Crystallographic Information Fiesta

AICS School 2019

Professor John R. Helliwell, Chair of the IUCr Committee on Data (CommDat) and IUCr Representative to CODATA, was unable to attend the School in person, but sent a video message to highlight the importance of the topic of the School, both within crystallography and in the wider context of social responsibility of science. The following is a transcript of that message.

Quality structural data in modern scientific culture

John R. Helliwell

Transcript

The IUCr Standing Committee on Data [CommDat], of which I am Chairman, is a new committee, but the traditions within the IUCr with respect to data are long-standing.

[Here are some] very important milestones. 1991: the Crystallographic Information Framework [CIF]. That led to, in 1998, the opportunity to introduce the *checkCIF* [service], another very vital milestone. Now [in] 2011 President Sine Larsen of the IUCr commissioned a working group, the Diffraction Data Deposition Working Group; so you look at the possibilities for raw diffraction data archiving – the primary data. So at that point we are at the experiment. And this is a really important milestone because of the initiatives outside our domain with respect to archiving. The digital archives and [their] capacities have expanded enormously. So the working group was charged with looking into the possibilities of that and whether we needed to do that and effect the hows, the whys and the whats of raw data archiving; and we deliberated for six years. I presented a final report in Hyderabad [at the 2017 IUCr Congress] concluding that it was a great opportunity – a revolutionary opportunity – and the crystallographic community should really seize that opportunity.

All that work led to the realization that actually over the decades a number of different committees [that had been] involved with data [possibly could] be resurrected. But mainly because of the raw data opportunities we should have a standing committee on data – it was too important to not have a body to which all the individual Commissions could turn (if they wished) for assistance on these new opportunities.

There is also the vital importance of the databases: we're approaching 1 million crystal structures in the Cambridge Structural Database initiated by Bernal and Olga Kennard and [under] Olga Kennard's leadership over many decades. (I was in email contact with Olga Kennard just a month ago.) [There is also the] Protein Data Bank with 150,000 depositions through the leadership for many years of Helen Berman and now Stephen Burley.

So the bringing together of all of these strands within the quite simple sentence of the terms of reference (there are more details and bullet points at the website): "The Committee on Data (CommDat) will advise the IUCr Executive Committee on all aspects of data with respect to policy and actions to be taken.

If you think about data as a "zoo", what do we have? We have our raw measurements – top right we have an example of a raw diffraction data pattern; and then we process that data and so in the middle there we have our processed structure factors; and finally we derive our structural information. So I'm deliberately using keywords here – raw, processed and derived – because they are then generic terms which other scientific disciplines also relate to. So in representing the IUCr at the International Council for Science Committee on Data [CODATA] where I meet particle physicists, astronomers and so on – other community areas which are leading with respect to data and the trust in their data – these are terminologies which we all understand. And of course, as a part of this, each one of these domains, the raw, the processed and the derived, requires careful metadata recording, because the metadata (data about our data) is absolutely vital if we are to proceed to use it productively and also to reuse it because at the heart of reuse of data is reproducibility and we're aware (we read in *Nature* and so on) of worries of irreproducibility in areas like psychology and so on.

These are really *major* concerns of society. And a big change is the merger of the International Council of Scientific Unions and the International Council of Social Sciences which took place last year to form the International Science Council.

Now the "Big Data" pyramid is shown here and it illustrates with our coordinates (the derived data) our files were kilobytes and still are kilobytes; our processed intensities are megabytes; and we've always taken the opportunities that the technology gives us to preserve what we can. So we started by preserving coordinates and then we extended that to intensities and at the base of this pyramid are our primary diffraction data – the raw data which are now gigabytes or even terabytes. The heart of what we're doing as scientists: we move closer and closer to objectivity. Objectivity itself is possible through the raw data. And that is a very vital transformation which I think is a revolution even for our science where we put great store on data.

As scientists it's the element of *trust* that is so important. Anything that undermines trust undermines the confidence of society and the funding agencies in funding us. So we have achieved trust in data transmission and exchange

through the Crystallographic Information File; trust in data consistency through *checkCIF*, firstly for coordinates and then extended to include structure factors. And the data provenance with the opportunities in the discussions within the Diffraction Data Deposition Working Group in those six years. And this simple title, "The Science is in the Data" which was the title that Gautam Desiraju, the most recent president before Sven Lidin, gave me for my keynote, and which led to this article at the bottom of the slide here, published in *IUCrJ*.

The traditions of IUCr have been recognized through awards. So I led the nomination for IUCr to receive the Association of Learned and Professional Society Publishers' award in 2006, and here you see a picture of Brian McMahon, an expert within the IUCr headquarters in Chester in data R&D and in the Crystallographic Information File. 2014 was the International Year of Crystallography. As the Representative to CODATA it was obvious to me that with a good nomination we could seek the award of the CODATA Prize. So with a worldwide set of letters of support the nomination was successful for Syd Hall from Perth [inventor of CIF] to receive the CODATA Prize 2014. "The implementation and evolution of STAR/CIF Ontologies: interoperability and preservation of structured data" – that's the article that came from Syd's Prize and here is Syd receiving the award in New Delhi at the biennial CODATA Congress.

Crystallography and crystallographers are recognized by all scientific communities for the work that we do and the work and trust that people have in our data. But it's not only the scientists; this European Union report [from] 2018, "Turning FAIR into reality", that data should be findable, accessible, interoperable and reusable, recognized crystallography – the International Union of Crystallography – for the many data standards that were maintained successfully by the Union and by the community and by the Commissions – by *you*.

So for you the take-home message is that the IUCr maintains the need for the highest quality of data management at all stages between the experimental data collection, through reduction and analysis, to publication and database deposition.

It is important.

Reading material

Helliwell, J. R., McMahon, B., Guss, J. M. & Kroon-Batenburg, L. M. J. (2017). The science is in the data. *IUCrJ* **4**, 714–722.

Helliwell, J. R., Minor, W., Weiss, M. S., Garman, E. F., Read, R. J., Newman, J., van Raaij, M. J., Hajdu, J. & Baker, E. N. (2019). Findable Accessible Interoperable Re-usable (FAIR) diffraction data are coming to protein crystallography. *IUCrJ* **6**, 341–343.

Online resources

Helliwell, J.R., McMahon, B., Androulakis, S., Szebenyi, M., Kroon-Batenburg, L., Terwilliger, T., Westbrook, J. & Weckert, E. (2017). Final report of the IUCr Diffraction Data Deposition Working Group. **https://www.iucr.org/resources/data/dddwg/final-report**

IUCr CommDat website. **https://www.iucr.org/resources/data/commdat**

Public input to CommDat. Discussion forum **https://forums.iucr.org/viewforum.php?f=39**