

This document has been prepared by John R Helliwell (UK), Chairman of the IUCr Committee on Data and IUCr Representative to CODATA, in consultation with the CODATA International Data Policy Committee and its subgroup on Data Rights and Responsibilities. JRH thanks the organisers of the two CommDat Workshops referred to below for helping to polish the precise wordings used namely Prof Simon Coles (UK) and Dr Aaron Brewster (USA), respectively highlighted in yellow and turquoise. The text highlighted in grey explains some basic definitions about data types as used by crystallographers as well as the importance of the metadata concerning a crystal (or powder or solution) sample.

January 26th 2022.

Approved by the CODATA IDPC in the week ending Feb 11th 2022.

The question: “Just who is the owner of science research data?” has arisen

(See Appendix 1 below for more detail) at our two Workshops of the IUCr Committee on Data linked with IUCr Congress Prague. Namely is the data owner i.e. controller:- the experimenter, their PI, the PI’s employer, the facility such as a synchrotron used by the PI, or the PI’s funding agency?

The two workshops mentioned above featured different communities within crystallography; one involved the chemical crystallographers who measure 95% of their diffraction data at their home ie university facilities and the other workshop the macromolecular crystallographers who measure about 90% of their diffraction data at central facilities such as a synchrotron.

In terms of definitions:-

- (i) Raw data are those directly from a detector, whose calibrations have been applied, but before processing (a step also sometimes known as data reduction). The final, third step, is molecular model determination and refinement, which form the final derived data.
- (ii) A proposal will include sample safety metadata, and which are potentially very useful in expediting effective unpublished data reuse when released to the public, typically after three years. Nb an abstract will not include such details. A category of synchrotron usage by macromolecular crystallographer research teams is the BAG method, Block Allocation Group, and will involve multiple samples and possibly multiple ie different proteins. Nevertheless BAG synchrotron beamtime usage requires the research team to state the sample safety data.

Clearly open science is important and formally declared to be so by the recent UNESCO declaration (<https://www.unesco.org/en/articles/unesco-sets-ambitious-international-standards-open-science>), amongst others. However, ownership of science data, and its openness, varies around the World as evidenced by different behaviours of the above who variously believe they have control of data release, or even deletion.

We, IUCr, believe this situation deserves clarification and after consultations made with the CODATA Executive Secretary and then the CODATA General Assembly, it was decided that the CODATA International Data Policy Committee (IDPC, <https://codata.org/initiatives/data-policy/international-data-policy-committee/>) should be a good venue to explore this in detail. Not least the IDPC has a working group addressing *Data Rights and Responsibilities*. The Chair of the CODATA IDPC, Mark

Leggott added: *"This working group will continue to consider the aspects of raw data sharing in different domain contexts, as well as from the perspective of different actors, and will publish an output after their work is complete."*

Two meetings of the IDPC and the working group have ensued and we have received helpful feedback, which I have summarised as follows:-

1. *Published raw data with articles and database depositions* protocols are straightforward and basically should be fully open to the readership, which ensures reproducibility and subsequently replicability of science research findings.
2. *"Unpublished data retention, release or deletion"* decisions should comply with the principle of "satisfy the taxpayer and their proxies (funding agencies)" via what is promised in the research team's data management plan, and then approved by the funding agency. Where a central facility is involved in the experimental feasibility then a research team in accepting such facility beamtime in effect transfers the data management plan's remit to be instead that of the facility's overall data management plan. A simple and seemingly quite practical criterion to decide deletion of unpublished raw data by the facility would be the operational one of "in the absence of access by anyone of a raw dataset after a set term (say 10 years) that dataset can be deleted". If an individual research team is unwilling to accept a facility's data management plan, then it can choose not to use that facility. It is very likely of course that a facility will have a highly professional data handling and archiving staff. Alternatively, if the research team know how to manage the datasets at their home facilities, and they can make arrangements for transport or effect a network data transfer, then they can archive unpublished data on their own. Indeed, this may be their only option when using facilities where raw dataset retention periods (see Table 1 in <https://zenodo.org/record/5155882#.YefRTy-nyhA>) expire on timescales shorter than they can complete their analyses.
3. Effective data reuse of unpublished raw data depends on adequate (instrument and sample) metadata being made available. Ideally this should include a release of the beamtime proposal document at the same time as the release of the raw data; an abstract alone may well not be adequate for effective data reuse. However, ideally such a protocol should be an agreed one by a facility with its individual research teams. Whilst not allowing a single overall policy it would be better than only making available abstracts to accompany raw data sets.

4. Whilst the person measuring raw data may not have definitive rights of control, which most likely rest with their PI i.e. Team Leader, they do have rights to help define the details of the data management plan, which could thereby include best efforts to publish raw data details (a Raw Data Letter style of publication) independent of a full analysis publication.

New questions arose from the IDPC members to me:-

- i. What is the carbon footprint of now saving all these raw data?
- ii. What obligations are on industry to release raw data measured at publicly funded central facilities like a synchrotron? An IDPC Member offered the comment that *"I think that the use of a publicly funded facility may not be sufficient to create an obligation; more so if the research was in part or in whole funded by public funds."*
- iii. What fraction of raw diffraction data leads to publication? and thereby justify its carbon footprint. [JRH: Raw Data Letters will contribute to improving this fraction of course. Funding agencies are keen on release to the public of unpublished raw data after three years to increase this fraction and increase the return on their funds invested in research.]

Appendix 1 (and provided as extra briefing for the CODATA IDPC at the request of the Chairman)

At our Workshop on Archiving and publishing raw data in chemical crystallography one speaker who runs a university data collection and crystal structure analysis service said she *"could see situations where raw data would be interesting to publish but she didn't own the data, the PI did and he/she would not likely give permission to spend her time publishing a Raw Data Letter"*. When she asked us *"Who owns the data?"* we had to answer we didn't know for sure. Her expertise and knowing which data to measure, and how to set up her apparatus, surely gave her some level of rights though.

At our Workshop on Macromolecular Crystallography and High Throughput Data Collection one speaker from the USA Dept of Energy's Linac Coherent Light Source stated that one solution to *"too much data due to high throughput"* would be for the experimenter to delete those data sets that were only tests and in the fullness of time also delete those that did not lead to publication. I pointed out that in Europe the experimenter did not own the data, but the Facility did. So, such as the European Synchrotron Radiation Facility (ESRF) in Grenoble retain all measured data and plan to release measured datasets after three years, if not published by then. One experimental run would be registered with one doi. Thus, a dataset would include publishable and unpublishable samples' datasets. This, the retention of uninteresting datasets, can be perceived as wasteful of energy in a climate change World. Synchrotron Data Policies also vary. So, comparing two examples within Europe the ESRF and the French national synchrotron Soleil one will publish the proposal text, very important metadata, but the other only the abstract. So, who owns the metadata, at the level of an abstract or a two-page proposal? The latter maybe being essential to allow effective data Reuse. Suffice to say I asked the USA colleague to get a ruling from the Dept of Energy on whether as experimenter he had the right to delete a dataset. No feedback yet.