Extending Cristallographic Data Management to Include the Routine Archiving of Raw Experimental Data

John R. Helliwell, School of Chemistry, Faculty of Engineering and Physical Sciences, University of Manchester
Brian Mcmahon, International Union of Crystallography

Abstract:

The determination of crystal structures and their publication have been associated for many decades with effective research data management practices. Structural models are collected in curated high-quality databases, such as the Cambridge Structural Database for organic structures and the Protein Data Bank (PDB) for biological macromolecules. The journals published by the IUCr validate and check submitted structural models. The processed experimental data ('structure factors') from which these models are determined are normally deposited with the PDB in the case of protein and nucleic acid structures; for organic or inorganic structures published in IUCr journals, the author must deposit the structure factors with the journal, which makes them freely available as supporting information. Currently the raw experimental data (most often in the form of X-ray diffraction images) are retained by the researcher or by large facilities such as synchrotron laboratories where the experiments are performed. There are no community policies for long-term preservation of the primary data.

However, there is growing interest in retaining such data sets. Significant scientific reasons for doing so include the following.

(1) Raw diffraction images include additional information (diffuse scattering between the discrete diffraction spots used in structure analysis) that could yield new science upon detailed reanalysis.

(2) Occasional incorrect assignment of space groups may be made during the early stages of data reduction. Retaining the raw data allows for a complete reanalysis if necessary.

(3) Raw data may show evidence of multiple crystal lattices that are discarded in the standard determination of the dominant crystal structure.

(4) Sometimes the structure cannot be determined. Retention of the experimental data could allow other research groups to attempt a determination; or a new attempt may be made when new methodologies are developed.

(5) Currently data beyond an arbitrary diffraction resolution limit are routinely discarded. Retention of the raw images would permit future re-evaluation that takes account of data beyond the published diffraction resolution.

(6) Retention of raw data for public scrutiny can help to safeguard against scientific fraud.

The International Union of Crystallography has commissioned a working group to investigate the feasibility, costs and technical requirements for routine deposition and archiving of such raw data. The group includes representatives of the PaN (Photon and Neutron Data Infrastructure) initiative (http://pan-data.eu), synchrotron laboratories and neutron sources, structural databases, university researchers and the IUCr itself. It is interested in best practice among experimental facilities and the role of institutional repositories in hosting their researchers' data sets.