MS053.O06

Crystallographic metadata from a front end perspective

Kamil Filip Dziubek[1]
[1]*LENS - European Laboratory For Non-Linear Spectroscopy, Sesto Fiorentino, Italy*
E-mail: dziubek@lens.unifi.it

One of the main goals of the newly created IUCr Committee on Data (CommDat) is the raw data and its metadata preservation. Indeed, the metadata should be considered as an indispensable and integral part of the raw collected data. Only by providing the proper context can the data be interpreted, reused, validated and analyzed. In 2015 at Rovinj, Croatia, the IUCr Diffraction Data Deposition Working Group (now subsumed by the CommDat) organized the seminal workshop addressing the metadata for raw data issues [1]. The participants unanimously agreed that among the most burning matters, there is a necessity to define minimum requirements for metadata, concerning not only the diversity of instrumentation and image formats, but also the specific items related to a range of experimental techniques in crystallography. This work is carried out through the IUCr Commissions, which were asked to define their particular metadata standards.

While in the field of macromolecular crystallography complete automation of the diffraction experiment and remote data acquisition is becoming a standard rather than just a trend, in many other areas workflow and the associated dataflow still needs to be organized in a way facilitating metadata harvesting and preservation. Creation of structured templates to aid information collection can improve the completeness and consistency of metadata ensuring that all the necessary items are preserved [2]. Such templates are particularly efficient complying with specific guidelines and protocols for scientific conduct known as good laboratory practice (GLP). Although in principle experimentalists should be aware that recording metadata along with raw data is essential for further meaningful retrieval of archived content, making the assumption that everyone would diligently follow all the rules falls into the category of wishful thinking. For this reason, two major challenges remain to be addressed: automated raw data and metadata validation tool, tentatively dubbed as 'checkCIF for raw data' [1] and developing a data and metadata management system considering specific technique dependent requirements. This is the next step after implementation of metadata standards, requiring a close collaboration with software managers from instrument vendors and large scale facilities. Such detailed GLP procedures have been already reported by some beamline scientists [3], demonstrating the need for creating general data management protocols. Moreover, while the validation tools are employed for an objective a posteriori assessment of raw data and metadata sets, the laboratory guidelines and structured templates help to keep track and prevent loss of critical metadata in course of the experiment. It is particularly important if the dataflow is not fully automated and the human factor comes into play, e.g. the synchrotron users stressed or tired from working around the clock.

[1] Kroon-Batenburg, L. M. J. et al. (2017). IUCrJ, 4, 87–99.
[2] Willoughby, C. et al. (2016). J. Cheminform. 8, 9.
[3] Zhang, D. et al. (2017). J. Vis. Exp. 119, e54660.

Keywords: metadata, structured template, good laboratory practice