Planning For the Future of Crystallographic Raw Data and Its Metadata

Dr. Herbert J. Bernstein[1], Frances C Bernstein[1]
[1]*Bernstein + Sons*
yayahjb@gmail.com

Crystallography is alive and well and thriving. The macromolecular side has been invigorated by a flood of AI-generated models from AlphaFold and RoseTTAFold making phasing with molecular replacement increasingly effective for high resolution studies of moderate-sized macromolecules. Xfel and synchrotron serial crystallography now produce data volumes that threaten to exceed funding for storage and need massive processing resources. In this world we must forget human-readability of our data and focus on compact efficient formats. Existing CBF-based, HDF5-based, and other raw data formats work and follow this advice to be efficient and are increasingly interoperable, but they are poorly understood black boxes for many users. How are we to channel this data flood into something which can be managed, understood and used by crystallographers, biologists and others? The answer we propose is essentially the same as the one the world has adopted for word processing. Our documents are stored in formats no sane person reads easily, such as Postscript and PDF, but clever software and AI guides us in working with the words and their metadata easily. Let us do the same with our crystallographic data and metadata by building the tools we need for browsing, searching and manipulating crystallographic data in CBF, HDF5, TIFF and other available formats without ever having to deal with the innards of these formats. This is a huge task and will need the help of many of our colleagues and may well go on forever, but it is clearly within the scope of available tools.

Acta Cryst. (2023). A79, a100