Data Management and Data Validation Best Practices from The Perspective of A Data Repository

Dr Matt P Lightfoot[1], Dr Ian J Bruno[1], Dr Natalie T Johnson[3], Yinka Olatunji-Ojo[1], Suzanna C Ward[1]

[1]Cambridge Crystallographic Data Centre
lightfoot@ccdc.cam.ac.uk

This talk will explore the importance of good data validation and management from the perspective of the Cambridge Structural Database (CSD), a trusted worldwide data repository.

The CSD is a collection of over 1.2 million experimental three-dimensional structures obtained through crystallographic analyses. These structures are determined by crystallographers worldwide and undergo curation and enhancement by scientists at the Cambridge Crystallographic Data Centre (CCDC) prior to their addition to the database.

In this talk we will cover our current best practices for how we manage data at the CCDC including our current data preservation policy, our data repository certification, how we add and validate the metadata associated with deposited structures and how we link to raw data and data in third-party repositories.

We will also cover our current work on ensuring that we have the correct data integrity checks and workflows in place, our work on improving the availability of our data and our ongoing work to ensure that our data is in a human readable form as well as being fit for machine learning and AI.

Finally, we would also like to use this opportunity to ask the ACA community what matters most to them as we look to evolve our data management processes in the future.