Committee on Data - CommDat

THE IUCR COMMITTEE ON DATA (COMMDAT); ORIGINS, WHAT IT IS AND WHAT IT DOES

John R Helliwell

School of Chemistry, The University of Manchester, Brunswick Street, Manchester, M13 9PL, United Kingdom
john.helliwell@manchester.ac.uk

Keywords: Crystallographic data.
Synopsis: CommDat was formed in 2017 and is IUCr's standing committee for crystallographic data preservation and reuse for raw, processed and derived data. It sprang from the IUCr Diffraction Data Deposition Working Group (DDDWG), which delivered its final report in 2017 after 6 years work.

1. Introduction

The Committee on Data (CommDat) was established by the IUCr Executive Committee at its meeting in Denver, USA in July 2016 [1]. CommDat works with the IUCr's Commissions, including the publishing and standardisation Commissions (Journals, *International Tables,* Crystallographic Nomenclature), having a coordinating and advisory role regarding data. CommDat reports directly to the IUCr's Executive Committee. It also embraced the functions of the IUCr Diffraction Data Deposition Working Group (DDDWG), which ran from 2011 to 2017 tasked *to address growing calls within the crystallographic community for the deposition of primary diffraction images, with some mechanism that allowed their retrieval and reanalysis by other scientists for such purposes as structure redetermination, software and methods development, validation and review*. While the main focus of the Working Group remained on diffraction images, it reassessed what other categories of experimental data needed to be treated in a parallel fashion - such knowledge helping to orchestrate work flows, metadata standards and other mechanisms to improve the management of experimental data across all of crystallography. CommDat also has taken over the data-related interests of the now discontinued Committees on Crystallographic Databases and on Electronic, Publishing, Dissemination and Storage of Information. The Membership of CommDat and its consultants currently comprise 21 people who actively discuss crystallography data matters via a committee members' email list. CommDat exists alongside the Committee for the Maintenance of the CIF Standard (COMCIFS), which has the technical brief to manage the CIF standard, including active development of the CIF specification and dictionaries. CommDat, like the DDDWG before it, has an active Public Forum. CommDat's recent activities include participation in the International Data Week events and at CODATA meetings and workshops. It is a member of the IUCr Programme Committee for the World Congress in 2020. A key output of CommDat and COMCIFS that is being devised in a joint effort is a checkCIF for Raw Diffraction Data.

2. Terms of reference

CommDat's terms of reference are very broad:to advise the IUCr Executive Committee on all aspects of data with respect to policy and actions to be taken. These include the specific interests listed below, and may extend to any related issues that arise in the future.

- Raw data and its metadata preservation and their digital object identifiers
- Data mining within individual and across two or more databases
- Data and software development
- Data and instrumentation developers
- Data policy drivers as received from policy makers (*e.g.* funding agencies)
- Data type domains (discrete *versus* diffuse, i.e. continuum scattering)
- Data and eScience
- Data and data publishing [*IUCrData;* recommendation of editors for *IUCrData*; linking of data to articles in IUCr publications; new article categories involving data]
- Data repositories

3. Some history

The world of crystallography has a long history in connecting data with its publications. A view of this is shown in Figure 1.

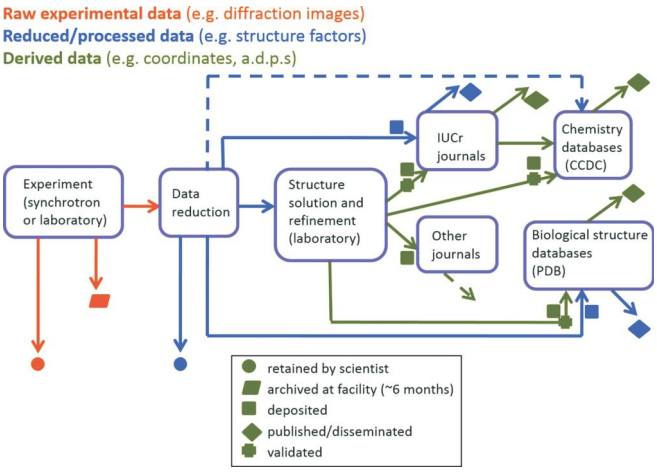

Figure 1. Crystallography's coherent approach to organising its data and publication flows: CIF ontologies at each stage of data processing as shown in Fig. 6 of reference [2].

The proposal for IUCr to have a Standing Committee on Data was made by JRH and accepted by the IUCr Executive Committee. CommDat sprang from the DDDWG having delivered its Final report and recommendations in 2017. Over a six year period the DDDWG addressed the hows, the whys and the whats of raw diffraction data archiving.

The DDDWG's Final report Recommendations [1] were, in brief:

- Authors should provide a permanent and prominent link from their article to the raw data sets which underpin their journal publication and associated database deposition, and the raw diffraction data sets should obey 'FAIR' principles (Findable, Accessible, Interoperable and Re-usable) (https://www.force11.org/group/fairgroup/fairprinciples).
- A registered Digital Object Identifier (DOI) should be the persistent identifier of choice.
- An archive of raw diffraction data sets for currently unsolved crystal structures should be pursued.
- An archive of raw diffraction data sets showing significant diffuse scattering should be pursued.
- Workshops for research data management training for the community should continue and be sponsored and organised by the IUCr.
- There should be continued regular checking by the IUCr Executive Committee of the progress of the IUCr Commissions logging of their raw diffraction data metadata.
- Archived raw diffraction data should be automatically validated wherever possible via a 'checkCIF for raw data approach', and be peer reviewed where necessary, at the minimum to include certain core metadata.
- Jointly with the IUCr Commission on Crystallographic Computing, the IUCr should pursue reproducibility of science objectives which require open source software and accurate versioning.
- IUCr should engage with vendors and the World Data System to promote the certification of raw diffraction data standards.
- IUCr's CommDat should continue the directory of data archives by adding any new data archives that are established in future. (Currently listed and described in [2].)
- IUCr should invite the community to alert CommDat of further case studies that document the value of archiving of raw diffraction data. (Current case study examples are included in [3]).
- IUCr recognises that metadata for the sample are clearly vital for all the IUCr Commissions (and are especially diverse in small angle scattering), and whose standardised descriptions should be actively pursued by the Commissions.
- CommDat should regularly monitor the evolution of technology.
- IUCr should actively support the neutron, synchrotron and X-ray laser facilities in their raw data archiving activities.

4. Publications

Publications directly arising from DDDWG Workshops are listed as references [2], [4-9].

4.3 Other activities of the DDDWG

The IUCr DDDWG Members John Helliwell and Brian McMahon with the IUCr President Marvin Hackert and the IUCr Secretary Treasurer Luc van Meervelt led the writing of a Response by IUCr (http://www.iucr.org/iucr/open-data) to the Science International 2015 Accord on *Open Data* in a *Big Data* World (https://council.science/publications/open-data-in-a-big-data-world). In the Response the IUCr acknowledged the importance of this Accord, and endorsed its analysis of the values of open data and its Principles of Open Data. The IUCr Response noted that the Accord is very general, with applicability across the entire panorama of science. Because the specific values, significance and implementation of Open Data principles will vary in detail between disciplines, the IUCr considered it useful to contribute a detailed response to the Accord as a case study of best practice emerging in one particular field. Specifically the IUCr holds 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 and processing routines are fully specified and quantified, statistical uncertainties are evaluated and declared).

4.4 Impact of the DDDWG Final Report

In moving towards implementation of the DDDWG recommendations CommDat is now actively working with several IUCr Commissions on their further consultations with their specific communities and/or their implementations of the DDDWG recommendations.

4. Public Fora

The DDDWG's Forum for Public Input was: http://forums.iucr.org/viewforum.php?f=21

CommDat's Forum for Public Input is:
http://forums.iucr.org/viewforum.php?f=39
These have both been very actively used.

5. Planned activities of CommDat

At ECM32 Vienna CommDat is hosting a Workshop on "Data Science Skills in Publishing: for authors, editors and referees"; details are here https://ecm2019.org/satellites/data-science-skills-in-publishing/. This is complemented by an Italian Crystallographic Association School (AICS) on Crystallographic Information in Naples http://cristallografia.org/contenuto/aics2019-crystallographic-information-fiesta/3311 which CommDat is assisting with. The AICS2019 offers "an intensive course in Crystallographic Information, covering the extraction, dissemination and use of scientific knowledge from the structure determination experiment to database-driven discovery".

At IUCr 2020 CommDat has proposed a Keynote by Dr Loes Kroon-Batenburg with the title: *Metadata and checkCIF for raw diffraction data in realising ultimate crystallographic objectivity* and three Microsymposia with titles of: A. *Pre publication peer review of crystallographic data*, B. *Post publication peer review of crystallographic data* and C. *Exemplary practice in chemical and biological database archiving.*

Acknowledgements

CommDat has received generous financial or in-kind support of several partners, who contribute to workshops and other activities of the Committee, namely IUCr Journals, CODATA, The Australian Nuclear Science and Technology Organisation (ANSTO), Bruker, the ICDD and WIley. CommDat is very grateful for this support.

References [1]

1. Helliwell, J.R., McMahon, B., Androulakis, S., Szebenyi, D. M., Kroon-Batenburg, L., Terwilliger, T., Westbrook, J. and Weckert, E. The Final report of the IUCr DDDWG (2017) http://forums.iucr.org/viewtopic.php?f=21&t=396
2. Kroon-Batenburg, L., 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. https://doi.org/10.1107/S2052252516018315 .
3. Helliwell, J. R., McMahon, B., Guss, J. M. & Kroon-Batenburg, L. M. J. (2017). The Science is in the Data *IUCrJ*, **4**, 714–722.
4. Terwilliger, T. C. (2014). Archiving raw crystallographic data. *Acta Cryst.* D**70**, 2500–2501.
5. 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.
6. Meyer, G. R., Aragao, 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). Operation of the Australian Store.Synchrotron for macromolecular crystallography. *Acta Cryst.* D**70**, 2510–2519.
7. Guss, J. M. & McMahon, B. (2014). How to make deposition of images a reality. *Acta Cryst.* D**70**, 2520–2532.
8. 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.
9. Bruno, I., Gražulis, S., Helliwell, J. R., Kabekkodu, S. N., McMahon, B. & Westbrook, J. (2017). Crystallography and Databases. *Data Sci. J.* **16**, p. 38.

1 For further information about the IUCr's scientific advisory committees see https://www.iucr.org/iucr/governance/advisory-committees .