The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC)

W. Minor
University of Virginia

The Integrated Resource for Reproducibility in Macromolecular Crystallography (IRRMC) has been developed as part of the BD2K (Big Data to Knowledge) NIH project to archive raw data from diffraction experiments and, more importantly, to extract metadata from diffraction images alone, or from a combination of information obtained from a PDB deposit and diffraction images. As of February 2017, the IRRMC resource contained indexed data from 3235 macromolecular diffraction experiments (6189 data sets), accounting for around 3% of all structures in the Protein Data Bank (PDB). The IRRMC utilizes a distributed storage system implemented with a federated architecture of many independent storage servers, which provides both scalability and sustainability. The resource, which is accessible via the web portal at **https://www.proteindiffraction.org**, can be searched using various criteria. All data are available for unrestricted access and download. The resource serves as a proof of concept and demonstrates the feasibility of archiving raw diffraction data and associated metadata from X-ray crystallographic studies of biological macromolecules. The goal is to expand this resource to include data sets that have failed to yield X-ray structures in order to facilitate collaborative efforts that will improve protein structure-determination methods and to ensure the availability of 'orphan' data left behind for various reasons by individual investigators and/or extinct structural genomics projects. Every dataset in the IRRMC resource is assigned a DOI (Digital Object Identifier), which should provide a reliable mechanism of data location, even if the URL or the maintainer of the data changes.