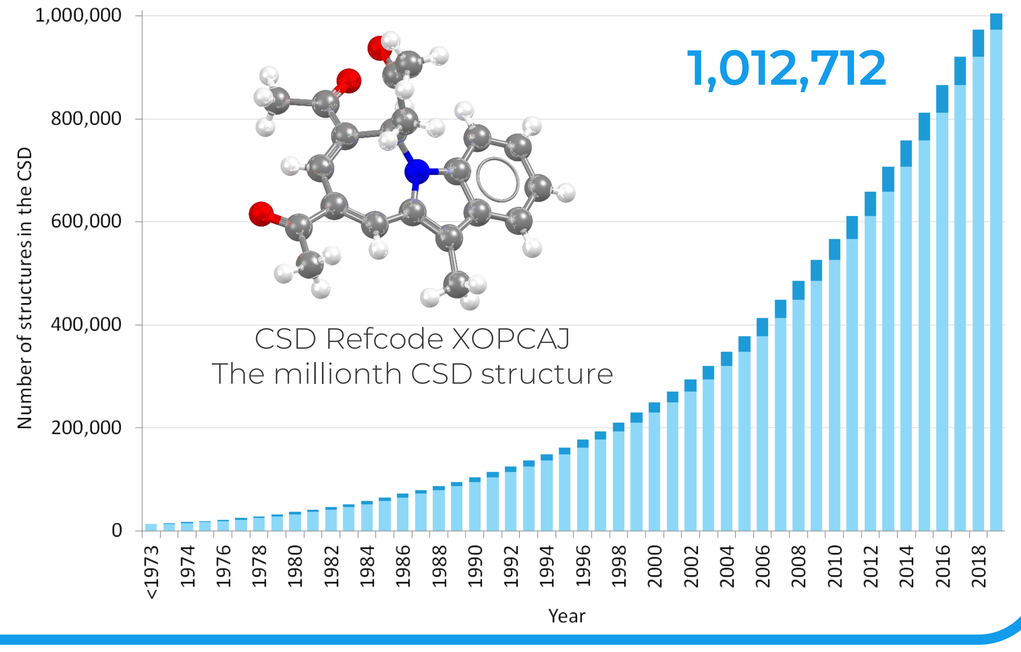
Using the CSD to increase data science skills in the publication of structural data

Suzanna C. Ward,* Natalie T. Johnson,* Amy Sarjeant

*The Cambridge Crystallographic Data Centre (CCDC) Email: ward@ccdc.cam.ac.uk, njohnson@ccdc.cam.ac.uk

Introduction

The knowledge derived from structural databases is used worldwide in research, education and structure validation. One key area where databases can help the community set standards is through interactive deposition processes. At the CCDC we have made changes to our online deposition service to facilitate this and in partnership with FIZ Karlsruhe have extended this service to cover data in both the Cambridge Structural Database (CSD) and the Inorganic Crystal Structure Database (ICSD). During deposition the CCDC can help researchers adhere to best practices by requiring data to be in standard file formats and by mandating or at least encouraging the inclusion of associated information such as structure factors. Deposition processes may also include checks and measures to help crystallographers identify and fix issues prior to publication. To help facilitate validation and data provenance the CCDC has started capturing raw data DOIs and crystallographer details. This poster demonstrates a number of ways in which a structural database, namely the CSD, can work with the community to comply with FAIR data principles and to help set standards from validation to publication.



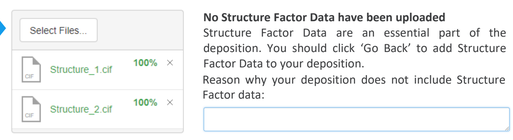
Deposition Standards

1 Login → 2 Upload → 3 Check Syntax → 4 Validation → 5 Add Publication → 6 Enhance Data → 7 Review → 8 Submit

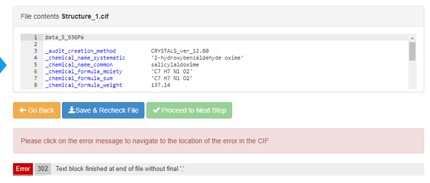
Joint service

CCDC FIZ Karlsruhe
Leibniz Institute for Information Infrastructure

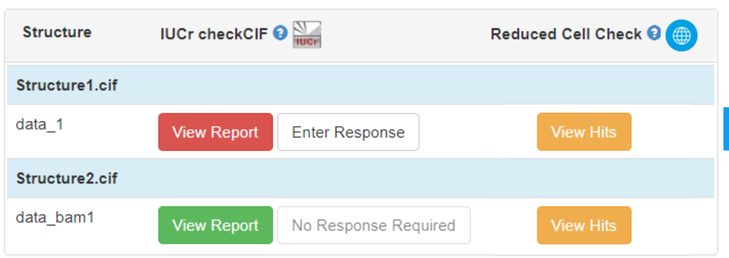
Standard file format

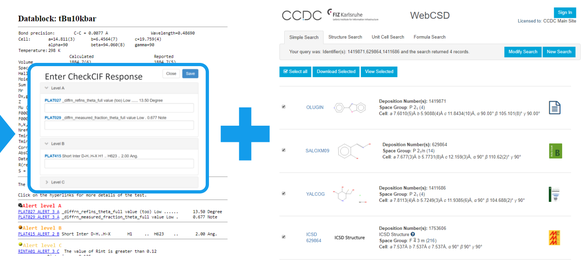


Syntax checking

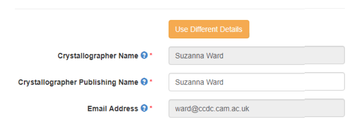


Validation





Attribution

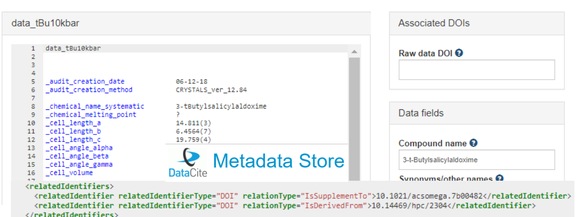


CCDC

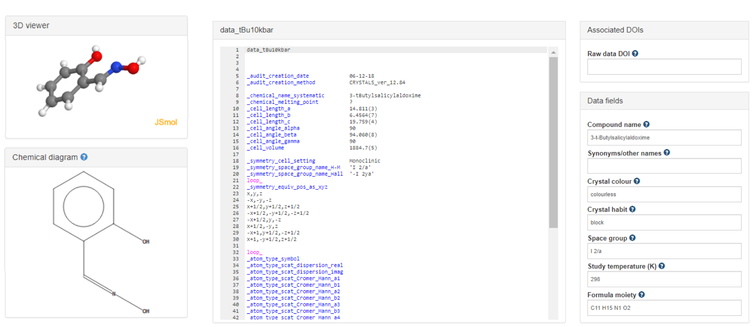
Crystallographer details added on deposition are visible on the CCDC Referee Service, Access Structures and WebCSD

Crystallographer	
Name	Suzanna Ward
Affiliation	CCDC, 12 Union Road, Cambridge, UK

Access to raw data



Enhancing data



Deposition guidelines

The CCDC CIF Deposition Guidelines

When preparing your CIF for deposition please include as much information as possible and check it carefully. This is especially true for *CSD Communications* where there is no paper to describe the chemistry and experimental details leading to your structure. If you choose to publish your data as a *CSD Communication* please remember to provide all the authors/crystallographers/chemists who contributed to the crystallographic experiment as authors of the data. If we are unable to validate your structure from the information you have provided we may contact you. If we cannot resolve the issue, unfortunately, we may not be able to add your structure to the CSD.

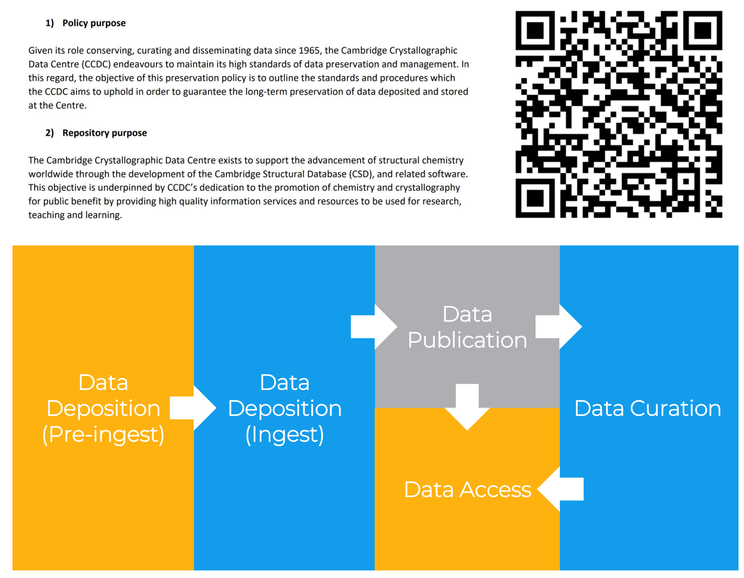
Guidelines in Chinese

All experimental CIF files (including those from powder diffraction experiments) should contain an R-factor. This should be consistent with the crystallography being performed correctly and to the best ability that would be expected from the material and equipment used. We prefer single block CIF files and would like all experimental CIFs to contain:

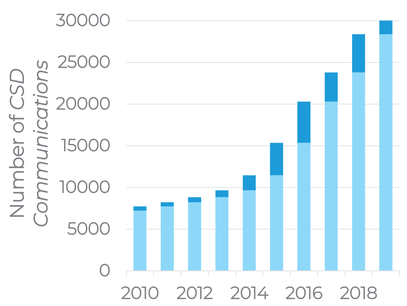
- R-factors (R1, wR2, Rint)
- GooF
- Shift/ESD (to show that the refinement has converged)
- Explanation of any problems with numbers of reflections and parameters
- Any residual electron density
- Details of squeeze/solvent masking
- Atomic Displacement Parameter (ADP) values
- Temperature – cell and data collection temperatures match

Data Sharing

CCDC Data Preservation Policy



CSD Communications



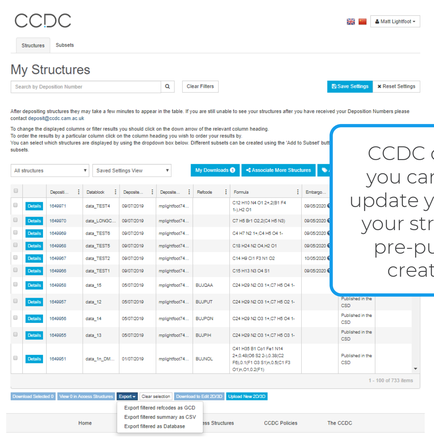
FAIR Data

The CCDC works to follow the FAIR guiding principles when sharing and preserving data

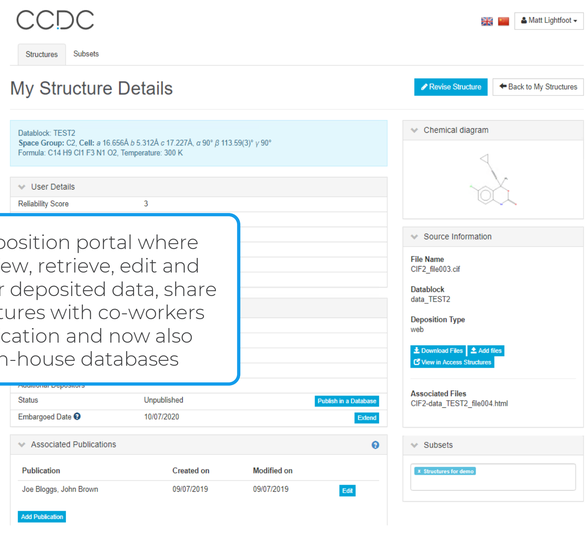
DATA SHOULD BE
Findable | Interoperable
Accessible | Reusable
BY HUMANS AND MACHINES

Wilkinson, M. D. *et al.*, The FAIR Guiding Principles for scientific data management and stewardship, *Sci. Data*, 2016, **3**, DOI: 10.1038/sdata.2016.18

My Structures



CCDC deposition portal where you can view, retrieve, edit and update your deposited data, share your structures with co-workers pre-publication and now also create in-house databases



Supporting the Community

This poster shows how standards can be enforced and encouraged through deposition to a domain-specific repository. While data repositories like the CSD can help to encourage good data sharing practices, researchers have to take responsibility to apply these practices to all their scientific data.
A highly curated database can be used to check the integrity of new data; at the CCDC are taking steps towards ensuring the integrity of crystallographic data in the CSD. We are working on incorporating checks on data completeness, identifying cases of misconduct, as well as methods to allow users to apply a wider range of quality filters to select data fit for their research needs. We would like to hear your thoughts on how we should continue to support the community from data collection to publication.

Colin R. Groom, *et al.*, The Cambridge Structural Database, *Acta Cryst. B*, 2016, **72**, 171, DOI: 10.1107/S2052520616003954