A Gold Standard for the archiving of macromolecular diffraction data

Herbert J. Bernstein[1], Andreas Förster[2], Aaron S. Brewster[3], Graeme Winter[4]

[1]Ronin Institute for Independent Scholarship, c/o NSLS II, Brookhaven National Laboratory, Upton, NY, USA;
[2]DECTRIS Ltd., Täfernweg 1,5405 Baden-Dättwil, CH;
[3]Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 94720, USA;
[4]Diamond Light Source Ltd, Harwell Science and Innovation Campus, Didcot OX11 0DE, UK;

hbernstein@bnl.gov

Macromolecular crystallography (MX) is the dominant means of determining the three-dimensional structures of biological macromolecules. Over the last few decades, most MX data have been collected at synchrotron beamlines using a large number of different detectors produced by various manufacturers and taking advantage of various protocols and goniometry. These data came in their own formats, some proprietary, some open. The associated metadata rarely reached the degree of completeness required for data management according to Findability, Accessibility, Interoperability and Reusability (FAIR) principles. Efforts to reuse old data by other investigators or even by the original investigators some time later were often frustrated.

In the culmination of an effort dating back more than two decades, a large portion of the research community concerned with High Data-Rate Macromolecular Crystallography (HDRMX) agreed in 2020 to an updated specification of data and metadata for diffraction images produced at synchrotron light sources and X-ray free electron lasers (XFELs) [1]. This Gold Standard builds on the NeXus/HDF5 NXmx application definition and the International Union of Crystallography (IUCr) imgCIF/CBF dictionary and is compatible with major data processing programs and pipelines. It will ensure effortless automatic data processing, facilitate manual reprocessing of data independent of the facility at which they were collected, and enable data archiving according to FAIR principles, with a particular focus on interoperability and reusability.

Direct consequences of the Gold Standard are an unambiguous definition of the experimental geometry, a record of the synchrotron and beamline where the data were collected, and additional optional metadata that will make subsequent submission of the structural model to the PDB more straightforward. Just as with the IUCr CBF/imgCIF standard from which it arose and to which it is tied, the Gold Standard is intended to be applicable to all detectors used for crystallography. In particular, the application of the Gold Standard does not require the use of HDF5. Corresponding metadata definitions exist in CBF/imgCIF. All hardware and software developers in the field are encouraged to adopt and contribute to the standard.

The Gold Standard provides a convenient and consistent way to record the essential minimal data and metadata needed to process a wide range of macromolecular diffraction experiments including single axis, single crystal rotation experiments using single-module detectors, XFEL serial crystallography experiments using powerful multi-module detectors producing tens of thousands of images from huge numbers of small crystals, as well as synchrotron experiments producing large number of wedges from micro-crystals. Examples from all of these and more will be discussed.

[1] Bernstein, H.J., Förster, A., Bhowmick, A., Brewster, A.S., Brockhauser, S., Gelisio, L., Hall, D.R., Leonarski, F., Mariani, V., Santoni, G., Vonrhein, C. and Winter, G. (2020). *Gold Standard for macromolecular crystallography diffraction data*. IUCrJ, 7(5) 784 -- 792.

The work was supported in part by funding from Dectris Ltd., from the U. S. Department of Energy (BES KP1605010, KP1607011, DE-SC0012704), from the U. S. National Institutes of Health (NIGMS P30GM133893, R01GM117126).

Keywords: FAIR, MX data, Gold Standard, Archiving