.4948613

soon be able to, deposit their raw data and associated metadata for anyone in the world to view and download, subject of course to the natural constraints of file size and network bandwidth.

Two major publicly funded repositories are the Integrated Resource for Reproducibility in Macromolecular Crystallography (http://www.proteindiffraction.org) and the Zenodo repository (https://zenodo.org) for general scientific data. The former has been developed by the Minor group at the University of Virginia (http://olenka.med.virginia.edu/CrystUVa) and is supported by the US National Institutes of Health Big Data to Knowledge Initiative (https://datascience.nih.gov/bd2k). Zenodo has been developed by CERN (http://www.cern.ch) as part of the European Union OpenAIREplus initiative (http://www.openaire.eu).

Two additional private repositories are available for general use. The Harvard-based SBGrid organization (https://sbgrid.org) has developed a Structural Biology Data Grid (https://data.sbgrid.org) that can be used by any member of SBGrid to archive raw data and metadata. The ResearchGate scientific networking site (https://www.researchgate.net) allows researchers to share data (https://www.researchgate.net/blog/post/present-all-your-research-in-a-click).

2.2.1. The Integrated Resource for Reproducibility in Macromolecular Crystallography. The Integrated Resource for Reproducibility in Macromolecular Crystallography (Grabowski *et al.*, 2016) is a protein diffraction database that addresses the need for archival of crystallographic raw images, as outlined in the discussion above and in the *Acta Cryst. D* group of articles published recently (Terwilliger, 2014). This database currently includes over 2900 raw crystallographic data sets and associated metadata. Most of these are linked with a deposit in the Protein Data Bank (http://www.pdb.org; Berman, 2000) and many of them represent work from structural genomics projects (http://csgid.org, http://ssgcid.org, http://www.jcsg.org, http://mcsg.anl.gov, http://thesgc.org). The database is highly structured, with crystallographic metadata associated with each data set. A very useful feature of this service is that the web interface to the database shows a representative diffraction image from each data set, allowing a researcher to note quickly the characteristics of the diffraction from the crystals used in each data set, for example the order in the diffraction pattern, the presence of diffuse scattering and the extent of anisotropy in the diffraction pattern. The database can be searched based on PDB ID, resolution of diffraction, the location where data were collected, authors, and many other characteristics. It is planned for the database to be available for deposits and downloads by anyone. Every entry in the database has an assigned DOI that can be used to refer to the data and which provides a stable permanent link to

the data, and the data deposited are not limited in file size. The metadata associated with the raw data are an integral part of the database, so that it may be practicable in the future to reprocess automatically much of the raw data in the database as new algorithms for data analysis become available (*cf.* Terwilliger & Bricogne, 2014).

2.2.2. Zenodo. The Zenodo archive is a general scientific archive developed by researchers at CERN as part of a European Union Framework 7 initiative. It provides a repository for scientific data sets in any field and has the unique feature that, as part of CERN, it has access to exceptional capacity for data storage and archiving. Though it is supported by the EU, researchers from anywhere in the world can archive their data and anyone can access the data. The Zenodo archive is designed to provide a resource for the many small scientific projects in the world that do not have an easy way to make their data available to the scientific community and, unlike the other databases discussed here, plans to charge a fee for larger-scale users. The archive currently has over 2500 data sets from all fields of science. Data sets can have multiple files, normally up to a total size limit of 50 GB; individual files can be a maximum of 2 GB in size. Each data set is assigned a DOI for permanent archiving and discovery, and is linked with metadata provided by the researcher.

2.2.3. Structural Biology Data Grid. The SBGrid organization provides access for researchers at many structural biology laboratories around the world to a packaged set of software that can be used in many areas of structural biology, including X-ray crystallography, cryo-electron microscopy, electron diffraction, small-angle scattering and other areas. SBGrid also provides access to cloud-based computing resources that carry out structural biology calculations. The Structural Biology Data Grid is a service recently started by SBGrid that allows any SBGrid researcher to archive raw data from any of the SBGrid structural biology areas. This database currently has over 240 data sets from 62 different institutions. The data can be viewed by anyone and crystallographic data sets can be downloaded by anyone, with cut-and-paste scripts for easy downloading of individual data sets. Each data entry has a unique DOI assigned, there are no limitations on file sizes, and metadata describing how to analyse the data are provided.

2.2.4. ResearchGate. ResearchGate is a commercial scientific social networking service that provides a simple mechanism for researchers to post their scientific papers and information about themselves, and for researchers to communicate about and discuss scientific topics. ResearchGate additionally allows researchers to archive scientific data sets for anyone to download. The data sets are assigned a DOI, and the size of individual files is limited.

2.3. Synchrotron, neutron and X-ray laser facility options

There are now several striking examples of current and evolving practice in data capture and management across a range of large-scale facilities accommodating a variety of techniques and sciences. Among those we are aware of are the Australian Synchrotron (Clayton, Victoria, Australia), the ESRF, the Institut Laue–Langevin (ILL, Grenoble, France), the Diamond Light Source (Didcot, UK) and the ISIS neutron source at the Rutherford Appleton Laboratory (Didcot, UK). The Australian Synchrotron has led the world's synchrotrons on data archival with its Store.Synchrotron data storage service for macromolecular crystallography (Meyer *et al.*, 2014). As well as diffraction image data archiving, it also supports users in their publications with linking to raw data sets *via* DOI registrations and, finally, the release of data sets for public analysis – something that, in the neutron community, the ILL is doing as well. There are also fine examples like Diamond that has so far retained all of its measured data. The ESRF has published a summary of its views on the era of Big Data at synchrotron radiation facilities in general and the challenges that today face the ESRF itself (ESRF, 2013). In an encouraging recent statement, it has announced a proactive data archiving policy (Andy Götz and colleagues from ESRF, personal communication).

There are still very significant challenges of data management in home laboratories and for medium-scale service providers such as the UK National Crystallography Service (Southampton, UK). In all these places, all the data from an experiment must be handled in the context of resource management, provenance, validation and bulk storage, all of which require ever greater volumes of metadata that should conform to widely accepted standards.

2.4. The data deluge

One caveat that we apply to our encouraging survey of repository solutions is that, as technology advances, so the volume of data collected is increasing at a dramatic rate. Hence, while the entire download total from Utrecht University in 2015 was 31 GB, a single data set produced by an Eiger 16M detector currently operating on a synchrotron beamline could be over 70 GB. This suggests that centralized experimental facilities, with their large data storage capacities and gigabit internal networks, will continue to play an important role as first-choice repositories for quasi-routine retention of data sets. However, it may also become necessary to apply principles of 'triage', either at the point of data collection or in subsequent long-term storage allocation. Such triage might either delete certain data sets or retain some subset, according to a variety of possible criteria. An initial suggestion for a set of such criteria was proposed in the DDDWG online forum in 2011 (http://forums.iucr.org/viewtopic.php?f=21&t=57) but has yet to be developed by the community.

3. Metadata for raw data requirements

3.1. A holistic metadata framework for crystallography

Crystallography and related structural sciences are fortunate in having a standardized approach to data characterization and management, known as the Crystallographic Information Framework (CIF; Hall & McMahon, 1995). This

has two components: a standard file format and data model (Hall *et al.*, 1991; Bernstein *et al.*, 2016), which facilitate data exchange between software programs, structural databases and publishing systems; and a set of 'dictionaries' that control the meaning of the tags associated with data values, and which can impose restrictions on data types and values where appropriate. These dictionaries collectively constitute the controlled vocabulary and associated definitions that represent the semantic meaning of a data file or stream – what is fashionably called the 'ontology' of a particular scientific domain.

Each CIF dictionary contains definitions relevant to a particular field or topic area, such as small-unit-cell structures determined by single-crystal diffractometry (the so-called 'core' dictionary), powder diffraction, biological macromolecular structures, modulated incommensurate structures, multipole electron density or diffraction images (Hall & McMahon, 2016). These compilations by topic take a comprehensive view of what may be termed 'data'. Thus, the core dictionary contains items as diverse as a single atomic positional coordinate, the ambient temperature at the time the experiment was conducted, the convergence metrics of the least-squares refinement, the software used for generating molecular graphics, or the entire text of an associated scientific publication. That is, there is no differentiation between items that might normally be categorized as 'raw', 'processed' or 'derived' data, or that might be characterized as 'metadata'.

The advantage of this lack of differentiation is that *all* the information needed to interpret, validate or reuse a data set can be stored in a single file; and this can make it easier to collect and verify such information during the course of an experimental workflow. Fig. 6 illustrates how the CIF ontologies inform the 'coherent information flow' at every stage of the information processing lifecycle in a typical structure determination experiment. In practice, not all real-world workflows use CIF as their actual mechanism for capturing data and metadata. For example, in large instrumental facilities, information about a particular experiment might be collected within a unified content management system developed by the facility to accommodate a wide range of different scientific experiments (Matthews *et al.*, 2010). Similarly, to manage the high-throughput data acquisition requirements of modern detectors, images may be generated as binary HDF5 files, or in proprietary formats.

Nevertheless, all raw data sets and associated metadata can, in principle, be converted into CIF representations, which might be a practical benefit for archiving purposes (*i.e.* to use a single standard representation), or at the very least can demonstrate what important metadata are missing, by comparison with the comprehensive CIF dictionary compendia of what can and should be collected.

Various IUCr Commissions are continuing to compile metadata definitions relevant to their field of interest in the form of CIF dictionaries. In addition to those listed by Hall & McMahon (2016), a small-angle scattering dictionary (sasCIF) has recently been published (Kachala *et al.*, 2016); work is well advanced by the IUCr Commission on Magnetic Structures to characterize magnetic structures and their underlying symmetries (magCIF); and the Commission on High Pressure has an active working group defining essential aspects of the experimental setup needed in non-ambient crystallography.

As mentioned before, the imgCIF dictionary describes an actual format for storing raw diffraction data. However, it also includes a rather complete set of data items that, if fully populated and used in conjunction with other items in the core or macromolecular CIF dictionaries, can fully describe the experimental apparatus and operating parameters, thus permitting a complete interpretation of archived images in this format. The imgCIF format itself is relatively little used, largely because of the speed requirements in modern detectors which require different data acquisition strategies. However, there is an ongoing effort to define metadata terms in the increasingly common NeXus format (Könnecke *et al.*, 2015) that are in concordance with the experimental metadata items defined in the imgCIF dictionary.

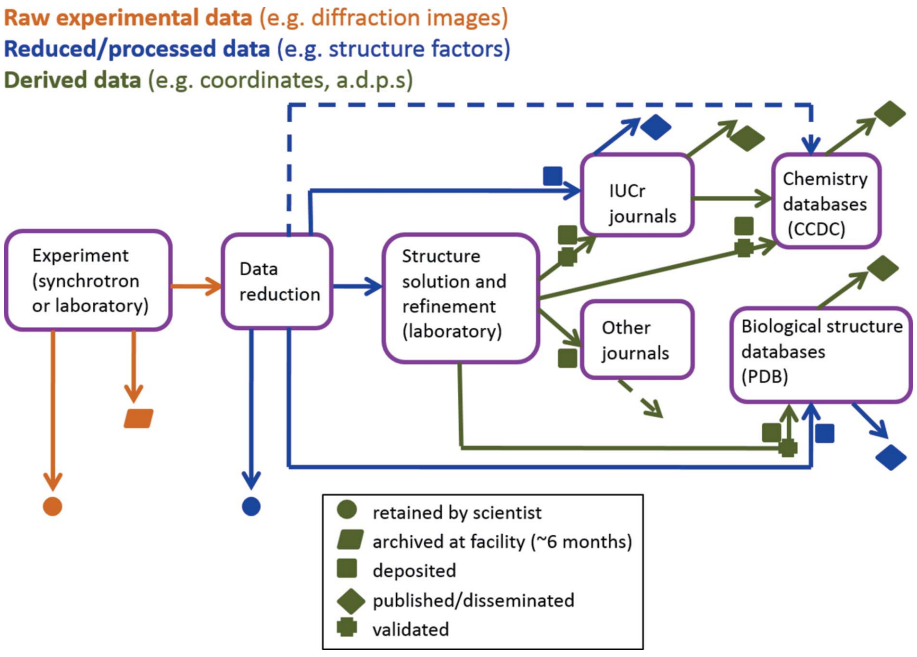


Figure 6
A coherent information flow in crystallography. CIF ontologies characterize data at every stage of the information processing life cycle, from experimental apparatus to published paper and curated database deposit.

3.2. The diversity of instrumentation

In this section we examine the specifics of some of the problems encountered in practice with missing or poorly characterized metadata. The availability

of metadata in image headers and their interpretation by software developers has been discussed previously (Tanley, Schreurs *et al.*, 2013; Kroon-Batenburg & Helliwell, 2014). It can safely be concluded that metadata information is often lacking or is ambiguous, *i.e.* can be interpreted in different ways. Hardware manufacturers may use different words for the same physical parameter or its units, and it is all in the hands of the software developers to make correct use of the metadata information and fill in the missing parts, simply by acquired knowledge or by trial and error. We refer to the supporting information in the paper by Kroon-Batenburg & Helliwell (2014) for a discussion between Kay Diederichs, Toine Schreurs and Loes Kroon-Batenburg about φ scans around an axis not perpendicular to the X-ray beam on a fixed χ goniometer. Though sufficient information was available in the header, the *XDS* software (Kabsch, 2010) ignored most of it and used knowledge of the (usual) instrumental set-up, which in this case did not suffice. Initially the raw data, which are now on the Manchester University Library archive, were stored on a website at Utrecht University (http://rawdata.chem.uu.nl) and we added a photograph of the experimental set-up as metadata to resolve the ambiguity of the goniometer, *e.g.* is the spindle axis pointing up or down?

We should distinguish between diffraction equipment designed to be used in combination with the manufacturer's software, which adequately handles metadata information, and assembled instruments like those on a synchrotron beamline. In the first case, taking the data to another place for use with third party software may give rise to problems, as described by Tanley, Diederichs *et al.* (2013). The image headers at best contain the type of goniometer (*e.g.* '`MACH3 with KAPPA`' for Bruker Proteum) but rarely are the orientations and dependencies of the four axes given. In the second case, commercial detectors (*e.g.* the Pilatus from Dectris) are installed on a beamline and it is the beamline control software, in close interaction with the detector software, that is responsible for writing information in the image headers. In this mixed environment not all metadata are captured. Usually, but not always, the wavelength, detector-to-sample distance, pixel size and number of pixels in either direction, rotation start angle and increment, and exposure time are given.

The most common problems with metadata, however, are related to the orientations of the goniometer axes and rotation directions, and the definition of the faster and slower directions in pixel coordinates with respect to the laboratory axes and the origin of the pixel coordinates; especially disturbing is the absence of or an incorrect beam centre (see below). Table 2 gives the goniometer definitions known to the *EVAL* software (Schreurs *et al.*, 2010) and shows their large variety.

An interesting tabulation of beamline settings for running *autoPROC* (Vonrhein *et al.*, 2011) is given at the website http://www.globalphasing.com/autoproc/wiki. Values such as `BeamCentreFrom = header:x,-y`, `ReversePhi = 'yes'` and `TwoThetaAxis = '-1'` are given in order to cope with similar problems to those mentioned above (Table 2). There are eight possible ways in which the pixel values in the image file relate to the physical detector face, and detector vendors use all eight possible conventions (Wladek Minor, private communication). A wrong beam centre can hamper the indexing step. One can estimate the beam centre by manual inspection, by calibration using powder diffraction, by taking a direct beam shot or by removing Bragg spots and using the solvent diffuse ring to find the beam centre (Vonrhein *et al.*, 2011); otherwise one has to resort to trial and error. Fig. 7 shows the mini-CBF header that is used by Dectris for Pilatus detectors. Most of the information is present but some parameters are ambiguous: `Beam_xy`: see discussion above; `Oscillation_axis` is given as '`X`': what is the X direction? `Polarization` is `0.990`: which plane has the strong intensity? We encountered an especially confusing situation where a Bruker fixed-χ goniometer was mounted with 90° rotation on Argonne beamline 15ID-B, while the images were converted to the normal Bruker instrument orientation. The strong polarization direction therefore appeared to be along the oscillation axis, but it was not (Jozef Kožíšek, private communication); only the string `TARGET SYNCHROTRON` in the header warned us.

Table 2
Implementation of goniometer types in *EVAL* (Schreurs *et al.*, 2010).

Goniometer	Axes, directions, off-set
Kappa	Axes: omega = z, kappa = k, phi = z, swing = z Rotation direction -1 -1 -1 -1 Values: omega, kappa, phi, swing, dist Kappa support angle
Euler	Axes: omega = z, chi = x, phi = z, swing = z Rotation direction 1 1 1 1 Values: omega, chi, phi, swing, dist
Horax	Axes: omega = y, chi = x, phi = z, swing = y Rotation direction 1 1 1 1 Values: omega, chi, phi, swing, dist
DTB	Axes: omega = z, chi = $-x$, phi = z, swing = y Rotation direction -1 -1 -1 1 Values: omega, chi, phi, swing, dist
X8	Axes: omega = z, chi = x, phi = z swing = z Rotation direction 1 -1 -1 1 Values: omega+180, chi, phi+90, swing
X8C	Axes: omega = z, chi = x, phi = z, swing = z Rotation direction 1 -1 -1 1 Values: omega+180, chi, phi+90, swing
Raxis	Axes: omega = z, chi = x, phi = z, swing = z Rotation direction -1 1 -1 1 Values: omega, chi, phi, swing
Kappa180	Axes: omega = z, kappa = k, phi = z, swing = z Rotation direction: -1 -1 -1 -1 Values: omega+180, kappa, phi, swing Kappa support angle

More *a priori* knowledge is often needed to interpret diffraction image data. For example, there are different conventions on how to record dead regions on the detector: strips between detector panels on Pilatus detectors are indicated by '`-1`', whereas in ADSC detector image files these are indicated by '`0`'. Data processing software has to interpret such pixel data correctly. Dark image and non-uniformity corrections may lead to negative intensities and some detector read-out handlers use a so-called baseline offset: a fixed integer number has been added to all pixel intensities to avoid having to store negative numbers. Removing the baseline offset is important in estimating the standard deviations of net

```
###CBF: VERSION 1.5, CBFlib v0.7.8 - SLS/DECTRIS PILATUS detectors

data_thaumatin1_collect_1_003

_array_data.header_convention "SLS_1.0"
_array_data.header_contents
;
# Detector: PILATUS 6M Prosport+, S/N 60-0100 Diamond
# 2010/Dec/06 16:53:40.416
# Pixel_size 172e-6 m x 172e-6 m
# Silicon sensor, thickness 0.000320 m
# Exposure_time 0.097500 s
# Exposure_period 0.100000 s
# Tau = 199.1e-09 s
# Count_cutoff 244849 counts
# Threshold_setting 6340 eV
# N_excluded_pixels = 1128
# Excluded_pixels: (nil)
# Flat_field: (nil)
# Trim_directory: p6m0100_T5p9_vrf_m0p2_090717
# Wavelength 0.9778 A
# Energy_range (0, 0) eV
# Detector_distance 0.28930 m
# Detector_Voffset 0.00000 m
# Beam_xy (1262.93, 1290.58) pixels
# Flux 0.0000 ph/s
# Filter_transmission 1.0000
# Start_angle 110.0000 deg.
# Angle_increment 1.0000 deg.
# Detector_2theta 0.0000 deg.
# Polarization 0.990
# Alpha 0.0000 deg.
# Kappa 0.0000 deg.
# Phi 0.0000 deg.
# Chi 0.0000 deg.
# Oscillation_axis X, CW
# N_oscillations 1
;

_array_data.data
;
--CIF-BINARY-FORMAT-SECTION--
Content-Type: application/octet-stream;
     conversions="x-CBF_BYTE_OFFSET"
Content-Transfer-Encoding: BINARY
X-Binary-Size: 6231565
X-Binary-ID: 1
X-Binary-Element-Type: "signed 32-bit integer"
X-Binary-Element-Byte-Order: LITTLE_ENDIAN
Content-MD5: JbLt2HJ+eYksuL3SBy3rFw==
X-Binary-Number-of-Elements: 6224001
X-Binary-Size-Fastest-Dimension: 2463
X-Binary-Size-Second-Dimension: 2527
X-Binary-Size-Padding: 4095
```

Figure 7
Mini-CBF header of the Dectris Pilatus detector.

Bragg reflection intensities and for measuring diffuse intensities between the Bragg peaks. Spatial distortion corrections are usually carried out and cannot be undone or corrected by processing software, but they affect standard deviations (Waterman & Evans, 2010) and this information should be conveyed in the metadata.

Detector hardware is being developed for high-speed serial crystallography experiments at X-ray free-electron laser (XFEL) installations or high-flux synchrotron beamlines that require ultra-fast data acquisition. A container HDF5 format, often with a NeXus data format layer on top, is designed for flexible and efficient input/output (I/O) for such high volumes of data. New data processing software packages such as *CrystFEL* (White *et al.*, 2012), *cctbx.xfel* (Sauter *et al.*, 2013) and *DIALS* (Waterman *et al.*, 2013) for serial crystallography are under development and this provides the opportunity to address the metadata issues anew.

Dectris has installed the Eiger detector at several synchrotron beamlines. Metadata are contained in a separate file (`master.h5`) linking to the image data files. The NeXus data representation (Könnecke *et al.*, 2015), like CIF, is very flexible and all metadata required can be captured by defining NeXus groups, fields and attributes. A good example of how consistent and comprehensive metadata can be stored in an imgCIF/CBF file is provided in Fig. 8 (Jörg Kaercher, Bruker AXS, private communication). In the proprietary Bruker `.sfrm` format the starting angles 2θ, ω, φ and χ are given (‘`ANGLES: ...`’). Their axis directions are not defined, whereas they are in the CBF format: the orientations and dependencies are given in the left-hand panel of Fig. 8(*b*). In `.sfrm` the rotation axis ‘`AXIS: 2`’ indicates ω, and the starting angle and increment are found at ‘`START:`’ and ‘`INCREME:`’; equivalent values are found in the CBF header at ‘`_diffrn_scan_axis.displacement_angle`’ and ‘`_diffrn_scan_axis.displacement_increment`’ (Fig. 8*b*, right-hand panel).

4. A concern and an action arising from the Rovinj Diffraction Data Deposition Workshop

A concern was voiced during open discussion at the workshop *via* the question ‘*Can we move away from the knowledge base in the various software packages, and make use of well developed metadata formats such as in CIF or NeXus?*’, *i.e.* a standardized raw diffraction image data format would make life easier for software developers but would require coordination between detector manufacturers. This has led directly to renewed calls for a standardized image format of appeal across the whole community. In conjunction with this question, the DDDWG is working on defining minimum requirements for metadata. We acknowledge that there will continue to be a great diversity of image formats (not least because of the existing installed base of detectors and the legacy data sets that have been archived), and conversion utilities such as *eiger2cbf* (https://github.com/biochem-fan/eiger2cbf) will continue to be needed. Nevertheless, it is important that anyone seeking to develop further new formats should be acutely aware of the need for adequate metadata characterization and interoperability that we have described above, and such an awareness may temper the proliferation of more new formats without particular demonstrable value.

In a separate discussion it was agreed that there is a need for a set of criteria for capturing and validating the essential experimental metadata for reproducibility of scientific results from any given raw data set. The proposal referred to this as ‘*checkCIF* for raw data’ and a close collaboration on this matter has been established with the IUCr COMCIFS (chaired by James Hester, who also attended the Rovinj Workshop). To develop these ideas further, a workshop run by the DDDWG is to take place at the ACA 2017 Conference in New Orleans in May 2017.

5. Concluding remarks

In this topical review we have provided descriptions of the rapidly developing interest in and storage options for the preservation and reuse of raw data within the scientific domain supervised by the IUCr and its Commissions. We have highlighted the initiatives of science policy makers towards an ‘Open Science’ model within which crystallographers will work in the future; this will bring new funding opportunities but also new codes of procedure within open science frameworks. Skills education and training for crystallographers and frank discussion will be needed. Overall, we now have the means and the organization for preservation of our raw data,

```
ELAPSDR:2.500000
ELAPSDA:2.564948
OSCILLA:0
NSTEPS :1
RANGE  :0.150000
START  :-8.850000
INCREME:-0.149994
NUMBER :60
NFRAMES:100
ANGLES :0.000000          351.149994        0.000000          54.782002
NOVER64:0                       0                       0
NPIXELB:1                                 1
NROWS  :1024
NCOLS  :1024
WORDORD:0
LONGORD:0
TARGET :CU
SOURCEK:0.000000
SOURCEM:0.000000
FILTER :Osmic "greens" Cu-24-48-6 (He purge +)
CELL   :1.000000      1.000000      1.000000      90.000000     90.000000
CELL   :90.000000
MATRIX :1.000000      0.000000      0.000000      0.000000      1.000000
MATRIX :0.000000          0.000000          0.000000          1.000000
LOWTEMP:0                       0                       -5474
ZOOM   :0.000000                0.000000                1.000000
CENTER :509.500000        514.000000        0.000000          0.000000
DISTANC:5.000000
TRAILER:0
COMPRES:none
LINEAR :1.000000                                  0.000000
PHD    :0.000000                                  0.000000
PREAMP :6                                         0
CORRECT:8000G6H500
WARPFIL:8000G6H500
WAVELEN:1.541840          0.000000          0.000000          0.000000
MAXXY  :339.000000                        370.000000
AXIS   :2
:
```

(*a*)

```
_diffrn_radiation_wavelength.wt 1.0
_diffrn_radiation_wavelength.polarizn_source_ratio 0
_diffrn_radiation_wavelength.polarizn_source_norm 0

_diffrn_radiation.wavelength_id SOURCE

loop_
_axis.id
_axis.depends_on
_axis.equipment
_axis.type
_axis.vector[1]
_axis.vector[2]
_axis.vector[3]
_axis.offset[1]
_axis.offset[2]
_axis.offset[3]
 OMEGA . goniometer rotation 1 0 0 0 0 0
 CHI OMEGA goniometer rotation 0 0 1 0 0 0
 PHI CHI goniometer rotation -1 0 0 0 0 0
 TWOTHETA . detector rotation 1 0 0 0 0 0
 DX TWOTHETA detector translation 0 0 -1 0 0 0
 YAW DX detector rotation 1 0 0 0 0 0
 PITCH YAW detector rotation 0 -1 0 0 0 0
 ROLL PITCH detector rotation 0 0 1 0 0 0
 H ROLL detector translation 0 -1 0 0 0 0
 V H detector translation -1 0 0 0 0 0
 ELEMENT_X V detector translation 0 -1 0 -45.698 -45.698 0
 ELEMENT_Y ELEMENT_X detector translation 1 0 0 0 0 0

loop_
:
```

```
 TWOTHETA ? 0
 OMEGA ? 351.149994
 PHI ? 0
 CHI ? 54.782002
 H -0.223134594787576 ?
 V 0.178507675830061 ?
 PITCH ? 0
 ROLL ? 0
 YAW ? 0

loop_
_diffrn_scan_axis.axis_id
_diffrn_scan_axis.displacement_start
_diffrn_scan_axis.displacement_increment
_diffrn_scan_axis.displacement_range
_diffrn_scan_axis.angle_start
_diffrn_scan_axis.angle_increment
_diffrn_scan_axis.angle_range
 DX 53 0 0 ? ? ?
 TWOTHETA ? ? ? 0 0 0
 OMEGA ? ? ? 351.149994 -0.149993999999992 -0.149993999999992
 PHI ? ? ? 0 0 0
 CHI ? ? ? 54.782002 0 0
 H -0.223134594787576 0 0 ? ? ?
 V 0.178507675830061 0 0 ? ? ?
 PITCH ? ? ? 0 0 0
 ROLL ? ? ? 0 0 0
 YAW ? ? ? 0 0 0

_diffrn_measurement.id GONIOMETER
_diffrn_measurement.diffrn_id BRUKER
:
```

(*b*)

Figure 8
Comparison of header data in Bruker (*a*) `.sfrm` and (*b*) CBF formats.

but still the need for careful thought about the metadata descriptors for each of the IUCr Commissions continues to be pressing. We note that the Commissions work within a diversity of instrumentation, and so a range of actions is required to improve on this current situation.

We have identified specifically the need to revisit the imperative for the community to adopt a standardized image format, and to agree at least a minimal set of essential metadata for reproducibility. The imgCIF dictionary (Hammersley *et al.*, 2005) is the natural starting point for the former, and the interaction between COMCIFS and NIAC (Könnecke *et al.*, 2015) demonstrates the feasibility of applying a common ontology across differing physical formats. There are also grounds for optimism that the idea of '*checkCIF* for raw data' will appeal to both researchers and instrument vendors, given the enthusiastic representation of both at the Rovinj Workshop. As with all such initiatives, the rate of uptake will depend on drivers within the community. In the case of the original '*checkCIF*' for derived data, structural science journals (especially those of the IUCr) that demanded relevant metadata and consistency checking provided one such important driver. In the case of raw data, which underpins all subsequent scientific deductions and derivations, we are encouraged by the emerging policies on research data management that we have summarized in this article, and by the many archiving initiatives that have sprung up around X-ray diffraction images in the space of the past few years.

Acknowledgements

We are grateful to the IUCr for continuing support of DDDWG activities, including the Workshop in Rovinj that led

to this and a number of other articles. We are very grateful to various research institutes and universities who sent their staff to take part in that Workshop. Support for technical services and associated staffing costs was contributed by Dectris, IUCr Journals, CODATA, the Cambridge Crystallographic Data Centre, Bruker, FIZ Karlsruhe/ICSD, Oxford Cryosystems and Wiley, to whom we are very grateful. We are also indebted to the Croatian Association of Crystallographers for their active help in securing the best possible Workshop to address this important topic.

References

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

Bernstein, H. J. (2005). *Classification and Use of Image Data*. *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall and B. McMahon, pp. 199–205. Dordrecht: Springer.

Bernstein, H. J., Bollinger, J. C., Brown, I. D., Gražulis, S., Hester, J. R., McMahon, B., Spadaccini, N., Westbrook, J. D. & Westrip, S. P. (2016). *J. Appl. Cryst.* **49**, 277–284.

Bernstein, H. J., Sloan, J. M., Winter, G., Richter, T. S., NIAC & COMCIFS (2013). *Coping with BIG DATA Image Formats: Integration of CBF, NeXus and HDF5*. American Crystallographic Association Meeting, 20–24 July, 2013, Honolulu, Hawaii, USA. Poster T-16.

Diederichs, K. & Karplus, P. A. (2013). *Acta Cryst.* D**69**, 1215–1222.

Digital Curation Centre (2016). *Overview of Funders' Data Policies.* http://www.dcc.ac.uk/resources/policy-and-legal/overview-funders-data-policies.

ESRF (2013). *ESRFnews*, December ed., pp. 14–21. ESRF, Grenoble, France.

Grabowski, M., Langner, K. M., Cymborowski, M., Porebski, P. J., Sroka, P., Zheng, H., Cooper, D. R., Zimmerman, M. D., Elsliger, M.-A., Burley, S. K. & Minor, W. (2016). *Acta Cryst.* D**72**, 1181–1193.

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

Gutmanas, A., Oldfield, T. J., Patwardhan, A., Sen, S., Velankar, S. & Kleywegt, G. J. (2013). *Acta Cryst.* D**69**, 710–721.

Hall, S. R., Allen, F. H. & Brown, I. D. (1991). *Acta Cryst.* A**47**, 655–685.

Hall, S. R. & McMahon, B. (1995). Editors. *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*. Dordrecht: Springer.

Hall, S. R. & McMahon (2016). *Data Sci. J.* **15**, 3.

Hammersley, A. P., Bernstein, H. J. & Westbrook, J. D. (2005). *Image Dictionary (imgCIF)*. *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall and B. McMahon, pp. 444–458. Dordrecht: Springer.

Hester, J. R. (2016). *Data Sci. J.* **15**, 12.

International Structural Genomics Organization (2001). *Report of Task Force on Numerical Criteria in Structural Genomics.* http://www.isgo.org/organization/members07/010410.html.

Jacques, D. A., Guss, J. M., Svergun, D. I. & Trewhella, J. (2012). *Acta Cryst.* D**68**, 620–626.

Jones, B. (2015). *Towards the European Open Science Cloud*. http://doi.org/10.5281/zenodo.16001.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.

Kachala, M., Westbrook, J. & Svergun, D. (2016). *J. Appl. Cryst.* **49**, 302–310.

Könnecke, M. *et al.* (2015). *J. Appl. Cryst.* **48**, 301–305.

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

Marsh, R. E., Kapon, M., Hu, S. & Herbstein, F. H. (2002). *Acta Cryst.* B**58**, 62–77.

Matthews, B., Sufi, S., Flannery, D., Lerusse, L., Griffin, T., Gleaves, M. & Kleese, K. (2010). *Int. J. Digit. Curation*, **5**, 106–118.

Messori, L. & Merlino, A. (2016). *Coord. Chem. Rev.* **315**, 67–89.

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

Meyer, P. A. *et al.* (2016). *Nat. Commun.* **7**, 10882.

Minor, W., Dauter, Z., Helliwell, J. R., Jaskolski, M. & Wlodawer, A. (2016). *Structure*, **24**, 216–220.

National Science Foundation (2010). *Data Management and Sharing Frequently Asked Questions (FAQs).* http://www.nsf.gov/bfa/dias/policy/dmpfaqs.jsp.

Northwestern University Library (2016). *Data Management: Federal Funding Agency Requirements.* http://libguides.northwestern.edu/datamanagement/federalfundingagency.

Ravel, B., Hester, J. R., Solé, V. A. & Newville, M. (2012). *J. Synchrotron Rad.* **19**, 869–874.

Research Councils UK (2015). *Guidance on Best Practice in the Management of Research Data.* http://www.rcuk.ac.uk/documents/documents/rcukcommonprinciplesondatapolicy-pdf/.

Sauter, N. K., Hattne, J., Grosse-Kunstleve, R. W. & Echols, N. (2013). *Acta Cryst.* D**69**, 1274–1282.

Schreurs, A. M. M., Xian, X. & Kroon-Batenburg, L. M. J. (2010). *J. Appl. Cryst.* **43**, 70–82.

Science International (2015). *Open Data in a Big Data World.* Paris: International Council for Science (ICSU), International Social Science Council (ISSC), The World Academy of Sciences (TWAS), InterAcademy Partnership (IAP).

Shabalin, I., Dauter, Z., Jaskolski, M., Minor, W. & Wlodawer, A. (2015). *Acta Cryst.* D**71**, 1965–1979.

Strickland, P. R., Hoyland, M. A. & McMahon, B. (2005). *Small-Molecule Crystal Structure Publication Using CIF*. *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall and B. McMahon, pp. 557–569. Dordrecht: Springer.

Tanley, S. W. M., Diederichs, K., Kroon-Batenburg, L. M. J., Levy, C., Schreurs, A. M. M. & Helliwell, J. R. (2015). *Acta Cryst.* D**71**, 1982–1983.

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

Tanley, S. W. M., Schreurs, A. M. M., Helliwell, J. R. & Kroon-Batenburg, L. M. J. (2013). *J. Appl. Cryst.* **46**, 108–119.

Tanley, S. W. M., Schreurs, A. M. M., Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2016). *Acta Cryst.* F**72**, 253–254.

Terwilliger, T. C. (2012). *Continuous Improvement of Macromolecular Crystal Structures*. *ICSTI Insights: The Living Publication*, pp. 16–29 (http://www.icsti.org/IMG/pdf/Living_publication_Final-2.pdf). Paris: ICSTI.

Terwilliger, T. C. (2014). *Acta Cryst.* D**70**, 2500–2501.

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

Toby, B. H. (2005). *Classification and Use of Powder Diffraction Data*. *International Tables for Crystallography*, Vol. G, *Definition and Exchange of Crystallographic Data*, edited by S. R. Hall and B. McMahon, pp. 117–130. Dordrecht: Springer.

Vonrhein, C., Flensburg, C., Keller, P., Sharff, A., Smart, O., Paciorek, W., Womack, T. & Bricogne, G. (2011). *Acta Cryst.* D**67**, 293–302.

Waterman, D. & Evans, G. (2010). *J. Appl. Cryst.* **43**, 1356–1371.

Waterman, D. G., Winter, G., Parkhurst, J. M., Fuentes-Montero, L., Hattne, J., Brewster, A., Sauter, N. K. & Evans, G. (2013). *CCP4 Newsl. Protein Crystallogr.* **49**, 16–19.

White, T. A., Kirian, R. A., Martin, A. V., Aquila, A., Nass, K., Barty, A. & Chapman, H. N. (2012). *J. Appl. Cryst.* **45**, 335–341.

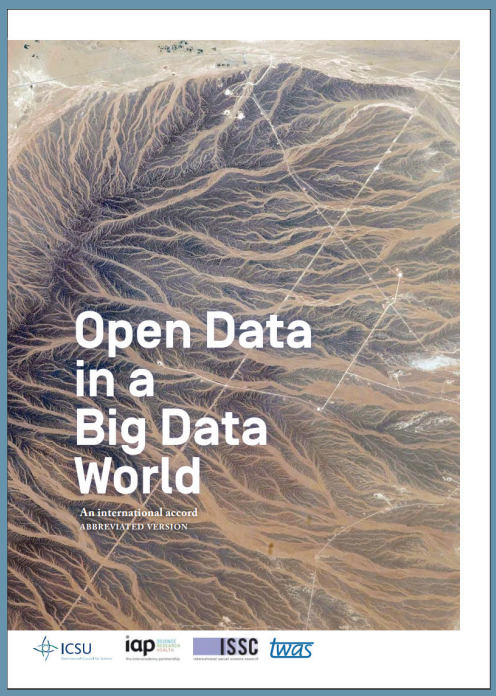
Open Data in a Big Data World

A position paper for crystallography

The 2015 Science International Accord

The International Union of Crystallography (IUCr) notes the publication by the International Council for Science (ICSU), the InterAcademy Partnership (IAP), The World Academy of Sciences (TWAS) and the International Social Science Council (ISSC) of an international Accord on *the values of open data in the emerging scientific culture of big data*, following the 2015 Science International meeting. The IUCr *acknowledges the importance* of this Accord, and *endorses the analysis of the values of open data and the Principles of Open Data* set out in the document *Open Data in a Big Data World*, published in short and long forms on the ICSU website at http://www.icsu.org/science-international/accord.

The Accord is very general, and has applicability across the entire panorama of science, which it defines as embracing 'all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines'. Because the specific values, significance and implementation of Open Data principles will vary in detail between disciplines, the IUCr considers it useful to contribute this *detailed response* to the Accord, as a case study of best practice emerging in one particular field.



This text was prepared for the IUCr by

Marvin L. Hackert, *IUCr President and IUCr Representative to ICSU*
Department of Molecular Biosciences, University of Texas at Austin, Austin, TX 78712, USA

Luc Van Meervelt, *IUCr General Secretary and Treasurer*
Chemistry Department, Katholieke Universiteit Leuven, Celestijnenlaan 200F, BE-3001, Leuven, Belgium

John R. Helliwell, *IUCr Representative to CODATA*
School of Chemistry, University of Manchester, Oxford Road, Manchester M13 9PL, UK

Brian McMahon, *IUCr Research and Development Officer*
International Union of Crystallography, 5 Abbey Square, Chester CH1 2HU, UK

August 2016

Executive Summary

Science is best served when access barriers to data (and publications) are low. However, the maintenance of the highest levels of quality in collecting, evaluating, storing and curating data is a very expensive component of the scientific process. Crystallography has a diverse ecosystem of approaches to sustainability and quality assurance.

Technological advances in scientific instrumentation and computer technology have dramatically increased the quantities of data involved in scientific inquiry. This Accord expresses the dependence of scientific assertions on supporting data. Inasmuch as science is by its nature interrogative, the Accord asserts that 'openness and transparency have formed the bedrock on which the progress of science in the modern era has been based'. The IUCr supports this assertion, but notes that openness alone (by which we mean the ability to access and re-use scientific data with little or no restriction) is not sufficient. The data openly accessed must be subject to critical scrutiny, through peer review and automated validation where possible. There is also a case to be made for raw data to be retained for as long a time as feasible, to permit re-evaluation that takes account of novel analytic techniques or that periodically employs different methodologies to eliminate systematic procedural error.

All scientific data must be subject to rigorous first analysis to exclude or quantify systematic bias or error; all software implementations should employ open algorithmic procedures and their results should ideally be cross-checked by independent implementations. An overlooked challenge in handling ever-growing volumes of data is the need to apply the same level of critical evaluation as has been applied to historically smaller volumes.

Openness alone is not sufficient. Data openly accessed must be subject to critical scrutiny, through peer review and automated validation where possible

The Accord does not formally define 'Open Data', but implies certain properties throughout its careful discussions. We hold that the essential component of openness is that the data supporting any scientific assertion should be

- **complete** (*i.e.* all data collected for a particular purpose should be available for subsequent re-use); and
- **precise** (the meaning of each datum is fully defined, processing parameters are fully specified and quantified, statistical uncertainties evaluated and declared).

Together, these properties include the criteria of Paragraph 8 of the Accord (long form), that open data should be *discoverable*, *accessible*, *intelligible*, *assessable* and *usable*. We note, however, that a full understanding of the data may depend on associated scientific publications that discuss the details of data processing where these differ from routine practice. The full linking of article and data is another key element of openness.

Science is best served when access barriers to data (and publications) are low. A major barrier to access is cost, and the phrase 'open access' is often used to characterize access to data and publications that involve no charge to the end-user. However, the maintenance of the highest levels of quality in collecting, evaluating, storing and curating data is a very expensive component of the scientific process, and care must be taken to understand how to obtain the maximum benefit from public funding of science. In many fields, it may indeed be cost-effective to provide direct funding to repositories or publishing platforms that require no further payment to access. In other fields, the situation is less clear cut.

Crystallography has a diverse ecosystem of disciplinary databases, data repositories, experimental facilities and publishers. Several of these are sustained through subscription-based access; but the other side of the coin is that they ingest, evaluate and publish data and information at no charge to the author/depositor, and without imposing any additional charge on the public purse. At the present time, this variety of approaches to sustainability and quality assurance serves this discipline well.

The SACLA X-ray free-electron laser facility in Hyogo, Japan, makes it possible to observe atoms and molecules in real time – generating vast amounts of data in the process.

Photo credit: RIKEN/XFEL

APPENDIX: Annotated Accord

We illustrate the points made in the *Executive Summary* by annotating the relevant parts of the short form of the Accord (reproduced in red below). Where we make no explicit comment, it may be taken that we are in tacit agreement with that part of the Accord.

This accord is presented as an outcome of "Science International 2015", the first of a series of annual meetings of four top-level representatives of international science (the International Council for Science – ICSU, the InterAcademy Partnership – IAP, The World Academy of Sciences – TWAS and the International Social Science Council – ISSC) that are designed to represent the global scientific community in the international policy for science arena. The accord identifies the opportunities and challenges of the data revolution as today's predominant issue for global science policy. It proposes fundamental principles that should be adopted in responding to them. It adds the distinctive voice of the scientific community to those of governments and inter-governmental bodies that have made the case for open data as a fundamental pre-requisite in maintaining the rigour of scientific inquiry and maximising public benefit from the data revolution in both developed and developing countries. Science International partners will promote discussion and adoption of these principles and their endorsement by their respective members and by other representative bodies of science at national and international levels.

The IUCr welcomes the interest of high-level international stakeholders in presenting a united voice that stresses the importance of scientific inquiry world-wide. In a world of expensive research programmes, often largely dependent for funding on income raised from public taxation, it is important that the needs and opportunities of small and developing countries are considered alongside those of the developed world.

The IUCr welcomes the interest of high-level international stakeholders in presenting a united voice that stresses the importance of scientific inquiry world-wide

The IUCr represents the worldwide community of scientists in the field of crystallography and related structural sciences. It comprises 50 Adhering Bodies representing 58 distinct nations. The IUCr itself is a member of ICSU and of CODATA, ICSU's Committee on Data for Science and Technology.

1. The Big Data World

The digital revolution of recent decades is a world historical event as deep [as] and more pervasive than the introduction of the printing press. It has created an unprecedented explosion in the capacity to acquire, store, manipulate and instantaneously transmit vast and complex data volumes, with profound implications for science[1]. The rate of change is formidable. In 2003 scientists declared the mapping of the human genome complete. It took over 10 years and cost $1billion – today it takes mere days and a small fraction

Macromolecular structures in Protein Data Bank: > 125,000

of the cost ($1000). "Big data", in which unprecedented fluxes of data stream in and out of computational systems, and "Broad Data" in which numerous datasets can be semantically linked to create deeper meaning, are the engines of this revolution, offering novel opportunities to natural, social and human sciences.

[1] The word "science" is used to mean the systematic organisation of knowledge that can be rationally explained and reliably applied. It is used, as in most languages other than English, to include all domains, including humanities and social sciences as well as the STEM (science, technology, engineering, medicine) disciplines.

While the opening section correctly raises the subject of the large volume of research data routinely generated and collected nowadays, and its applicability across many subject areas, we note that the proper conduct of science has always depended on a deep understanding of the nature of the data collected in any research effort, and a careful and proper analysis of its accuracy, precision and validity.

While very high data volumes and data acquisition rates may make it increasingly difficult to treat data with the care and respect that it needs, it is nevertheless essential to good science that every effort is made to do so. The ubiquity of data is no substitute for proper and critical analysis.

2. The Opportunities

The scientific opportunities of this data-rich world lie in discovering patterns that have hitherto been beyond our reach; in linking and correlating different aspects of systems better to understand their behaviour; in characterising complexity; and in iterating between descriptions of the state of a complex system and simulations that forecast its dynamic behaviour. There are many areas of research where such capacities are deeply relevant: in weather and climate forecasting; in understanding the workings of the brain; in the behaviour of the global economy; in evaluating agricultural productivity; in demographic forecasts; in unravelling histories; and in many contemporary global challenges such as those of environmental change, infectious disease and mass migration that require combined insights and data from many disciplines.

The proper conduct of science has always depended on a deep understanding of the nature of the data collected … and a careful and proper analysis of its accuracy, precision and validity

Crystallography abounds in examples of scientific laws and applications being derived from collecting and correlating data. Historic classification of naturally occurring crystal forms led to the understanding of lattice symmetries, packing arrangements and energetics. Subsequent probing of crystal structures by X-ray diffraction and other techniques

Molecular structures in Cambridge Structural Database: > 800,000

yielded vital information on the nature of chemical bonds, molecular structures and solid-state properties of materials. The elaboration of the structures of DNA and proteins created great insights into biological processes, and the availability of large and growing databases of nucleic acid and protein structures feeds into the enormous advances being made in genetics and therapeutics. Pioneering developments in time-resolved structural dynamics using synchrotrons and X-ray free-electron lasers probe the very nature of chemical reactions. Each of these developmental phases has involved

Inorganic and organic structures in Crystallography Open Database: > 365,000

ever-growing volumes and complexity of data that challenged the science of the time. We see the current 'Data Deluge' as just the latest (albeit large) step up in this evolutionary process. We welcome both the challenges it brings and the prospects for future discovery.

3. The Challenges

Grasping these opportunities poses serious challenges to the way science is done and organised. Open data are the common, enabling threads.

The Open Data Imperative

The fundamental role of publicly funded research is to add to the stock of knowledge and understanding that are essential to human judgements, innovation and social and personal wellbeing. The technologies and processes of the digital revolution provide a powerful medium through which scientific productivity and creativity can be enhanced by permitting data and ideas to flow openly, rapidly and pervasively through the networked interaction of many minds. If this social revolution in science is to be realised it is vital that we adopt a default position that publicly funded data should be made publicly accessible and re-usable when a

research project through which the data have been collected is completed.

While the bulk of this Accord focuses on publicly funded research – appropriately, as its main purpose is to help shape public policy – many of the principles it discusses are important to the proper conduct of science in any manifestation, including its practice within the private sphere. Some privately funded research

Inorganic crystal and powder diffraction data sets in Powder Diffraction File: > 384,000

does contribute directly to the public good, *e.g.* through academic publications; some is harnessed for commercial gain. While there are some additional pressures that reduce the ways in which such data are openly shared outside of the originating stakeholder, we nevertheless feel that the principles stated in this Accord should be considered as ideals towards which all scientific effort should aspire. We see below that, even in the area of publicly funded research, there can be factors that moderate the practical open dissemination of data. We urge the worldwide community of scientists, whether publicly or privately funded, always to have the starting goal to divulge fully all data collected or generated in experiments, and to temper this goal only so far as is absolutely necessary to allow the basic enterprise to be maintained in a sustainable way and with full scientific and ethical integrity.

Maintaining self-correction

Openness of the evidence (the data) for scientific claims is the bedrock of scientific progress. It permits the logic of an argument to be scrutinised and the reproducibility of observations or experiments to be tested, thereby supporting or invalidating those claims. When a paper making a scientific claim is published, it is essential that the evidentiary data, the related metadata that permit their re-analysis, and the codes used in computer manipulation are made concurrently open to scrutiny to ensure that the vital process of self-correction is maintained. Recent demonstrations in several disciplines of high rates of non-reproducibility of results of published papers emphasise the crucial need to re-invigorate open data processes for a big data world. Openness is not however enough. Data must be intelligently open, meaning that they should be: discoverable, accessible, intelligible, assessable and (re-)usable.

The reference to *evidentiary* data is a useful reminder that data collected in the course of a scientific inquiry may play a variety of roles. Our summary of ways in which crystallographic data has led to novel scientific hypotheses and conclusions (Section 2) largely refers to data sets that are themselves derived scientific models. (They are tabulations of atomic positional and displacement parameters, with associated information about chemical nature, modelling constraints or restraints, and many other metadata relating to provenance, analytical procedures, calculated precision of derived values, *etc.*) These models are constructed from experimental data, typically the diffracted intensities from a scattered collimated radiation or particle beam.

We urge the worldwide community of scientists, whether publicly or privately funded, always to have the starting goal to divulge fully all data collected or generated in experiments

Inorganic crystal structures in Inorganic Crystal Structure Database: > 185,000

These evidentiary (experimental) data supporting each structural data set should also be made available as part of the scientific record. For biological macromolecular structures, the community expects that such evidentiary data should be deposited in the same curated database (the Worldwide Protein Data Bank) as the structures themselves. These tabulations of intensities ('structure factors') are also required for small-molecule or inorganic structures published by the journals of the IUCr. They are not always required as part of the submission of structural results by other journal publishers, but their deposition in the relevant structural databases is increasingly encouraged as best community practice. There has long been a convention in crystallography that individual structural data sets and supporting sets of structure factors may be freely downloaded, even from journals or databases where subscriptions are required to access the published articles or complete database.

Pauling File: > 41,000 phase diagrams, > 290,000 crystal structures, > 106,000 physical property entries

More recently, there has been growing interest within the field of crystallography to retain the raw data for each structure determination experiment. These most commonly take the form of a collection of two-dimensional images capturing the diffracted beams as the crystal sample is rotated through all orientations relative to the incident beam. It is from these images that the more concise set of processed diffraction data (structure factors) is derived. While a set of structure factors is typically a few megabytes (MB) in size, the raw diffraction images may occupy many gigabytes (GB). For the traditionally small-scale data volume requirements of structural science, this does amount to a foray into 'Big Data'. Rapid improve-

ments in detector technology and developments in dynamical structure elucidation with intense synchrotron and X-ray laser sources will increase the volume of raw data collected by further orders of magnitude.

There has been considerable discussion in the last few years of the need to archive the primary data sets [1, 2]. Many researchers consider that the structure factors are adequate (in most cases) to validate the derived structural model. To the extent that crystal structure determination experiments are largely homogeneous in their methods and equipment, a high degree of confidence can be placed in the reduction processes that lead to the structure factors. However, errors of interpretation are sometimes made, and access to the raw data can help to mitigate this. Furthermore, diffuse scattering in the original images contains information about internal molecular dynamics or the correlated dynamics over different distance scales in the crystal. This is generally ignored during standard data reduction, and so potentially valuable scientific information is lost. There is a growing sense that at least some proportion of raw data images should therefore be retained for purposes of validation, new scientific discovery, and development or testing of novel software methods.

DDDWG

The Diffraction Data Deposition Working Group of the IUCr has been active since 2011, scoping the demand and practical requirements for routine deposition of diffraction images, the raw experimental data sets from many crystallographic experiments. Activities also involve the characterisation of essential metadata for describing the great variety of crystallographic and related structural experiments. Links to Workshops, discussion forums and other activities are at http://www.iucr.org/resources/data/dddwg.

There is also value in retaining primary data as a safeguard against the publication of results that are fraudulently derived. However, this should not be overstated; individuals motivated to fabricate scientific evidence may still find ways of doctoring primary experimental data. Although access to primary data can help to discourage unethical scientific behaviour, it cannot act as a complete preventative. This example serves to highlight the fact that more data, by itself, does not change the need to treat all data with appropriate care, respect and critical analysis.

CrystMet database of metals, alloys and intermetallics: > 161,000 entries

There has been growing interest within the field of crystallography to retain the raw data for each structure determination experiment

IUCr-sponsored data exchange standards provide for very detailed metadata to define precisely the content and context of a data set, and recent efforts have focused on the need to define (for all experiments) the metadata needed to understand fully the data collected, and to permit reproducibility of the experiment [3].

Adapting scientific reasoning

Many of the complex relationships that we now seek to capture through big- or broad- linked data lie far beyond the analytical power of many classical statistical methods. They require deeper mathematical approaches including topological methods to ensure that inferences drawn from big data and broad data are valid. Data-intensive machine-analysis and machine-learning are becoming ubiquitous, and have major implications for scientific discovery. The complexity of patterns that machines are able to identify are not easily grasped by human cognitive processes, posing profound issues about the human-machine interface and what it might mean to be a researcher in the 21st century.

As indicated in the introductory comments, crystallography is already an established leader in deriving complex relationships from extensive data collections, albeit much of this research successfully uses classical statistical methods. Further advances will be facilitated by harnessing the potential of 'broad-linked' data, *i.e.* by permitting text mining of publications and data mining of their associated data sets (including, as appropriate, the raw or processed experimental data that underpin the structural data models). Furthermore, automated discovery and analysis are assisted by the curation of a discipline-specific machine ontology (the Crystallographic Information Framework, CIF) [4] and the development of software that can use this ontology directly to test and follow linkages between granular concepts expressed within the data or associated publications.

In the case of IUCr journals, open-access articles are published through a Creative Commons attribution licence that permits text mining. For other articles whose publication is financed through journal subscriptions (a 'paywall'), the IUCr will provide free access for text-mining robots to *bona fide* researchers.

Work continues under the sponsorship of the IUCr to increase interoperability between the growing family of crystallography related ontologies ('CIF dictionaries') and cognate ontologies in related areas of science (*e.g.* Chemical Markup Language, CML, in the description of chemical structures and reactions; NMRStar for protein NMR conformation studies; macromolecular

structure application profiles in NeXus/HDF5 image acquisition files).

Bilbao Incommensurate Structures Database: > 130 incommensurate modulated and composite structures

Ethical constraints

The open data principle has ethical implications for researchers and research subjects. It can appear to override the individual interests of the researchers who generate the data, such that novel ways of recognising and rewarding their contribution need to be developed. The privacy of data subjects needs to be protected. In a regime of open sharing in which data are passed on from their originators, there is loss of control over future usage, whilst anonymisation procedures have been demonstrated to be unable to guarantee the security of personal records.

Crystallographic Information Framework

The Crystallographic Information Framework is a suite of machine-readable ontologies, data exchange formats and software applications and services developed by the IUCr since 1991 for data definition and exchange in crystallography and related structural sciences. The standards are fully documented on the web (cif.iucr.org) and in print (*International Tables for Crystallography Volume G: Definition and exchange of crystallographic data*). Specific ontologies exist for single-crystal and powder X-ray diffraction, biological macromolecular structures (proteins and nucleic acids), modulated and composite structures, electron density, twinning, symmetry and diffraction images.

Open global participation

Big data and open data have great potential to benefit less affluent countries, and especially least developed countries (LDCs). However, LDCs typically have poorly resourced national research systems. If they cannot participate in research based on big and open data, the gap could grow exponentially in coming years. They will be unable to collect, store and share data, unable to participate in the global research enterprise, unable to contribute as full partners to global efforts on climate change, health care, and resource protection, and unable fully to benefit from such efforts, where global solutions will only be achieved if there is global participation. Thus, both emerging and developed nations have a clear, direct interest in helping to fully mobilize LDC science potential and thereby to contribute to achievement of the UN Sustainable Development Goals.

A major project of the International Year of Crystallography in 2014 was to launch a series of capacity-building 'open laboratories' in many developing countries [5]. These often involved the loan of equipment from commercial vendors and hands-on training in the use of the equipment and the proper handling of generated data. For research in chemical crystallography, many results will be generated in local laboratories. Most equipment and software uses the open CIF standard; the IUCr and other crystallographic institutions provide free or open-source software for standard reduction and analysis of the experimental data, for characterizing the derived structural data sets, and for preparing articles for publication. Open-access article processing charges are reduced or waived for authors from developing countries. For larger-scale or more complex experiments, often conducted in synchrotron or neutron facilities, there are initiatives to develop regional resources (*e.g.* in the Middle East and Africa) that will provide access to the necessary equipment for LDCs. Through liaison with established facilities, IUCr working groups aim to encourage common modes of practice amongst the larger facilities with respect to data management and archiving.

Seizing the opportunity

Effective open data can only be realised if there is systemic action at personal, disciplinary, national and international levels. Although science is an international enterprise, it is done within distinctive national systems of responsibility, organisation and management, all of which need to respond to the opportunity. Research funders and research performing institutions should fund and implement processes that lighten the burden on researchers of making data intelligently open and that support open data processes. Increasing numbers of research communities have discovered the benefits of sharing data, in fields as varied as linguistics, bio-informatics and chemical crystallography, and have made major strides in realising benefit for their disciplines through international collaboration in facilitating access and use of open data. Responsibilities also fall on international bodies, such as the International Council for Science's (ICSU) Committee on Data for Science and Technology (CODATA), its World Data System (WDS) and the Research Data Alliance (RDA), to promote and support developments of the systems and procedures that will ensure international data access, interoperability and sustainability.

A major project of the International Year of Crystallography in 2014 was to launch a series of capacity-building 'open laboratories' in many developing countries

Discrete data items defined in the core CIF dictionary: 802

The IUCr formally documents every aspect of its data standardization programme on its website and through its journals and reference works [6]. It is an active member of CODATA, and seeks synergies with WDS and the RDA, and other international organizations such as the International Council for Scientific and Technical Information, ICSTI.

Open science and public knowledge

The idea of "open science" has developed in recognition of the need for stronger dialogue and engagement by the scientific community with wider society in addressing many current problems through reciprocal framing of the issues and the collaborative design, execution and application of research. There are, of course, legitimate limits to openness, such as the need to protect security, privacy and proprietary concerns through judiciously applied mechanisms. There are also countervailing trends towards privatisation of knowledge that are at odds with the ethos of scientific inquiry and the basic need of humanity to use ideas freely. If the scientific enterprise is not to founder under such pressures, an assertive commitment to principles of open data, open information and open knowledge is required from the global scientific community.

Discrete data items defined in the macromolecular CIF extension dictionary: 5631

The IUCr included many public outreach activities in its programme for the International Year of Crystallography, and is committed to maintain and expand such activities [7].

There are precedents (especially in structural biology) for retaining exclusive rights to access experimental data sets for a finite period of time. While such embargoes are permitted *e.g.* by the Worldwide Protein Data Bank, the relevant IUCr bodies are supportive of a move towards minimizing or removing them altogether.

We recommend caution in the use of terms such as 'privatisation'. While the IUCr supports the ideal of full open access to scientific data and knowledge, the proper maintenance and curation of databases, data repositories and publications is expensive. For a variety of reasons, not all such facilities are funded directly or fully from the public purse, and scientific endeavour remains a diverse ecosystem in terms of funding and business models. IUCr Journals, first published in 1948, grew according to the universal subscription model of the time. Although there is movement in the direction of open-access publication (the IUCr has two fully open-access titles), there are still many authors who will not or cannot pay the article processing charges necessary to sustain this publishing model. Therefore we currently offer a hybrid model where individual articles may be open access or behind a subscription paywall. Similarly, the most comprehensive database of small-molecule chemical structures (the Cambridge Structural Database, CSD) is funded through subscriptions, again for historical reasons: a previous national Government insisted that public funding of the academically based Cambridge Crystallographic Data Centre that maintains the CSD be replaced by a self-sustaining business model. In cases such as this, the 'private' status of scientific service providers does not imply that their primary objective is other than to advance the cause of science.

Structural data sets freely available from IUCr journals: > 58,800

Note also our comments near the start of Section 3 where we commend the principles set out in this Accord as equally applicable to publicly and privately funded research.

4. Principles of Open Data

Such is the importance and magnitude of the challenges to the practice of science from the data revolution that Science International believes it appropriate to promote the following statement of principles of open data.

The IUCr formally documents every aspect of its data standardization programme on its website and through its journals and reference works

Responsibilities

Scientists

i. *Publicly funded scientists have a responsibility to contribute to the public good through the creation and communication of new knowledge, of which associated data are intrinsic parts. They should make such data openly available to others as soon as possible after their production in ways that permit them to be re-used and re-purposed.*

ii. *The data that provide evidence for published scientific claims should be made concurrently and publicly available in an intelligently open form[2]. This should permit the logic of the link between data and claim to be rigorously scrutinised and the validity of the data to be tested by replication of experiments or observations. To the extent possible, data should be deposited in well-managed and trusted repositories with low access barriers.*

[2] *Refer to the full text document at http://www.science-international.org*

We reiterate that 'low access barriers' may involve payment from the end-user. This should be at a rate that ensures sustainability of the repository, and that allows for appropriate levels of quality control of the deposited data and associated services.

We note also that 'low access barriers' include the technical facilitation of reuse

through open standards, well-documented APIs (application programming interfaces), and rich metadata describing the nature, use and relevance of each data set – that is to say, the 'intelligently open' form alluded to in this principle.

***iii. Research institutions and universities** have a responsibility to create a supportive environment for open data. This includes the provision of training in data management, preservation and analysis and of relevant technical support, including library and data management services. Institutions that employ scientists and bodies that fund them should develop incentives and criteria for career advancement for those involved in open data processes. Consensus on such criteria is necessary nationally, and ideally internationally, to facilitate desirable patterns of researcher mobility. In the current spirit of internationalisation, universities and other science institutions in developed countries should collaborate with their counterparts in developing countries to mobilise data-intensive capacities.*

***iv. Publishers** have a responsibility to make data available to reviewers during the review process, to require intelligently open access to the data concurrently with the publication which uses them, and to require the full referencing and citation of these data. Publishers also have a responsibility to make the scientific record available for subsequent analysis through the open provision of metadata and open access for text and data mining.*

For chemical crystallography, IUCr journals require all derived structural models and the processed experimental data sets underpinning them to be submitted for peer review (and subsequent publication). The journals provide an automated service, *checkCIF*, that rigorously tests the completeness and internal consistency of the submitted data [8]. This service is openly available to authors in advance of submission as well as to the reviewers. In some of the journals, the synoptic *checkCIF* report on a structure is also provided as a supplement to the published article. Furthermore, because the service is openly available, anyone may generate a post-publication validation report to assess the precision of the determined structure. *checkCIF* is also used by other publishers of chemical crystallographic data sets.

Experimental intensity data sets (structure factors) freely available from IUCr journals: > 58,400

'low access barriers' include the technical facilitation of reuse through open standards, well-documented APIs (application programming interfaces), and rich metadata

For macromolecular structures, a validation report is created by database curators when a structural data set is deposited. (Within this discipline, such deposition typically occurs in advance of submission of research articles.) IUCr journals require authors to provide the validation report upon submission. Processed experimental data are also deposited with the structural databases; increasingly reviewers request this (and the raw experimental data) from authors. This is a voluntary process, but there is evidence that the community increasingly considers it as a necessary practice.

***v. Funding agencies** should regard the costs of open data processes in a research project to be an intrinsic part of the cost of doing the research, and should provide adequate resources and policies for long-term sustainability of infrastructure and repositories. Assessment of research impact, particularly any involving citation metrics, should take due account of the contribution of data creators.*

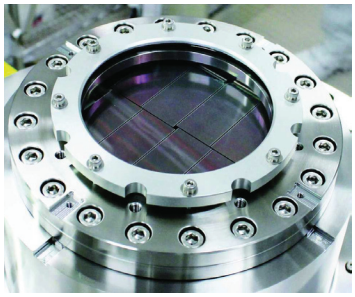
IUCr journals assign unique identifiers (DOIs) to all information supporting a publication, including derived and experimental data sets. This helps in providing citations for data. The IUCr has also launched a new service in 2016, *IUCrData*, which provides a fully citable form of short reports on crystallographic data.

***vi. Professional associations, scholarly societies and academies** should develop guidelines and policies for open data and promote the opportunities they offer in ways that reflect the epistemic norms and practices of their members.*

The IUCr does this consistently through its journal editorial guidelines, through the activities of advisory committees, working groups and Representatives on data and publishing organisations, through the community guidance and interactions of its Commissions, and since 2011 through the coordination and development activities of its Diffraction Data Deposition Working Group.

A data deluge

The camera head of a multi-port charge-coupled detector developed for serial femtosecond crystallography [Hatsui & Graafsma (2015), *IUCrJ* **2**, 371–383].



Time-resolved crystallography at high-flux radiation sources with a resolution of femtoseconds can generate hundreds of gigabytes of raw data per experiment.

***vii. Libraries, archives and repositories** have a responsibility for the development and provision of services and technical standards for data to ensure that data are available to those who wish to use them and that data are accessible over the long term.*

The IUCr develops metadata standards within its ontological framework that facilitate data characterization, archiving, validation and exchange. Although managed in a format that is almost unique to the discipline ('CIF'), care is taken to ensure ready interoperability with generic metadata standards in the library and repository worlds.

Boundaries of openness

***viii.** Open data should be the default position for publicly funded science. Exceptions should be limited to issues of privacy, safety, security and to commercial use in the public interest. Proposed exceptions should be justified on a case-by-case basis and not as blanket exclusions.*

All crystallographic data published by IUCr journals are openly available. Limited embargo practices are found in some areas, but the IUCr encourages full disclosure of all supporting data.

Enabling practices

***ix. Citation and provenance:** When, in scholarly publications, researchers use data created by others, those data should be cited with reference to their originator, to their provenance and to a permanent digital identifier.*

See comments under (v) regarding the assignment of permanent digital identifiers and opportunities for citation. The IUCr is a formal signatory to the Force11 principles on data citation [9].

Experimental powder profile data sets freely available from IUCr journals: > 1030

***x. Interoperability:** Both research data, and the metadata which allows them to be assessed and reused, should be interoperable to the greatest degree possible.*

This has been a principle of the IUCr since its inception in 1947, practised in the publication of derived and experimental data (in hard-copy form) since its journals were first published in 1948, and facilitated in the electronic age by the development of successive machine-readable standards, such as the Standard Crystallographic File Structure in 1981 [10], and the Crystallographic Information File in 1991 [11].

***xi. Non-restrictive reuse:** If research data are not already in the public domain, they should be labelled as reusable by means of a rights waiver or non-restrictive licence that makes it clear that the data may be re-used with no more arduous requirement than that of acknowledging the producer.*

***xii. Linkability:** Open data should, as often as possible, be linked with other data based on their content and context in order to maximise their semantic value.*

IUCr journals assign unique identifiers (DOIs) to all information supporting a publication, including derived and experimental data sets

Notes and References

[1] Significant community discussion moderated by the IUCr Diffraction Data Deposition Working Group is archived in an IUCr discussion forum http://forums.iucr.org/viewforum.php?f=21

[2] Terwilliger, T. C. (2014). Archiving raw crystallographic data. *Acta Cryst.* D**70**, 2500–2501.

[3] See the record of the two-day workshop on 'Metadata for raw data from X-ray diffraction and other structural techniques', http://www.iucr.org/resources/data/dddwg/rovinj-workshop

[4] Hall, S. R. and McMahon, B. (2016). The Implementation and Evolution of STAR/CIF Ontologies: Interoperability and Preservation of Structured Data. *Data Sci. J.* **15**, p. 3. DOI: http://doi.org/10.5334/dsj-2016-003

[5] http://www.iycr2014.org/openlabs

[6] Hall, S. R. and McMahon, B. eds. (2005). *International Tables for Crystallography, Volume G: Definition and exchange of crystallographic data.* First edition Dordrecht: Springer.

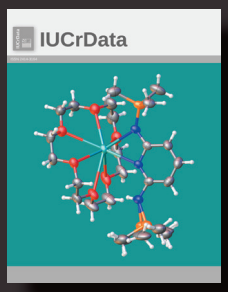
[7] http://www.iycr2014.org/into-the-future/conference/resolution

[8] Spek, A. L. (2009). Structure validation in chemical crystallography. *Acta Cryst.* D**65**, 148–155.

[9] https://www.force11.org/datacitation/endorsements

[10] Brown, I. D. (1988). Standard Crystallographic File Structure-87. *Acta Cryst.* A**44**, 232.

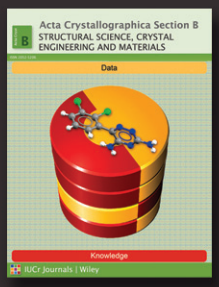
[11] Hall, S. R., Allen, F. H. & Brown, I. D. (1991). The Crystallographic Information File (CIF): a new standard archive file for crystallography. *Acta Cryst.* A**47**, 655–685.

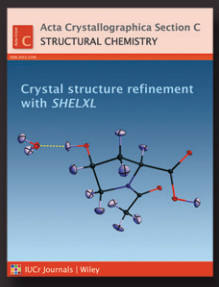


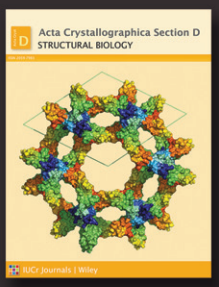
IUCrData is a peer-reviewed open-access data publication from the International Union of Crystallography (IUCr), first published in 2016. This innovative publication aims to provide short descriptions of crystallographic data sets and data sets from related scientific disciplines, as well as facilitating access to the data. The primary article category is *Data Reports*; these describe crystal structures of inorganic, metal-organic or organic compounds. Information on each crystal structure includes the crystallographic data (CIF and structure factors), a data validation report, figures and a text representation of the data.

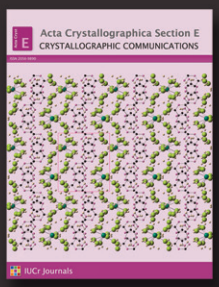
International Union
of Crystallography
2 Abbey Square
Chester CH1 2HU
UK
http://www.iucr.org

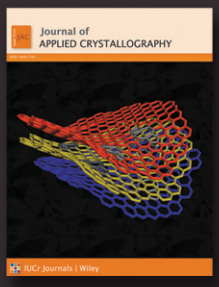
The IUCr is an International Scientific Union. Its objectives are to promote international cooperation in crystallography and to contribute to all aspects of crystallography, to promote international publication of crystallographic research, to facilitate standardization of methods, units, nomenclatures and symbols, and to form a focus for the relations of crystallography to other sciences.

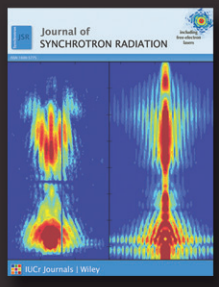












journals.iucr.org