Metadata for better data - Growing and improving the Cambridge Structural Database

Natalie Johnson, Seth B Wiggin, Suzanna C Ward, Ian Bruno

Cambridge Crystallographic Data Centre, Cambridge, United Kingdom.

njohnson@ccdc.cam.ac.uk

The Cambridge Structural Database (CSD)[1] is a database of over 1.1 million small molecule organic and metal-organic crystal structures. Each structure added to the database is curated to ensure important details about the structure are recorded alongside the entry. This curation process is particularly important for structures submitted directly to the CSD as a *CSD Communication*, with no accompanying journal article. As the CSD continues to grow and new techniques emerge it is essential that information is recorded consistently to ensure the data is findable. Consistency also allows the CSD to be utilised in data-driven approaches, such as machine learning, reducing the need for curation before it is ingested into models.

In addition to processing new data each year, the CCDC undertakes a series of improvement projects to assess the data stored in the CSD, ensure it is consistent and correct any errors. Ongoing projects also aim to capture additional information about the structure, such as if a specialist refinement technique is used. In addition, the CCDC is working towards updating the underlying format of the database, allowing new metadata about the structure to be stored. This poster will present highlights from work to continue to improve and grow the CSD.

[1] Groom, C., Bruno, I., Lightfoot, M., & Ward, S. (2016). *Acta Cryst. B* **72**, 171-179.

Keywords: metadata; database curation