Crystal structure data without a publication – how much more can be learnt?

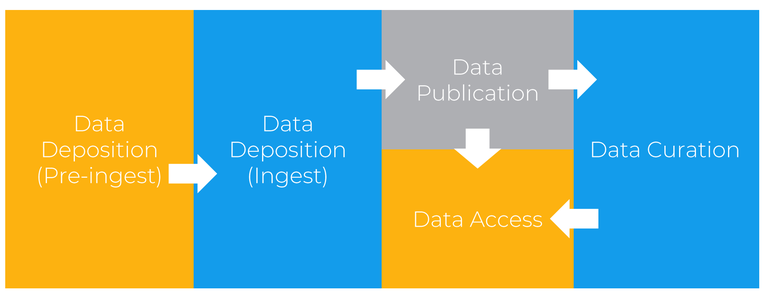
The Database Team, The Cambridge Crystallographic Data Centre (CCDC)

Email: hello@ccdc.cam.ac.uk Website: www.ccdc.cam.ac.uk Twitter: ccdc_cambridge Facebook: ccdc.cambridge YouTube: CCDCCambridge

Poster updated: March 2020

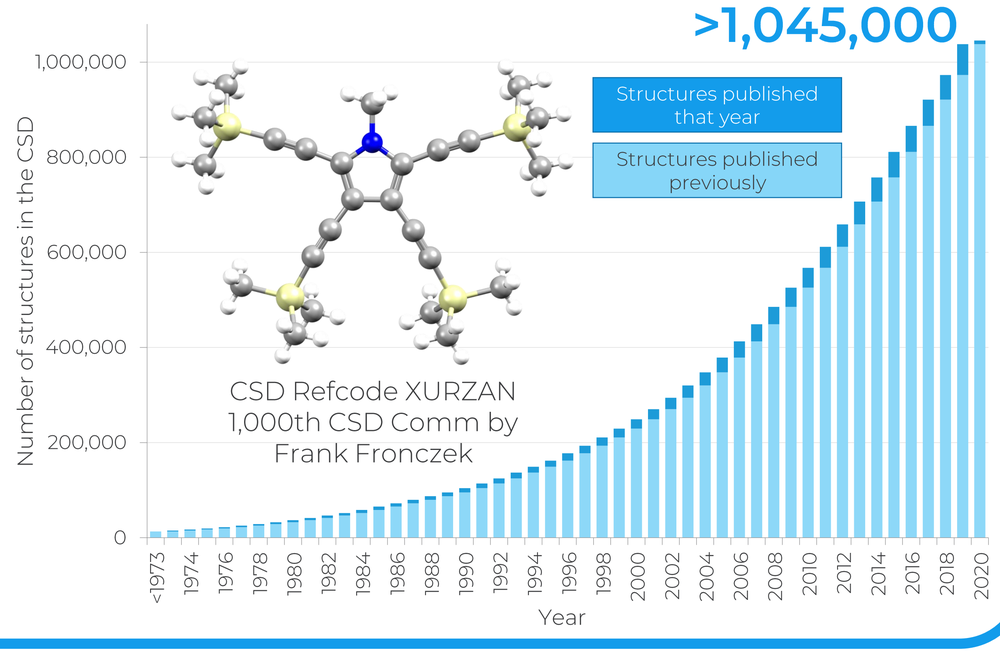
The CSD

The Cambridge Structural Database (CSD) is a curated database of over one million small molecule organic and metal-organic experimental crystal structures. Data is curated in house, validated and stored in a standardised format so that the structures are searchable, reusable and easily findable. The workflow shows the process of adding data to the database, from deposition to the CCDC, to data curation by one of the CCDC Editors once data has been published. The individual structures, as well as the knowledge derived from the collection, are used worldwide in research and education.



Data management processes for datasets deposited at the CCDC

One key area where databases can help the community is through guided deposition processes to ensure the data is of high quality. To assist with this, the CCDC has developed deposition guidelines to advise scientists of appropriate information to share with their data and through a web deposit system where the depositor can provide additional information during the process.



CSD Communications

ISSN 2631-9888

Scientists can share their crystallographic data through the CSD without the need for an accompanying journal article. These structures, called *CSD Communications*, undergo the same curation and validation processes as a structure associated with a traditional paper. Over the last few years we have taken a number of different approaches to increase the number of structures shared in this way.

Recognition and value

Scientists are frequently assessed on publications, impact factors and citations but these don't give a complete picture of a researchers scientific output. To this end *CSD Communications* now have a unique ISSN number and each dataset receives a full data citation including a Data DOI. This enables the data to be cited, to be added to researcher and university profiles and to be discovered and re-used by others.

Wordle showing authors with over 100 CSD Comms

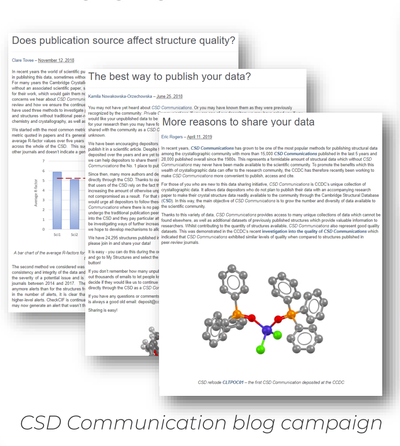
Frank Fronczek receiving a prize after becoming the first author of >1,000 CSD Comms

Logos showing how we use DataCite to assign DOIs, our integration with ORCID IDs, that the CCDC has been certified by CoreTrustSeal and CSD Communications have an ISSN Number

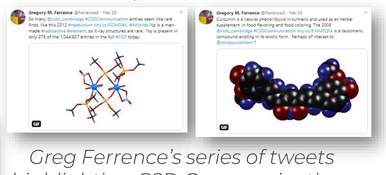
Community engagement

Strategies to increase engagement include:

- CCDC led promotional campaigns
- Engaging members of our community in promotion



CSD Communication blog campaign



Greg Ferrence's series of tweets highlighting CSD Communications

Bill Clegg talking about CSD Communications at ECM 2019

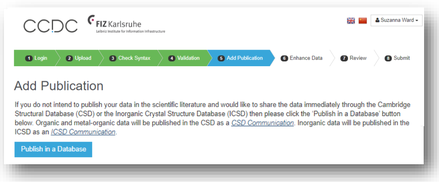
A competitive edge

Everyone loves a bit of friendly competition! To harness this we have introduced leaderboards at conferences. Each year the competition changes but there are always bonus points for *CSD Communications*! This year we are also introducing a monthly contributor table. Although a bit of fun it does help to remind the community about the initiative.

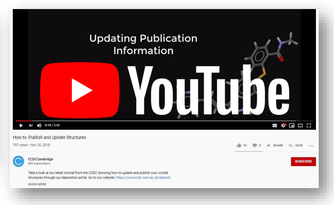
ECM 2019 conference leaderboard

Making it easy

Depositors can publish data as a *CSD Communication* during deposition or through our MyStructures portal. A YouTube video shows researchers how and guidelines are on our website. Each year we also send out reminder emails about unpublished data to all our depositors. After selecting to publish as a *CSD Communication* data is immediately shared through our free Access Structures service and then curated into the CSD.



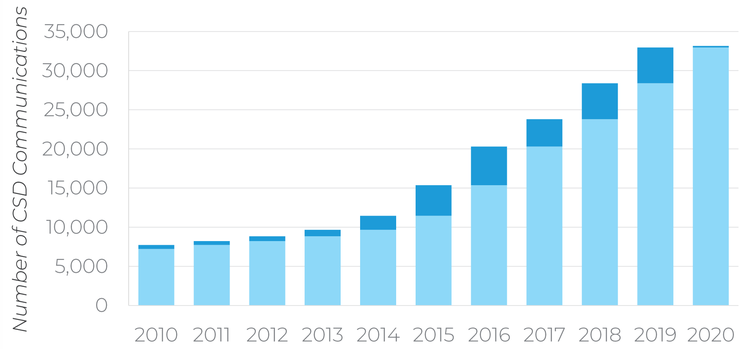
Depositing CSD Communications



Instructional YouTube videos

The results

Last year *CSD Communications* were the number one place to publish structural data with over 5,000 structures and a 10% increase on 2018. Efforts by a number of champions in the community have clearly had a huge impact.



Growth of CSD Communications since 2010

Supporting the Community

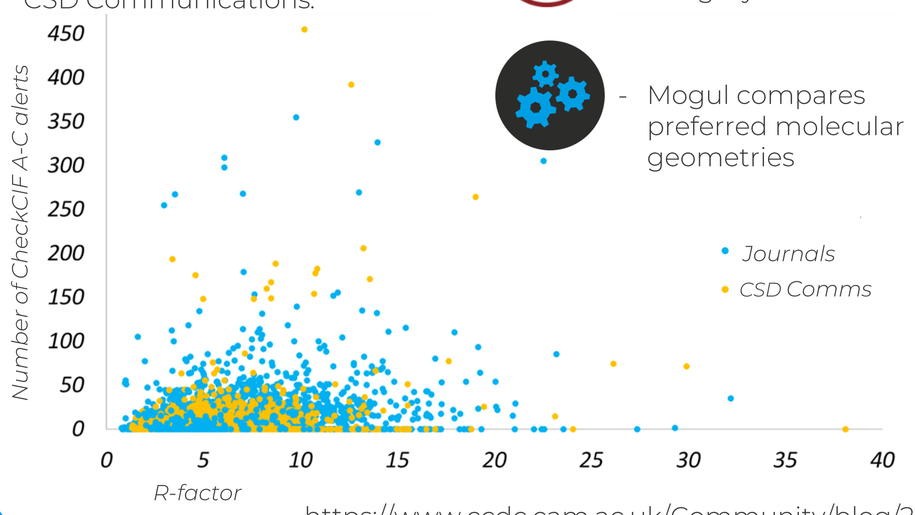
While adding more data we want to still maintain the quality of the CSD. To analyse how the quality of structures shared as *CSD Communications* compared with those from some popular journals we considered 3 key metrics.

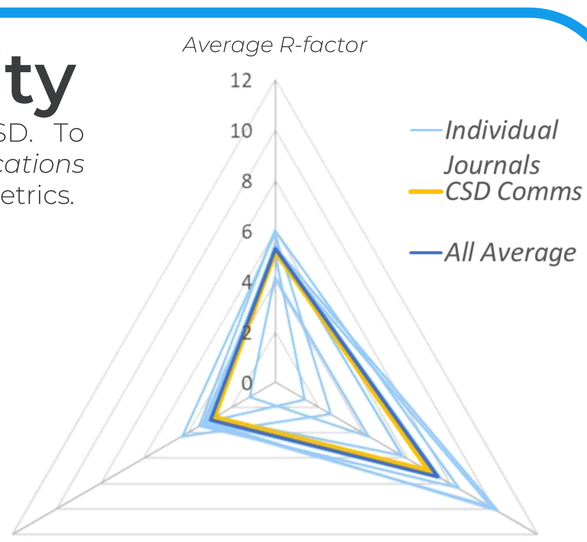
The results found that these methods do not show any clear correlation and that CSD quality does not appear to have been impacted by the increase in CSD Communications.

R-Factor - Standard reliability metric

checkCIF - Consistency and integrity check of data

Mogul compares preferred molecular geometries





CSD Communications sit close to the average values for all structures analysed over these metrics. From this we can be confident that the data being added is a valuable addition to the knowledge held in the CSD and can be reliably reused in other research. We are now working to develop a range of quality metrics and filters for the CSD.

https://www.ccdc.cam.ac.uk/Community/blog/2018-11-12-structure-quality/

Conclusions

With more of a focus on the importance of data and a resurgence of machine learning there is an even greater imperative to share chemical data and information reliably and efficiently. At the CCDC we have deployed a number of strategies to help the community share more data and we have seen a substantial increase in *CSD Communications* in recent years.

However, it is clear that there is still a significant amount of crystallographic data that does not get shared. Data can remain unpublished for a variety of reasons including ownership, permissions, data quality and time pressures. We will continue to evaluate how we can help you grow the amount of data shared worldwide and we would love to hear how we can better support you in your efforts.

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