Prepared 2019

Synchronized Searching:

Diving into CIF metadata to identify facilities for synchrotron studies within the CSD

N. T. Johnson, S. B. Wiggin and S. C. Ward
The Cambridge Crystallographic Data Centre, Cambridge, UK. E-mail: njohnson@ccdc.cam.ac.uk; Web: http://www.ccdc.cam.ac.uk

INTRODUCTION

The Cambridge Structural Database (CSD) is a fully validated database of organic and metal-organic crystal structures with > 950,000 entries. Alongside curating new structures, each year existing entries are enhanced with additional information or to improve the consistency of search terms. One planned enhancement is to ensure proper categorisation of synchrotron studies. To do this, data deposited to the CSD in Crystallographic Information File (CIF) format (~ 750,000 structures) was investigated using the CSD Python API for "synchrotron identifying information".[1]

As well as detecting additional synchrotron studies, the possibility of attributing these studies to particular facilities was investigated. Large facilities desire the ability to locate data measured on-site after it has been published, as research outcomes are not always reported back to them. There have been several projects in recent years focusing on the tracing of data in published research.[2]

METHODS

A list of keywords was created including: "syn", "tron", along with facility names and abbreviations. 8 different CIF attributes were searched for these keywords to identify studies undertaken with synchrotron radiation. The choice of attributes was determined based on a survey of a number of synchrotron structures and feedback from facilities.

Structures identified as synchrotron studies were then probed for a facility name or abbreviation. Beamlines were identified in a similar manner. Further information about the structures can be extracted from the CSD and can be used to create additional statistics, e.g. the year the structure was deposited in the database, the journal in which it was published.

Identification of synchrotron studies

A total of 10,095 structures were identified using the API script, including ~ 500 that were not previously flagged within the CSD.[3] The field that contained "synchrotron identifying information" varied between CIFs; Table 1 shows the number of synchrotron structures that had identifying information within certain attributes. Similar CIF attributes are grouped together: _diffrn_measurement_device/_diffrn_measurement_device_type and _diffrn_source/_diffrn_radiation_source.

	Radiation type	Source	Source type	Probe	Measurement	Monochromator
CIF Attribute	_diffrn_radiation _type	_diffrn_(source/ radiation_source)	_diffrn_source _type	_diffrn_radiation _probe	_diffrn_measurement _device/(_type)	_diffrn_radiation _monochromator
Number	8930	7960	705	12	867	615

Table 1 – CIF attribute(s) containing "synchrotron identifying information".

Identification of synchrotron facilities

A facility could be identified for 64% of structures, measured at 28 individual synchrotrons from across the world, indicating that the information is present in the majority of deposited CIFs. Figure 1 shows the number of unknown facilities by year. There is a generally increasing trend as the total number of structures increases, however, the % of structures where the facility is unknown is generally constant.

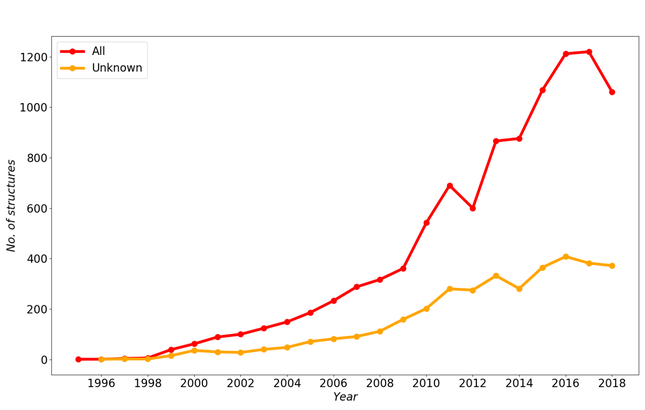

Figure 1 – Total and unknown facility synchrotron structures by year in the CSD.

An example of statistics which can be created for a facility is shown in Figure 2 – the number of structures appearing in the CSD attributed to the Synchrotron Radiation Source (SRS, Daresbury, UK) by year. The trend in the increase of structures up to 2008 mirrors the uptake of the CIF format, the numbers then decline after the synchrotron was decommissioned.

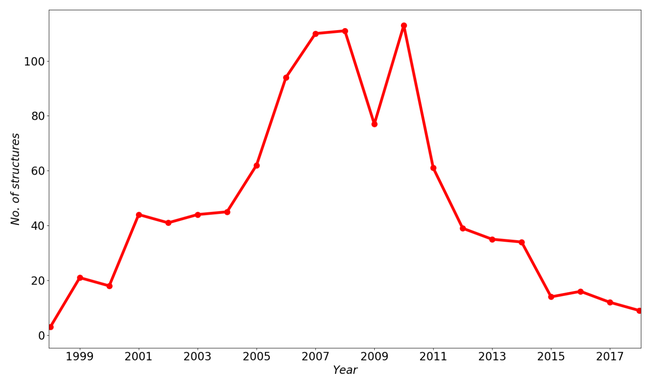

Figure 2 – Number of structures attributed to SRS by year.

The fields in which "facility identifying information" was most commonly identified are displayed in Table 2. This shows that information can be found in a variety of places.

Radiation type	Source	Source type	Probe	Measurement	Monochromator
75	5210	598	0	749	5

Table 2 – CIF attribute(s) containing "facility identifying information".

Identification of beamlines

Initial estimates show that > 45% of synchrotron studies also reported the beamline that was used. The % of structures with reported beamlines varies for each facility. Table 3 shows statistics for the SRS – where the beamline is reported for 97% of structures.

Beamline	# Structures
9.6	2
9.8	821
16.2	146
Unknown	35

Table 3 – Number of structures measured at SRS that can be attributed to a beamline.

Additional methods of identification

Not all structures measured using synchrotron radiation can be identified from textual information within the CIFs, therefore, other avenues were investigated. As laboratory X-ray sources for single crystal diffraction are usually reported at only a few known/standard wavelengths, structures measured with "non-standard" wavelengths could potentially have been collected at a synchrotron facility. 98% of CIFs contain a value for the wavelength.

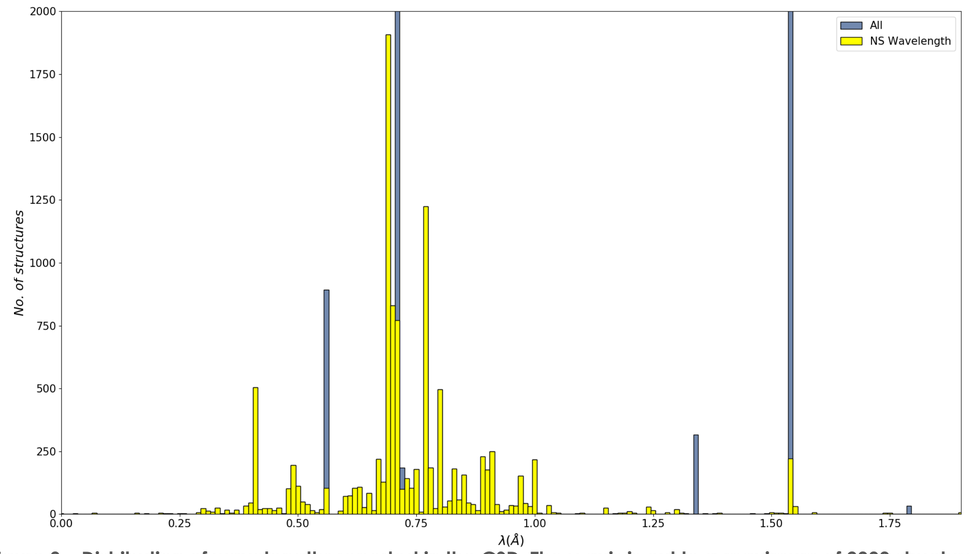

Figure 3 – Distribution of wavelengths reported in the CSD. The y axis is cut to a maximum of 2000 structures. Those with "non-standard" wavelengths are identified in yellow.

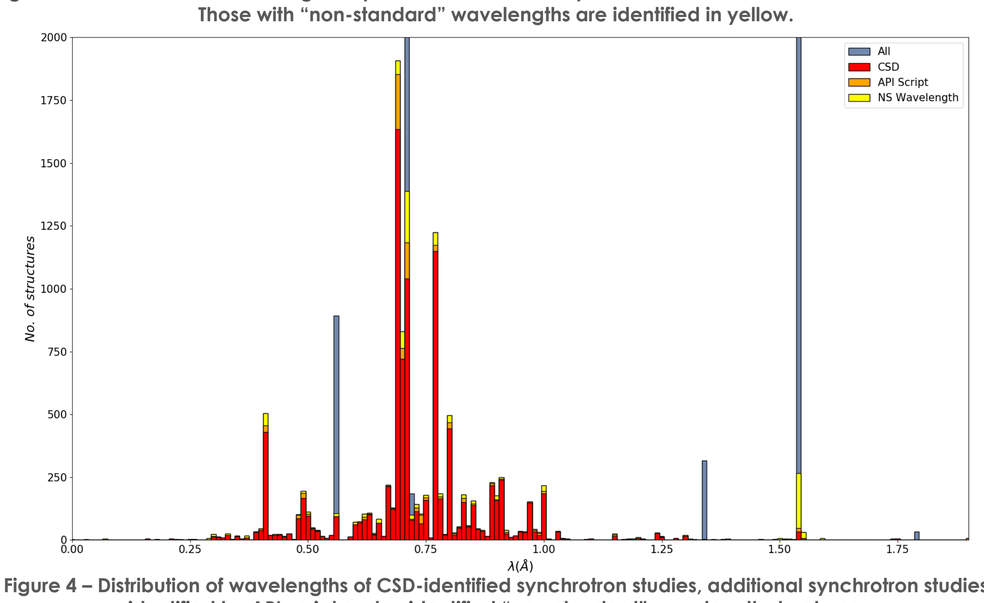

Figure 4 – Distribution of wavelengths of CSD-identified synchrotron studies, additional synchrotron studies identified by API script and unidentified "non-standard" wavelength structures.

Figure 3 shows the distribution of CIFs reporting "non-standard" wavelengths. Many of these structures are identified as synchrotron studies in Figure 4, indicating that this is likely a plausible method to identify other studies. An additional 801 structures were identified as potential synchrotron studies based on their wavelength. This is not a route to identify a potential facility, however, as many facilities use similar wavelengths. Other methods to identify synchrotron information could also be investigated.

CONCLUSIONS

This poster demonstrates that a search of CIF attributes could be an effective mechanism to identify facilities for data deposited in the CSD. This work additionally highlights the importance of including synchrotron and facility identifying information within CIFs in order to appropriately classify studies and to provide increased traceability for facilities. This could lead to the formation of new CIF guidelines in the synchrotron community.

ACKNOWLEDGEMENTS

We thank John Helliwell, Mike Hoyland, Mark Warren, Brian McMahon, Jason Price, Amy Sarjeant and Peter Strickland for helpful discussions relating to this work.

1: Groom, C R, et al., *Acta Cryst.*, 2016, B**72**, 171-179.
2: Haak, L L, et al., User Facilities and Publications Working Group: Findings and Opportunities Report. ORCiD. [Online], DOI:10.23640/07243.5623750.v1, (accessed January 2019).
3: Using CSD version 5.40 (November 2018)