

Research Data Management at CHESS

D. Marian Szebenyi*, Devin Bougie, Aaron Finke, Richard Gillilan, Jesse Hopkins, David Schuller, and Werner Sun

MacCHESS and CHESS, Cornell University, Ithaca, NY14853, USA.

*Corresponding author: MacCHESS, Cornell University, Ithaca, NY14853, USA. Tel: 1-607-255-5717, Fax: 1-607-255-9001, Email: dms35@cornell.edu

Historically, the Cornell High Energy Synchrotron Source, CHESS, with its relatively small number of beamlines, has relied on users to manage their own data. The facility has provided adequate RAID storage at each station for a 6-8 week run, with some longer term backup. The advent of increasing numbers of experiments involving massive amounts of data has strained this system. Accordingly, we have recently implemented a large, centralized, more organized, system (“CHESS DAQ”), with separate storage for raw data, metadata, and general user data. Nightly incremental backups and full archiving at the end of each run protect against data loss. This system is used for most experiments at CHESS, with individual variations to suit the needs of users and staff. Our primary goal has been, and remains, to facilitate research by our users, by providing them the means to collect, process, and store the most useful data possible, while avoiding excessive bureaucracy.

BioSAXS raw data from SAXS and WAXS detectors (the two detectors record images simultaneously), as well as metadata, are written directly to CHESS DAQ. Processing is carried out locally on copies of the raw data, and processed data are backed up locally as well as on the DAQ and to user-supplied media.

Raw crystallographic data from the dedicated MX station, i.e. diffraction images, are stored locally, with users responsible for processing data and transferring raw and processed data to their home labs. Raw data are kept on-line for a few weeks and off-line for several years. Limited metadata are stored in image headers. Implementation of a new database, to facilitate organization of raw data and metadata, is under development in parallel with adoption of a new user interface for data collection, based on JBlIce.