Synchrotron data in the CSD

Small molecule perspectives of synchrotron data and raw data sharing

Natalie Johnson

Data Integrity Research Scientist

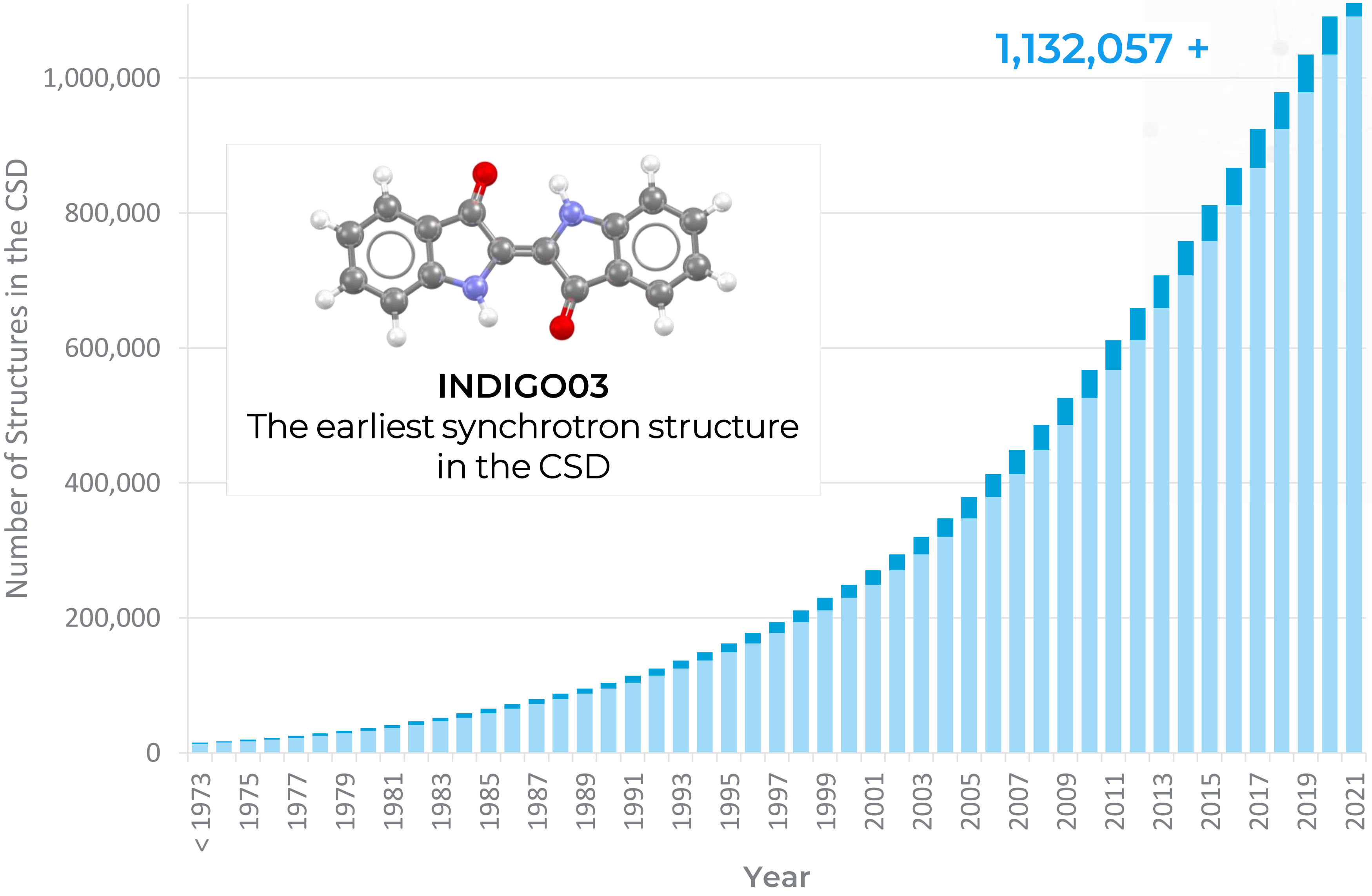
Workshop on MX raw image data formats, metadata and validation

14th August 2021

CCDC

advancing structural science

The Cambridge Structural Database (CSD)



INDIGO03
The earliest synchrotron structure in the CSD

The CSD is a database of small molecule organic and metal-organic crystal structures.

Every entry enriched and annotated by experts.

Every published structure:

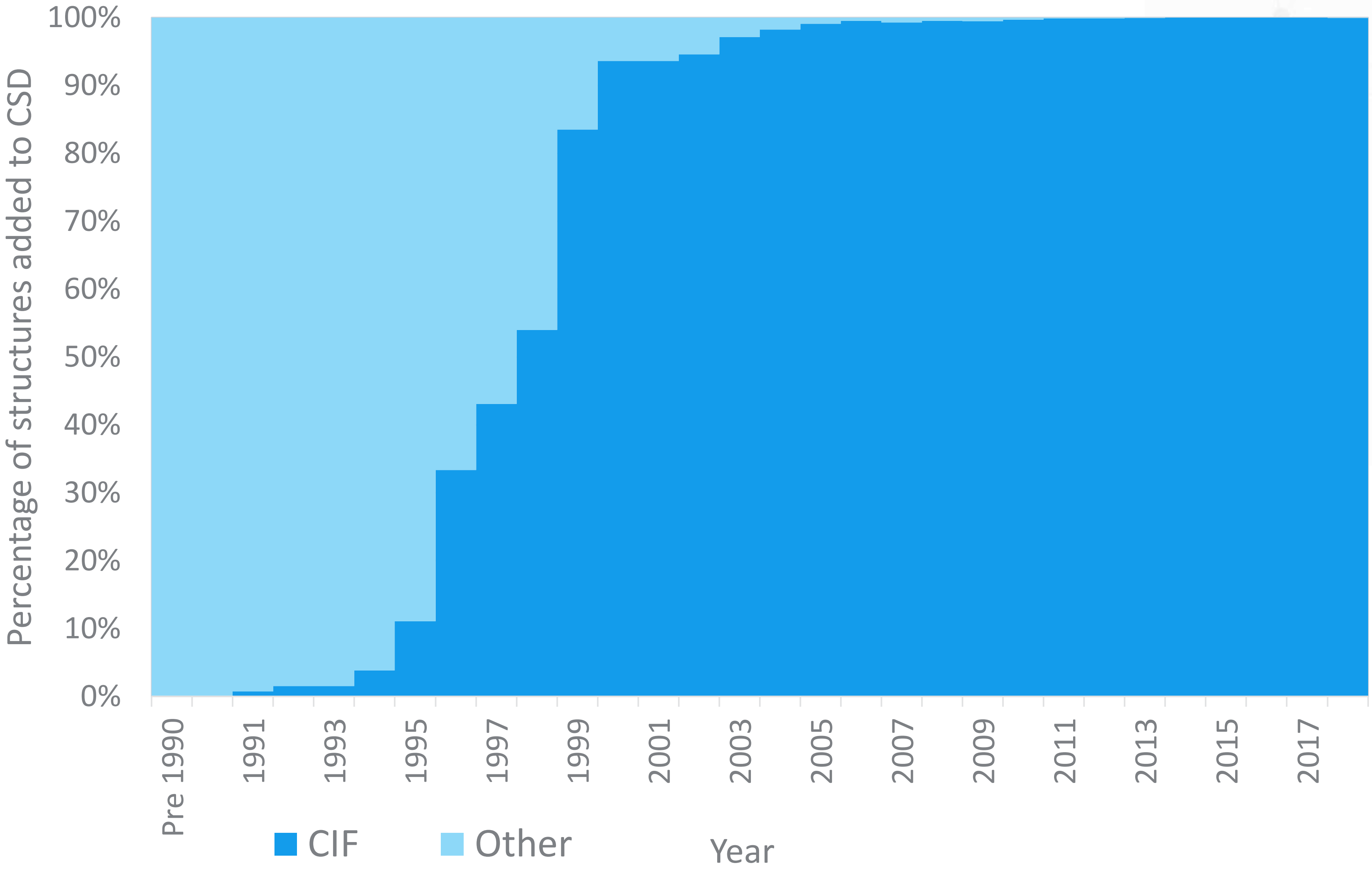
- Inc. ASAP & early view
- *CSD Communications*
- Patents
- University repositories

A trusted CoreTrustSeal repository

CORE TRUST SEAL

CCDC

The Cambridge Structural Database (CSD)



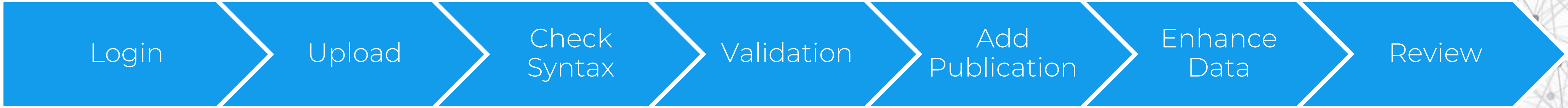
Data can be deposited to CCDC through an online deposition platform or via email.

Pre-CIF, entries were created by typing data from tables in papers.

When CIF is not available, CCDC manually creates a CIF from the accessible information.

365 structures (< 0.6%) added to the CSD with a manually created CIF in 2019.

Joint ICSD/CSD web deposition service



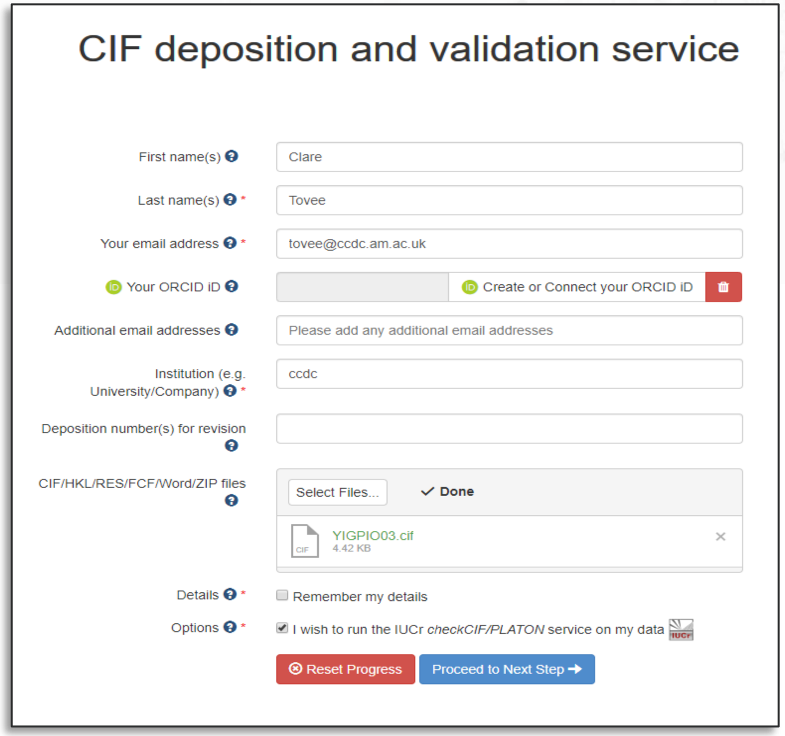
Login → Upload → Check Syntax → Validation → Add Publication → Enhance Data → Review

Tailored web deposition service enabling scientists to deposit crystal structures ready for publication

Scientists encouraged to deposit data pre-publication:

- Workflows with a number of publishers to assist with identifying publication details and allowing secure access to data for peer-review process.
- Multi-stage deposition progress to help scientist check and enhance their data.

https://www.ccdc.cam.ac.uk/deposit

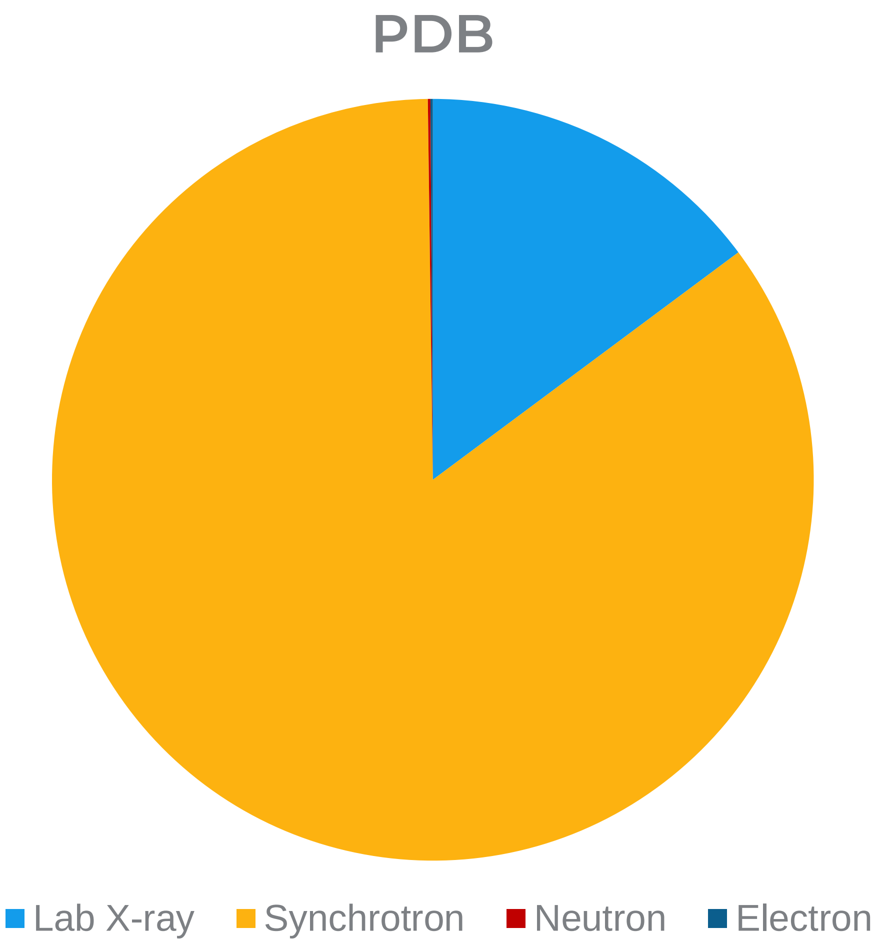


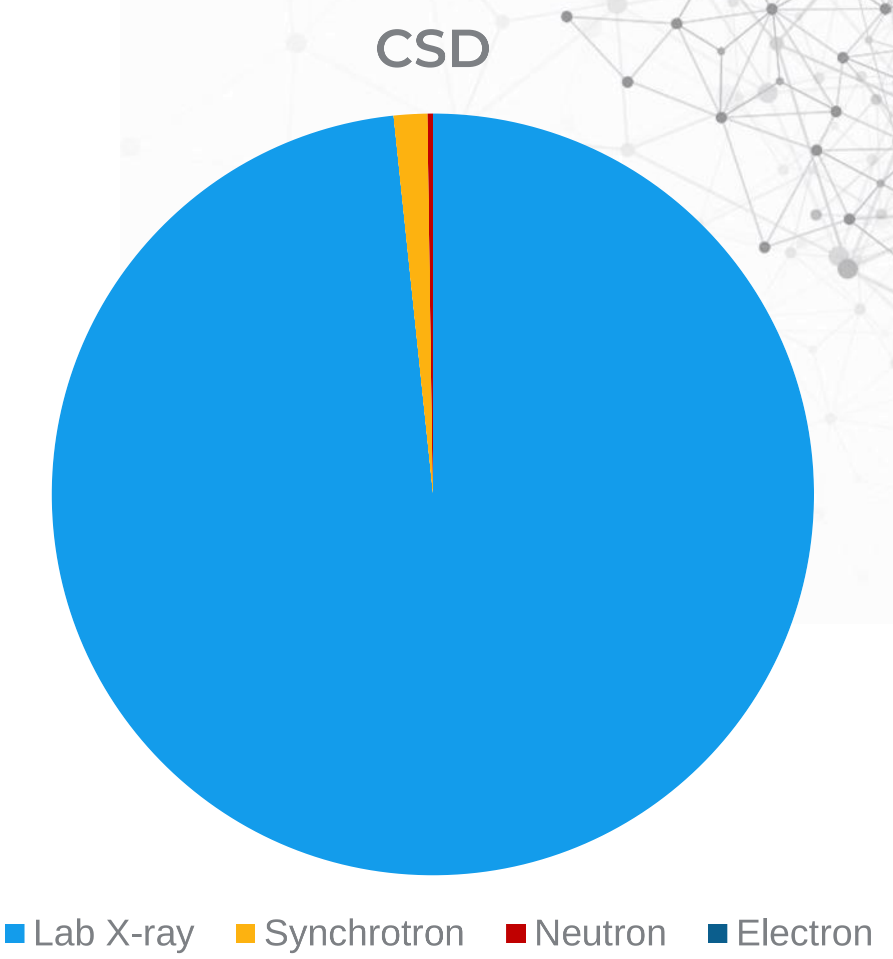
CIF deposition and validation service

First name(s) ?	Clare
Last name(s) ? *	Tovee
Your email address ? *	tovee@ccdc.am.ac.uk
Your ORCID iD ?	Create or Connect your ORCID iD
Additional email addresses ?	Please add any additional email addresses
Institution (e.g. University/Company) ? *	ccdc
Deposition number(s) for revision ?	
CIF/HKL/RES/FCF/Word/ZIP files ?	Select Files... ✓ Done YIGPIO03.cif 4.42 KB ×
Details ? *	□ Remember my details
Options ? *	☑ I wish to run the IUCr *checkCIF/PLATON* service on my data

Reset Progress | Proceed to Next Step →

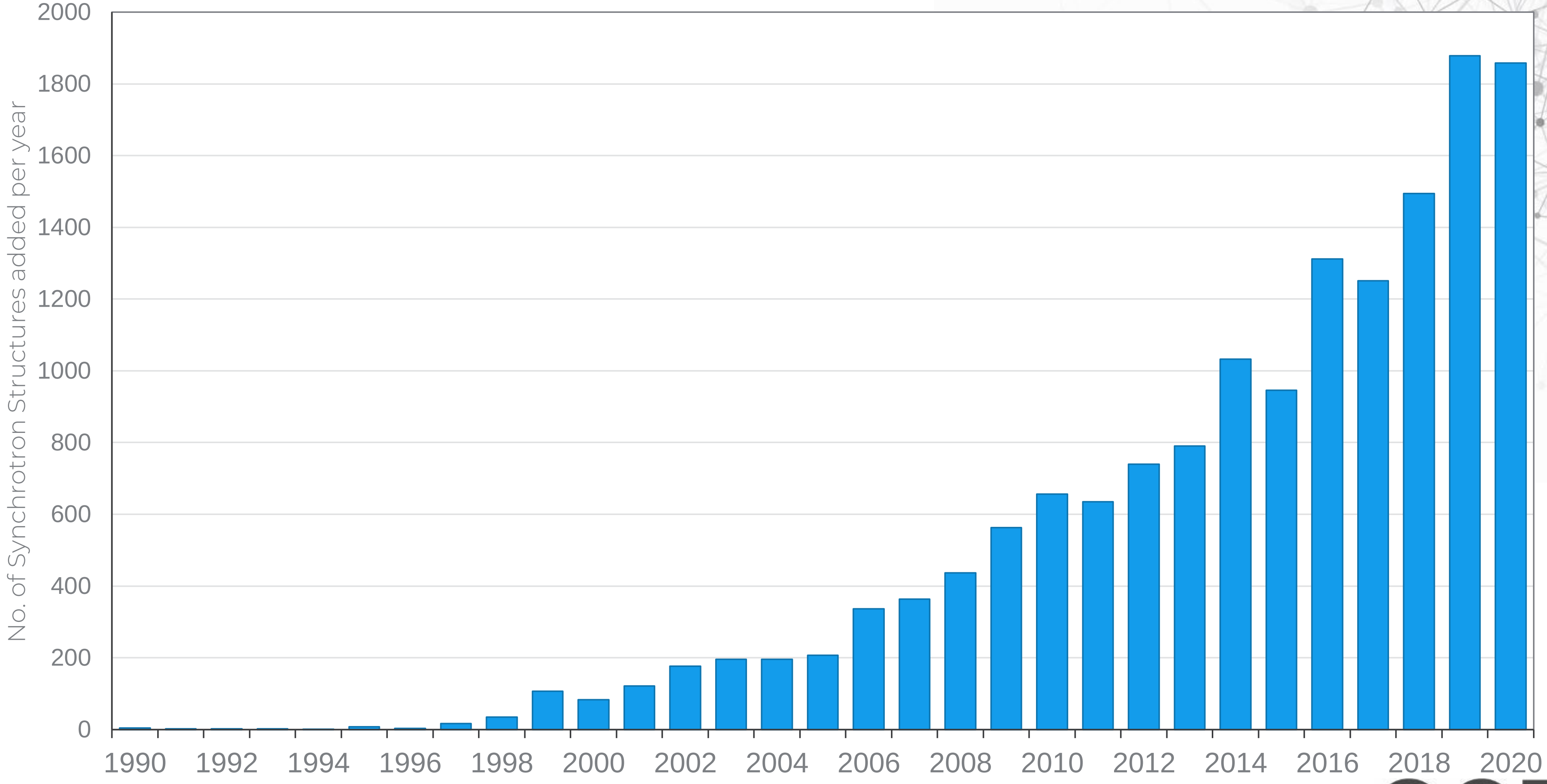
Synchrotron data in the PDB vs CSD





1 PDB Statistics from Biosync (http://biosync.sbkb.org/) and PDB website, accessed 25/06/2020
Diffraction data only.

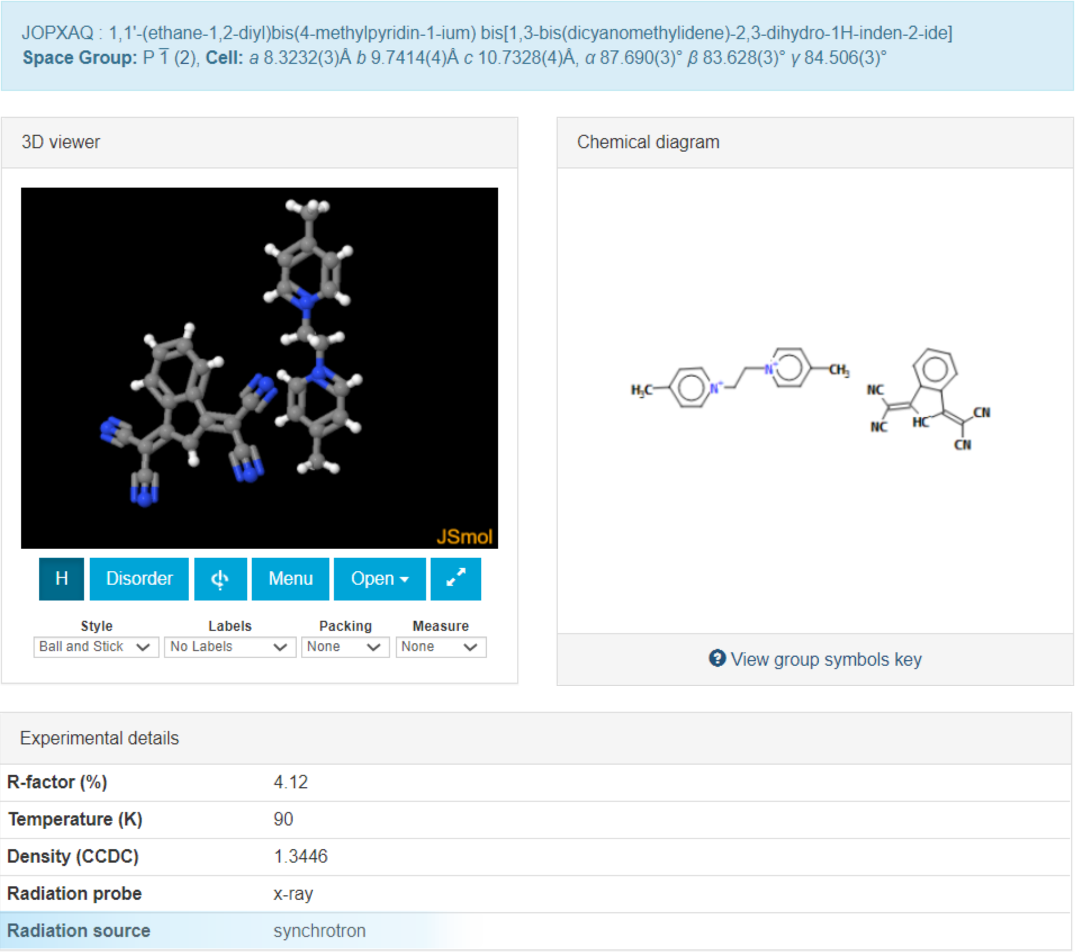
Synchrotron data in the CSD



2017: ~1500 structures from one paper removed from graph

Identification of synchrotron data

- Structures are flagged with ‘synchrotron’ in the CSD.
- These structures are identified during the data curation process either automatically by the curation software, which checks a select number of CIF fields, or manually labelled by an Editor.



JOPXAQ : 1,1'-(ethane-1,2-diyl)bis(4-methylpyridin-1-ium) bis[1,3-bis(dicyanomethylidene)-2,3-dihydro-1H-inden-2-ide]
Space Group: P $\bar{1}$ (2), **Cell:** *a* 8.3232(3)Å *b* 9.7414(4)Å *c* 10.7328(4)Å, *α* 87.690(3)° *β* 83.628(3)° *γ* 84.506(3)°

3D viewer

Style: Ball and Stick | Labels: No Labels | Packing: None | Measure: None

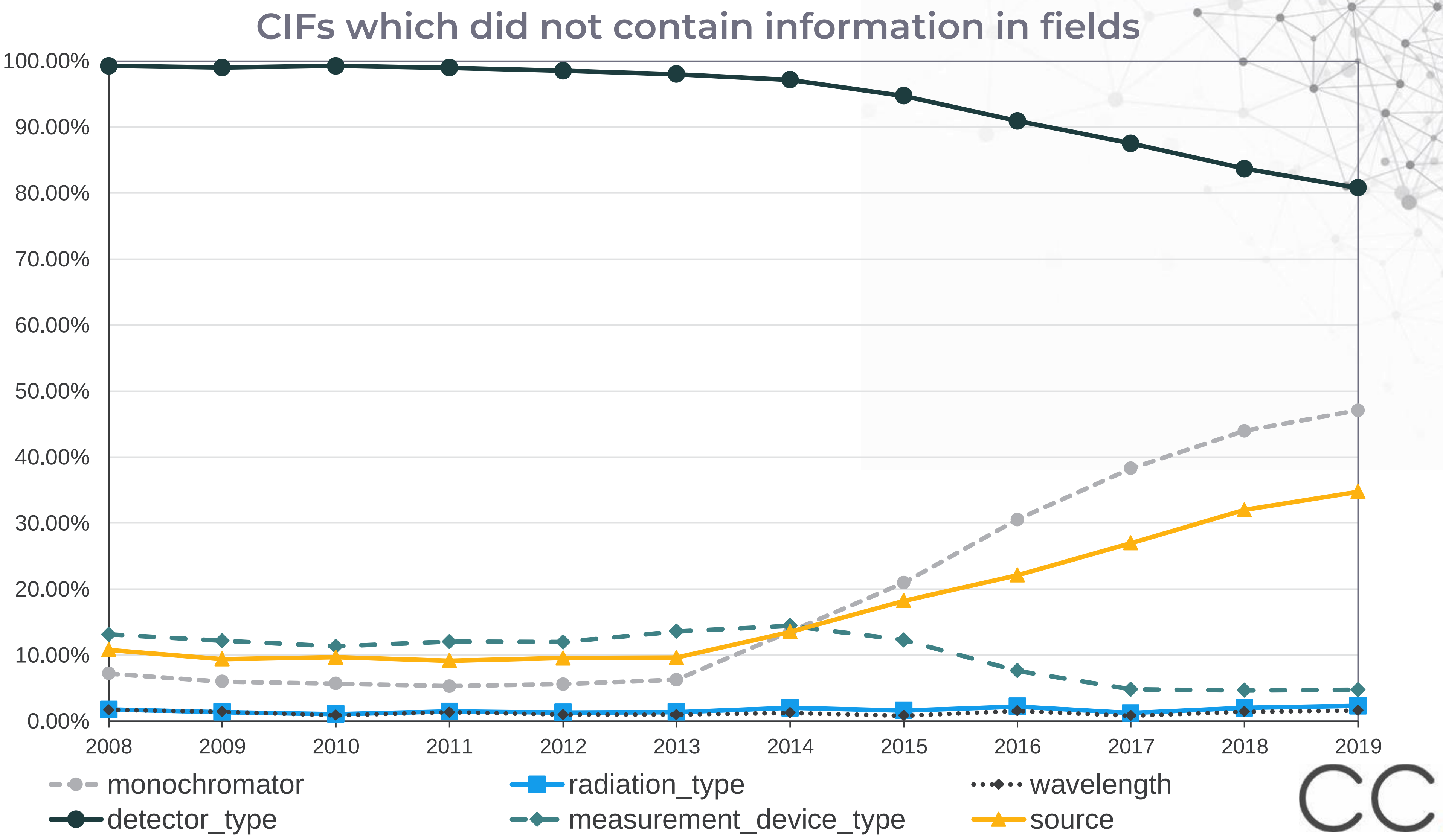
Chemical diagram

View group symbols key

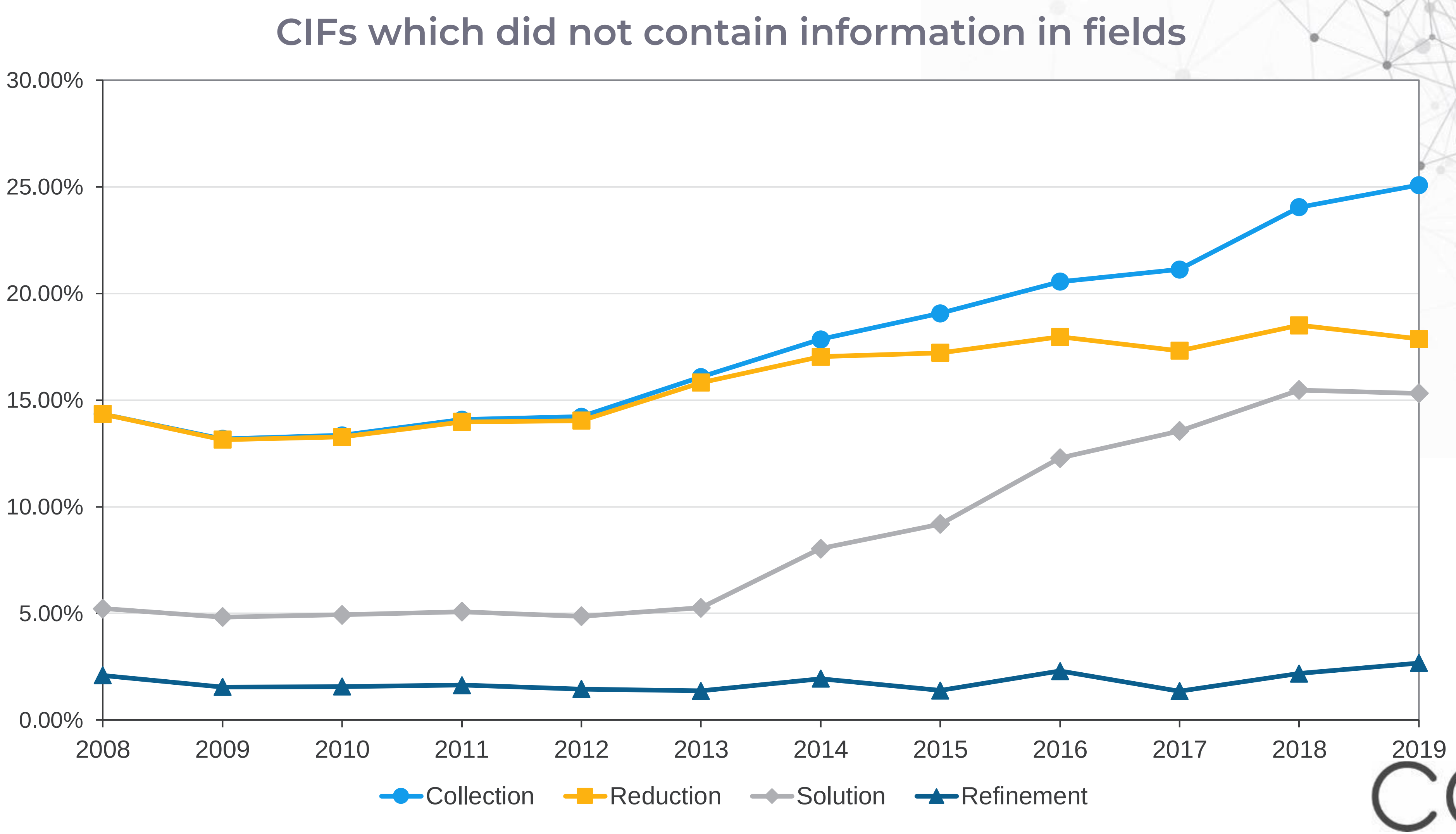
Experimental details

R-factor (%)	4.12
Temperature (K)	90
Density (CCDC)	1.3446
Radiation probe	x-ray
Radiation source	synchrotron

CIF completeness - Experimental

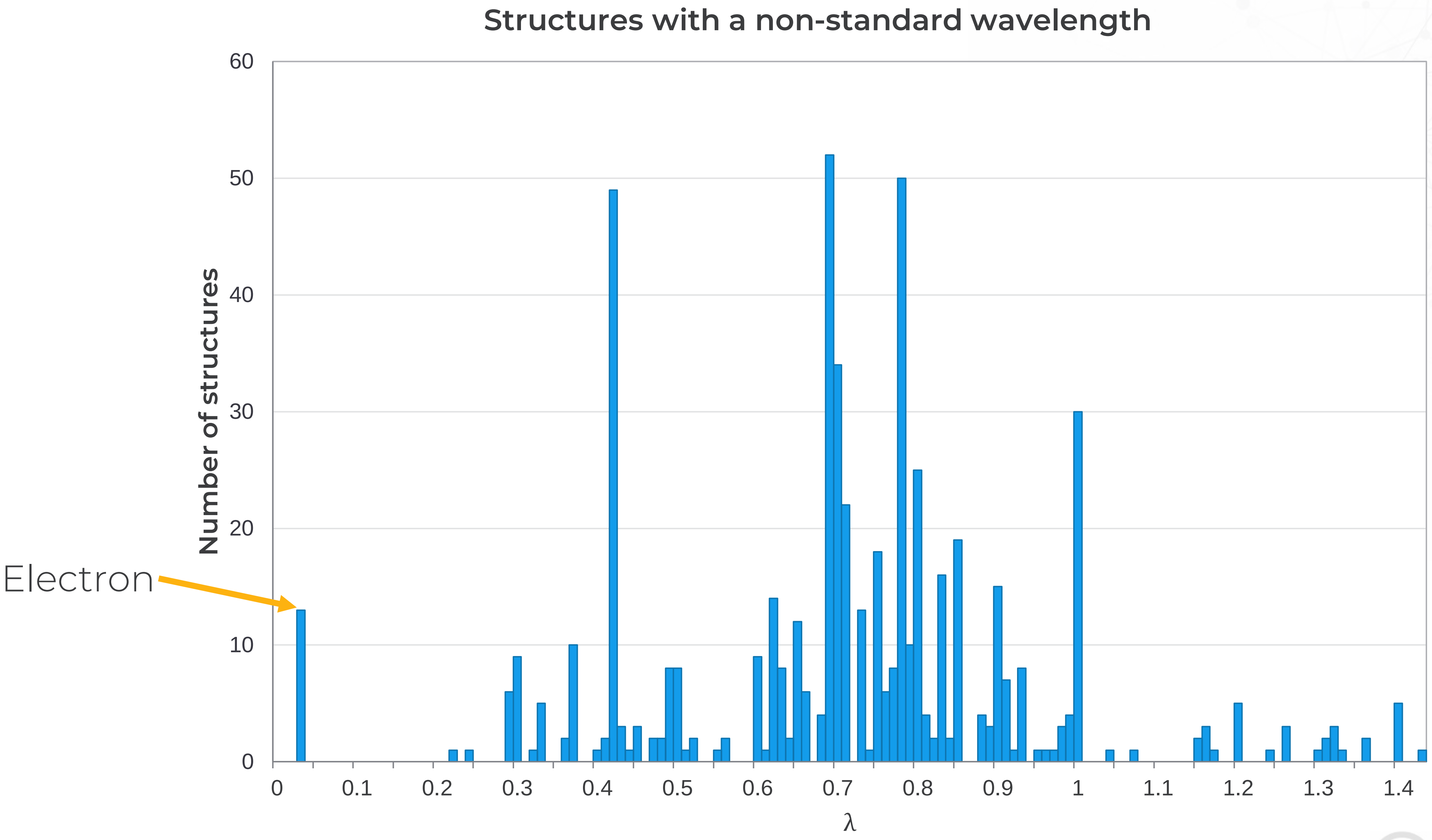


CIF completeness - Computational

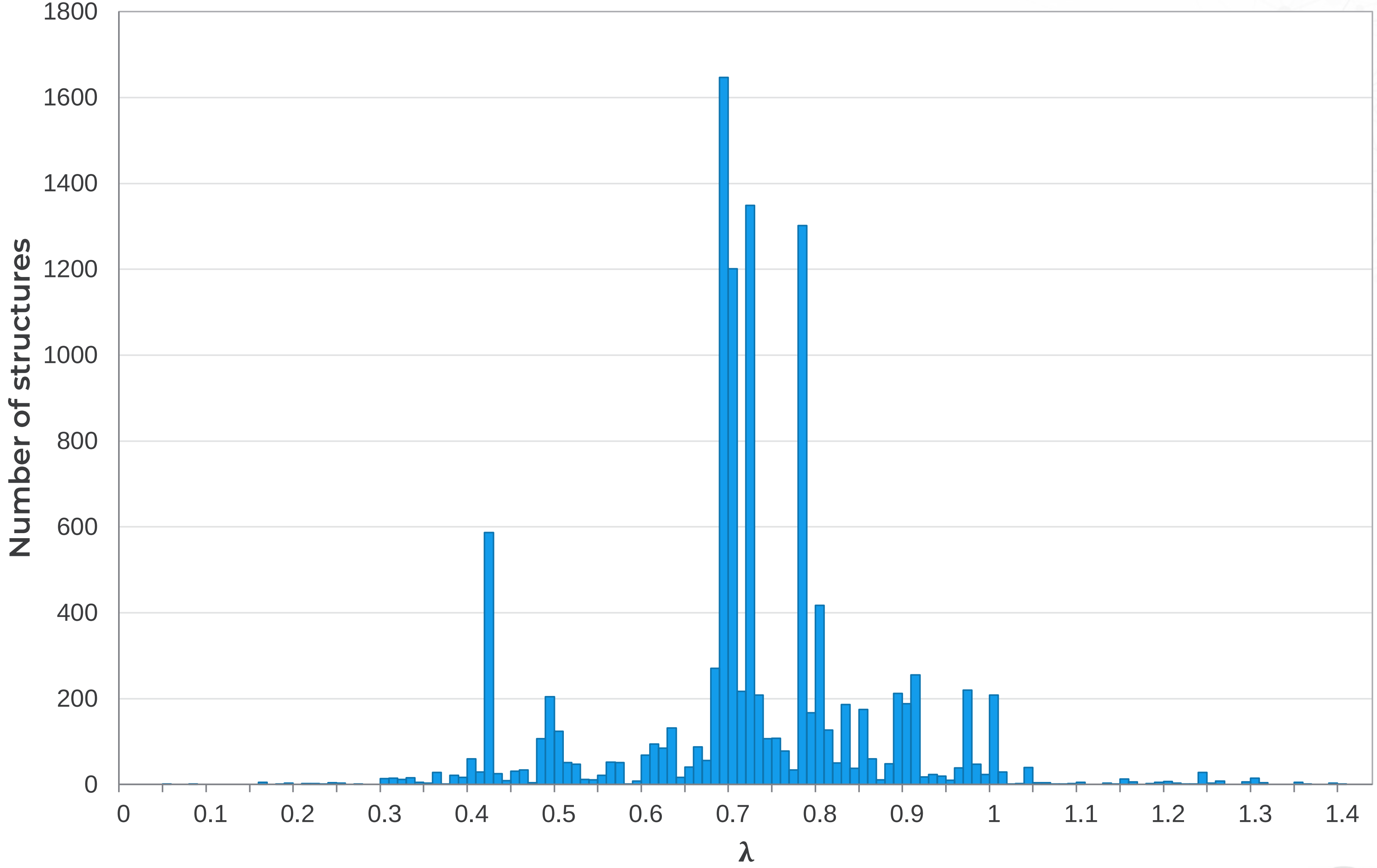


CIF completeness – Outcomes

- Some structures were not identifiable from the CIF as it did not contain complete information.
  - Additional structures found by cross-referencing papers associated with the facilities and structures in the CSD using CSD Python API and Publication DOIs.
  - CIF templates created for two facilities highlighting required information.
- Other inconsistencies found in a number of synchrotron CIFs e.g. radiation_type ‘synchrotron’ and _diffrn_source ‘sealed tube’/’rotating anode’

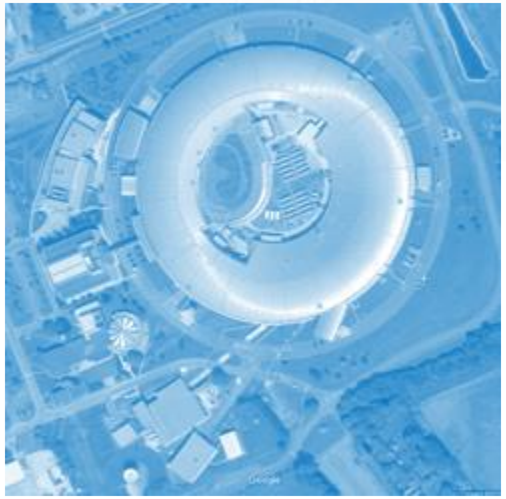

Structures with a non-standard wavelength
Number of structures
60
50
40
30
20
10
0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
1.4
λ
Electron

Wavelength of synchrotron structures



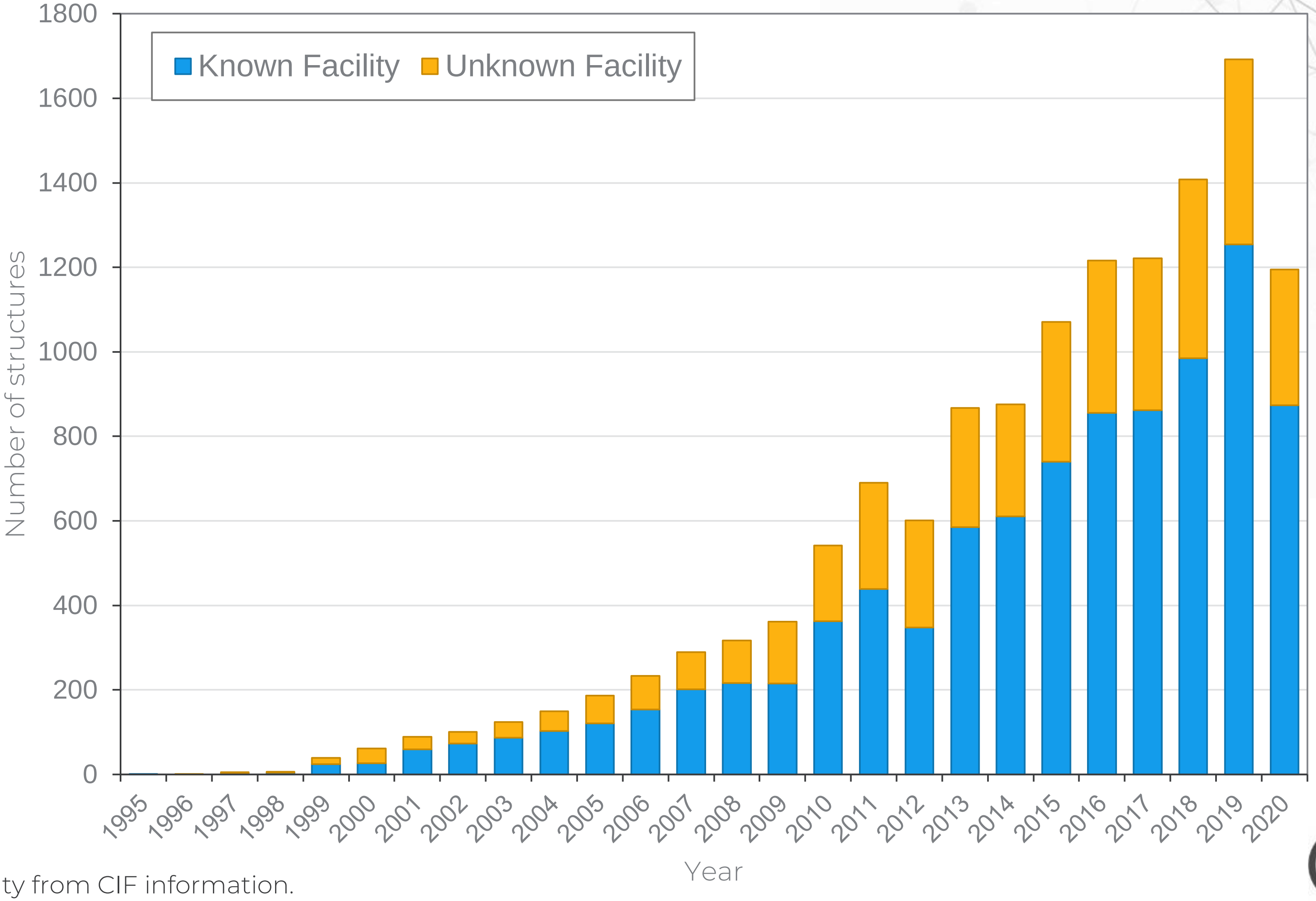
Facility information

- Using synchrotron studies identified with Python API.
- Attempted to associate with a particular synchrotron facility – using either names, acronyms or specific beamline information (e.g. 'Swiss-Norwegian beamline' at ESRF).
- Identified studies from 30 individual synchrotrons.
- 69% of studies could be attributed to a facility.



Imagery ©2019 Google, Map data ©2019

Known and unknown facility structures

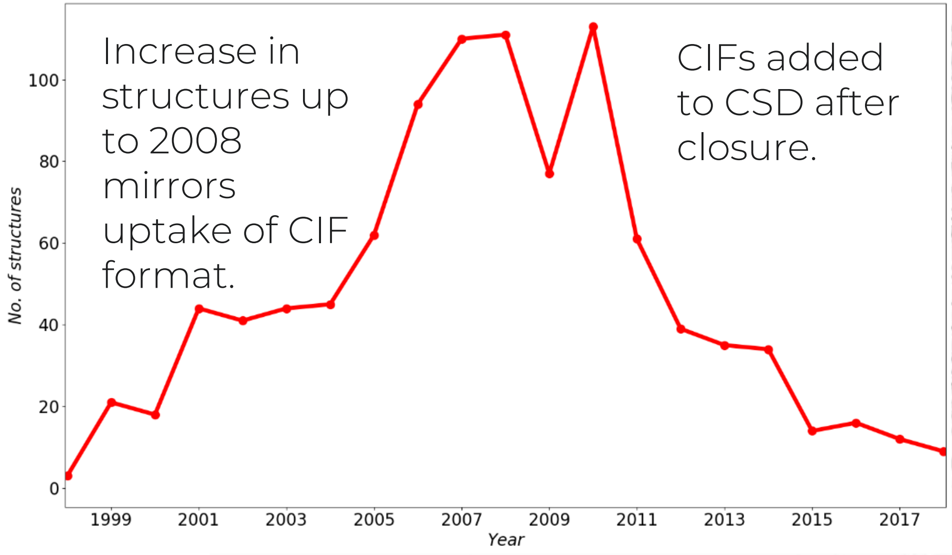


Identification of synchrotron facility from CIF information.
2020 numbers are incomplete (covers Jan-Oct 2020)

# Structures by Journal for 2008

Journal	# Structures
Dalton Transactions	17
Chemistry – A European Journal	14
Inorganic Chemistry	12
Journal of the American Chemical Society	8
European Journal of Inorganic Chemistry	8
Angewandte Chemie International Edition	6
Acta Crystallographica Section C	6
Chemical Communications	5
Crystal Growth and Design	4
Inorganic Chemistry Communications	3
Journal of Organometallic Chemistry	3
Organic and Biomolecular Chemistry	3
Acta Crystallographica Section E	2
CrystEngComm	2
Journals with 1 structure	10

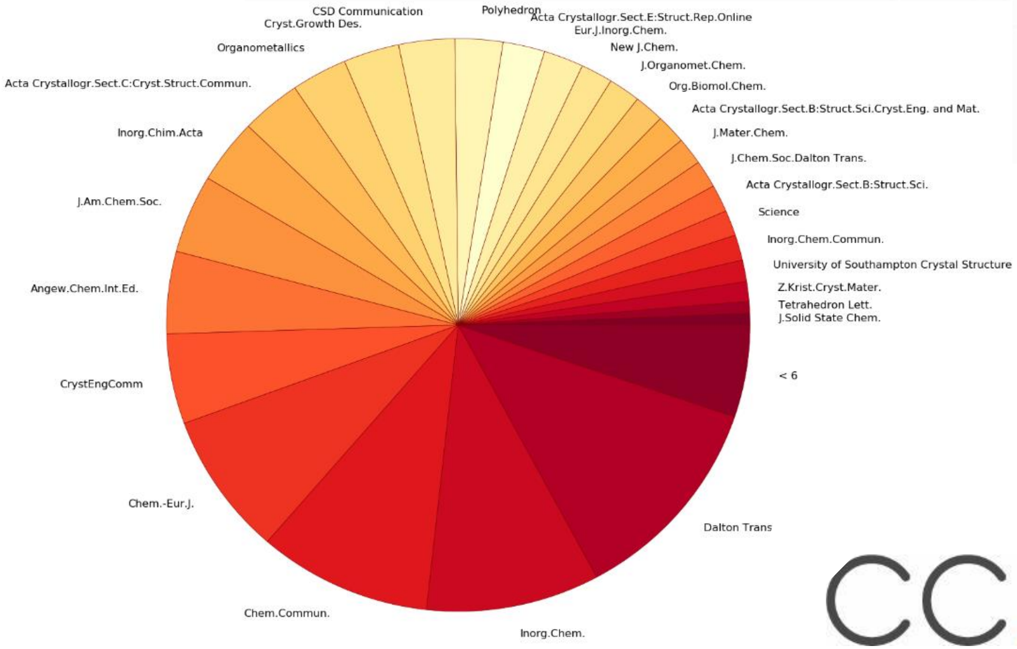
Structures per year



Increase in structures up to 2008 mirrors uptake of CIF format.

CIFs added to CSD after closure.

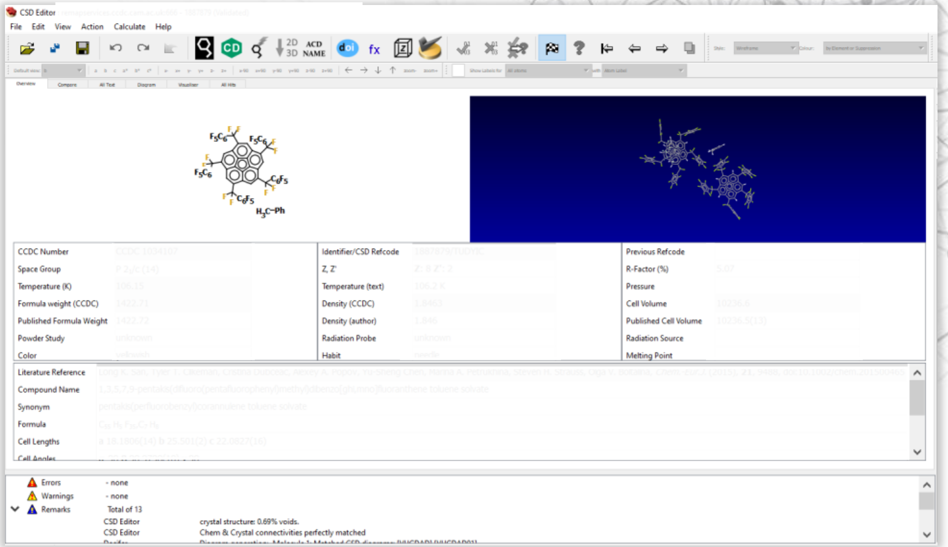
Overall Journal Information



Example of data insights: Synchrotron Radiation Source, UK

FAIR and FACT small molecule data

- Curation to ensure structures are labelled and presented consistently so are **findable** within the CSD.
- Option to download additional information associated with entry in Access Structures.



CCDC Number	CCDC 1034107	Identifier/CSD Refcode	1887879/TUDYIC	Previous Refcode	
Space Group	P 2_1/c (14)	Z, Z'	Z: 8 Z': 2	R-Factor (%)	5.07
Temperature (K)	106.15	Temperature (text)	106.2 K	Pressure	
Formula weight (CCDC)	1422.71	Density (CCDC)	1.8463	Cell Volume	10236.6
Published Formula Weight	1422.72	Density (author)	1.846	Published Cell Volume	10236.5(13)
Powder Study	unknown	Radiation Probe	unknown	Radiation Source	
Color	yellowish	Habit	needle	Melting Point	

Literature Reference: Long K. San, Tyler T. Clikeman, Cristina Dubceac, Alexey A. Popov, Yu-Sheng Chen, Marina A. Petrukhina, Steven H. Strauss, Olga V. Boltalina, *Chem.-Eur.J.* (2015), 21, 9488, doi:10.1002/chem.201500465

Compound Name: 1,3,5,7,9-pentakis(difluoro(pentafluorophenyl)methyl)dibenzo[ghi,mno]fluoranthene toluene solvate

Synonym: pentakis(perfluorobenzyl)corannulene toluene solvate

Formula: $C_{55} H_5 F_{35}, C_7 H_8$

Cell Lengths: a 18.1806(14) b 25.501(2) c 22.0827(16)

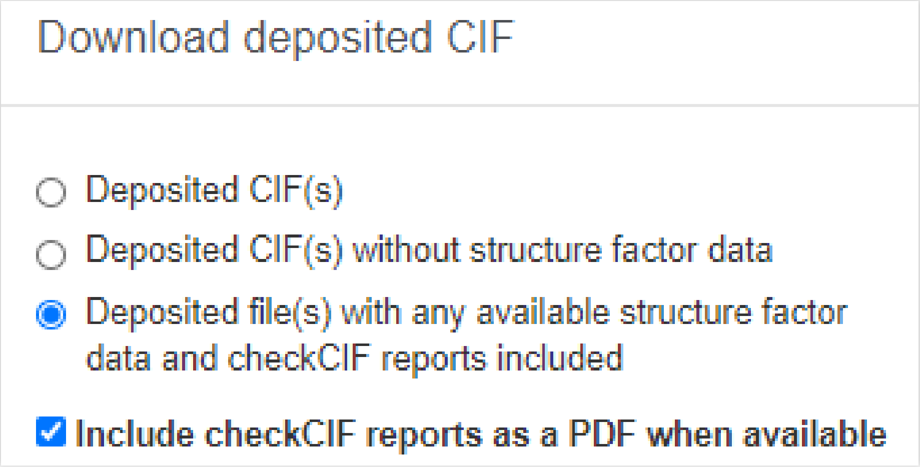
Errors - none

Warnings - none

Remarks Total of 13

CSD Editor crystal structure: 0.69% voids.

CSD Editor Chem & Crystal connectivities perfectly matched



Download deposited CIF

○ Deposited CIF(s)

○ Deposited CIF(s) without structure factor data

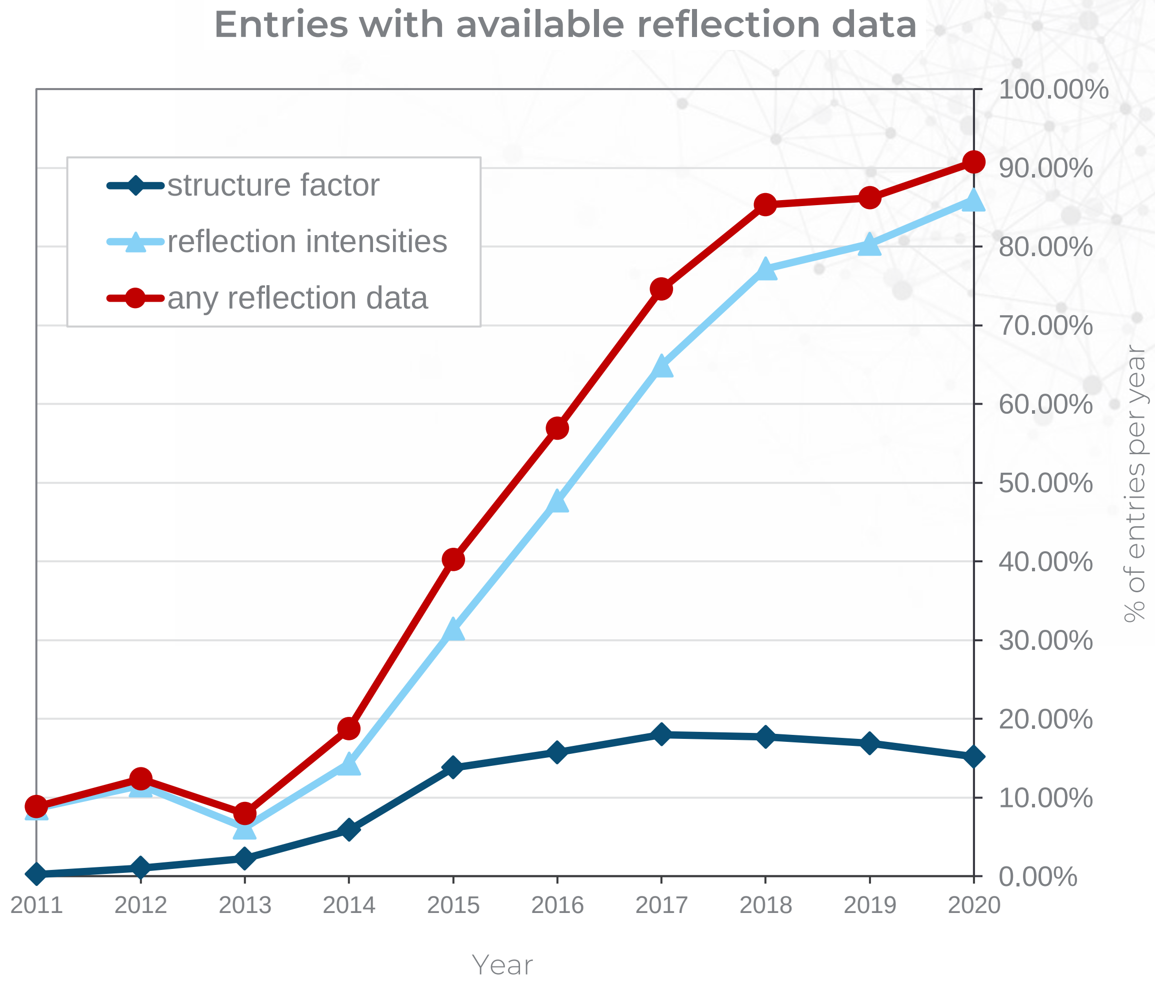
● Deposited file(s) with any available structure factor data and checkCIF reports included

☑ Include checkCIF reports as a PDF when available

Helliwell, J. (2019). *Structural Dynamics*. 6, 054306., DOI:10.1063/1.5124439

Structure factor information

- The deposition of structure factors is not explicitly required by CCDC.
- CCDC has accepted deposition of structure factor data since 2011.
- Reflection information can either be included in CIF or uploaded as a separate file.



https://www.ccdc.cam.ac.uk/Community/blog/10-years-of-structure-factors/

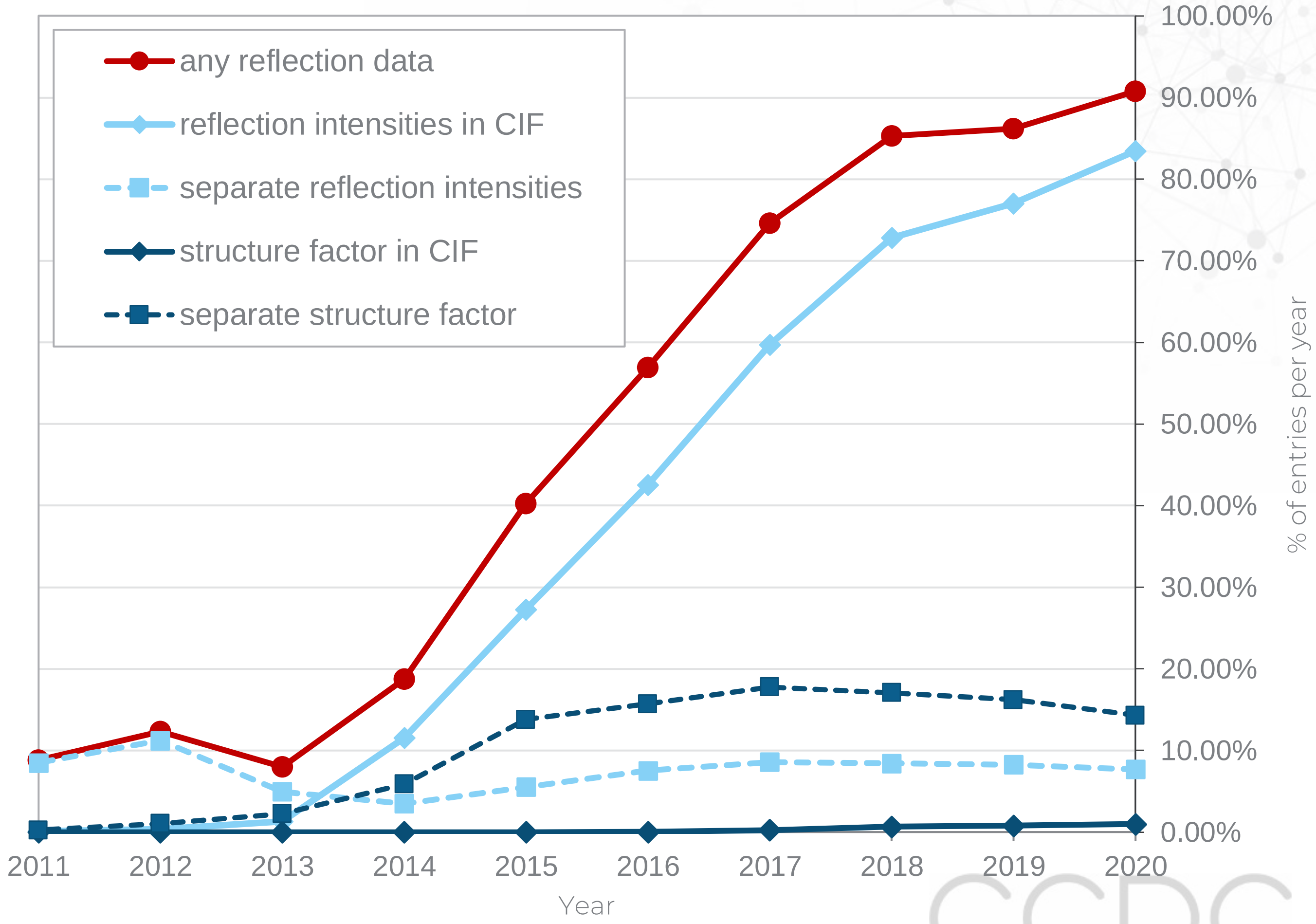
2020 data covers January to October

CCDC

Included in CIF vs separate files

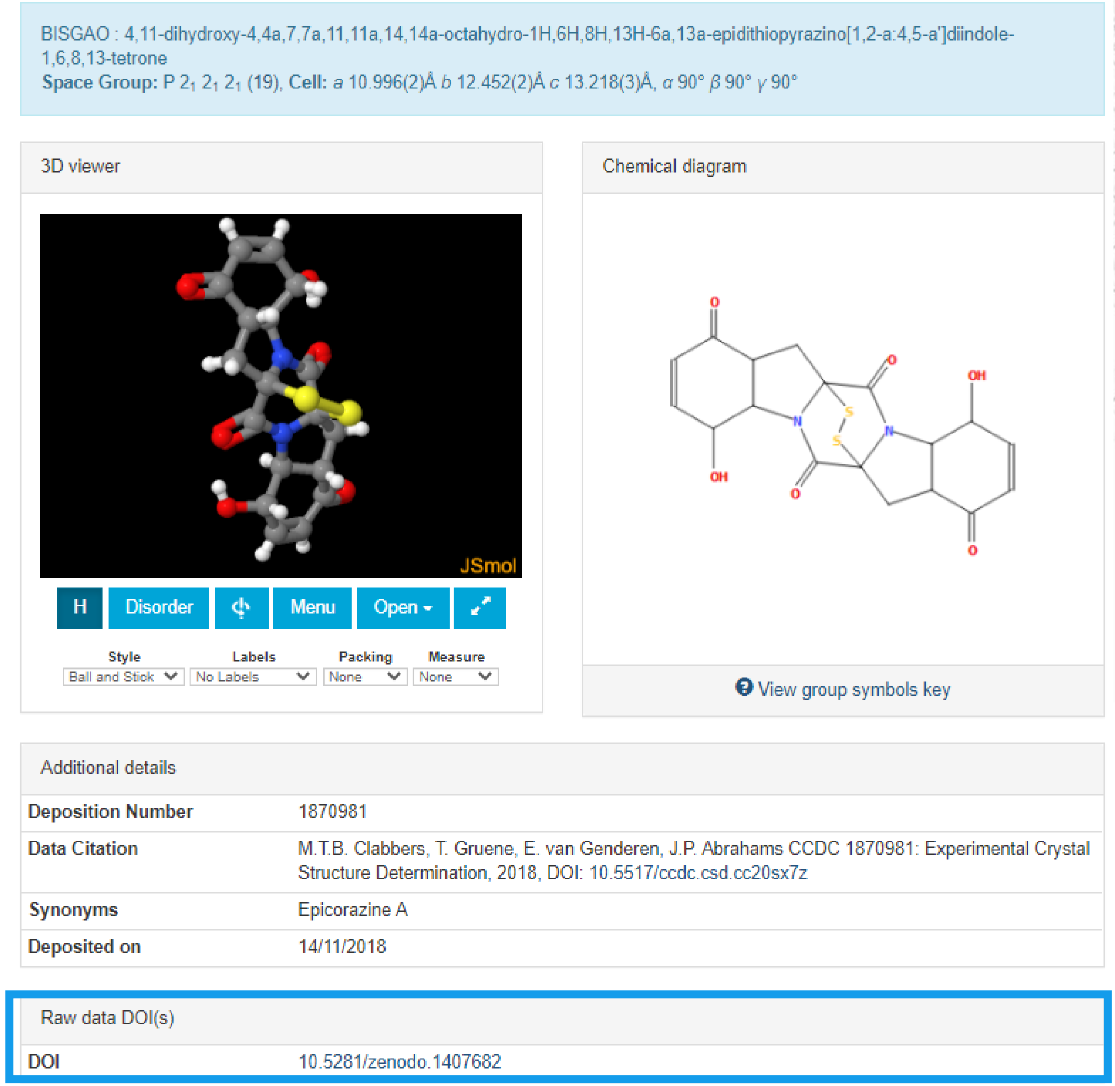
- Most reflection intensity information is within the CIF whereas most structure factor data is provided separately.
- Some journals require separate structure factor files.
- Not all software includes structure factors in CIF automatically.

Entries with available reflection data



Raw diffraction data

- Currently CCDC don’t store raw diffraction data for entries.
- **Raw Data DOI** – DOI link to raw raw diffraction data stored in a separate archive.
- The Raw Data DOI is associated with the entry in Access Structures.



BISGAO : 4,11-dihydroxy-4,4a,7,7a,11,11a,14,14a-octahydro-1H,6H,8H,13H-6a,13a-epidithiopyrazino[1,2-a:4,5-a']diindole-1,6,8,13-tetrone

Space Group: P 2_1 2_1 2_1 (19), **Cell:** *a* 10.996(2)Å *b* 12.452(2)Å *c* 13.218(3)Å, *α* 90° *β* 90° *γ* 90°

Additional details

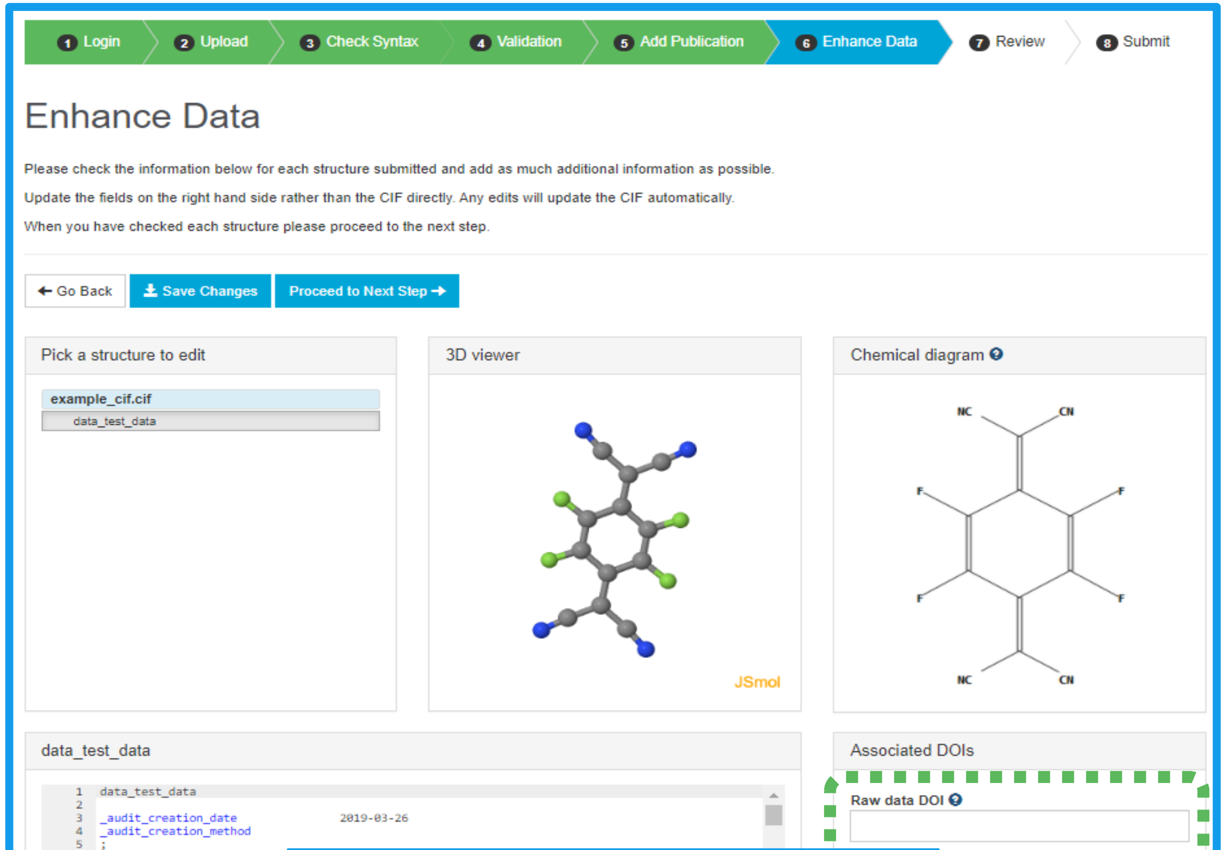
Deposition Number	1870981
Data Citation	M.T.B. Clabbers, T. Gruene, E. van Genderen, J.P. Abrahams CCDC 1870981: Experimental Crystal Structure Determination, 2018, DOI: 10.5517/ccdc.csd.cc20sx7z
Synonyms	Epicorazine A
Deposited on	14/11/2018

Raw data DOI(s)

DOI	10.5281/zenodo.1407682

https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=e1fc6d58-e3b7-e711-b787-005056977c87

Providing Raw Data DOI



Enhance Data

Please check the information below for each structure submitted and add as much additional information as possible.

Update the fields on the right hand side rather than the CIF directly. Any edits will update the CIF automatically.

When you have checked each structure please proceed to the next step.

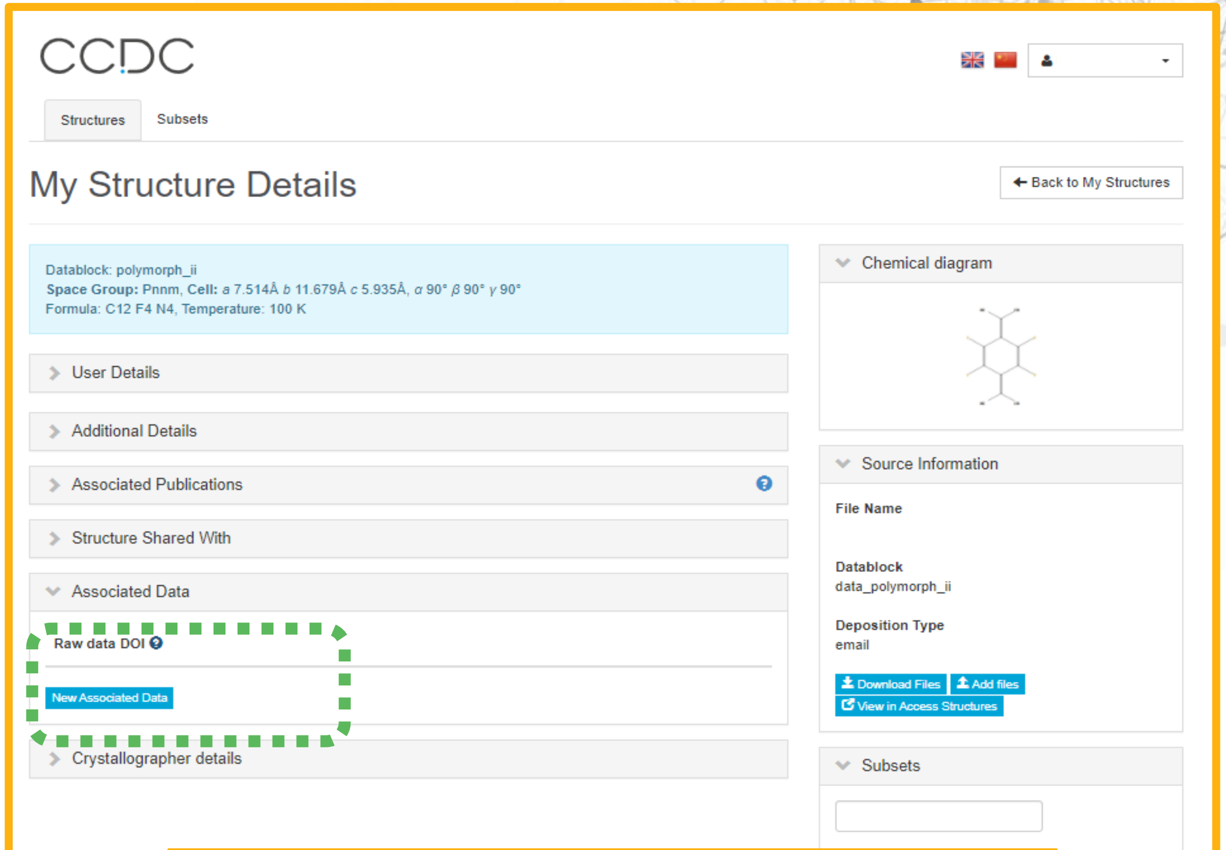
data_test_data

```
data_test_data

_audit_creation_date              2019-03-26
_audit_creation_method
;
Olex2 1.2
```

Raw data DOI

During Web Deposition



My Structure Details

Datablock: polymorph_ii
Space Group: Pnnm, **Cell:** *a* 7.514Å *b* 11.679Å *c* 5.935Å, *α* 90° *β* 90° *γ* 90°
Formula: C12 F4 N4, Temperature: 100 K

Raw data DOI

File Name

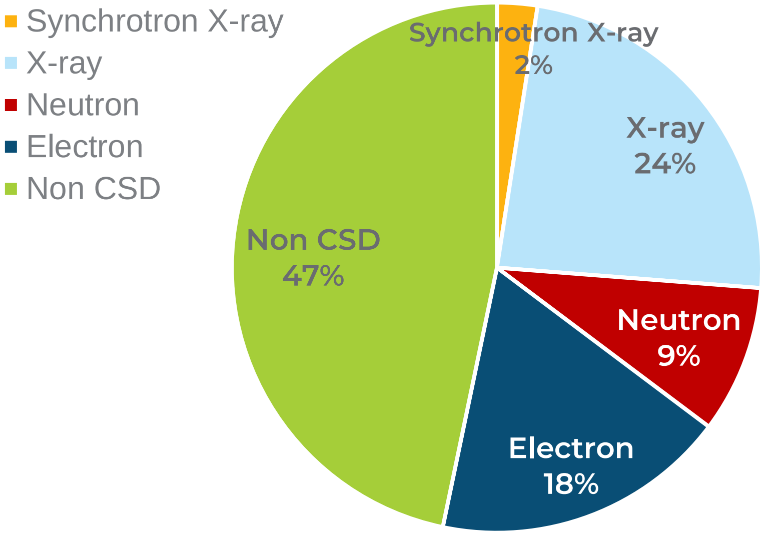
Datablock
data_polymorph_ii

Deposition Type
email

Post Deposition in My Structures

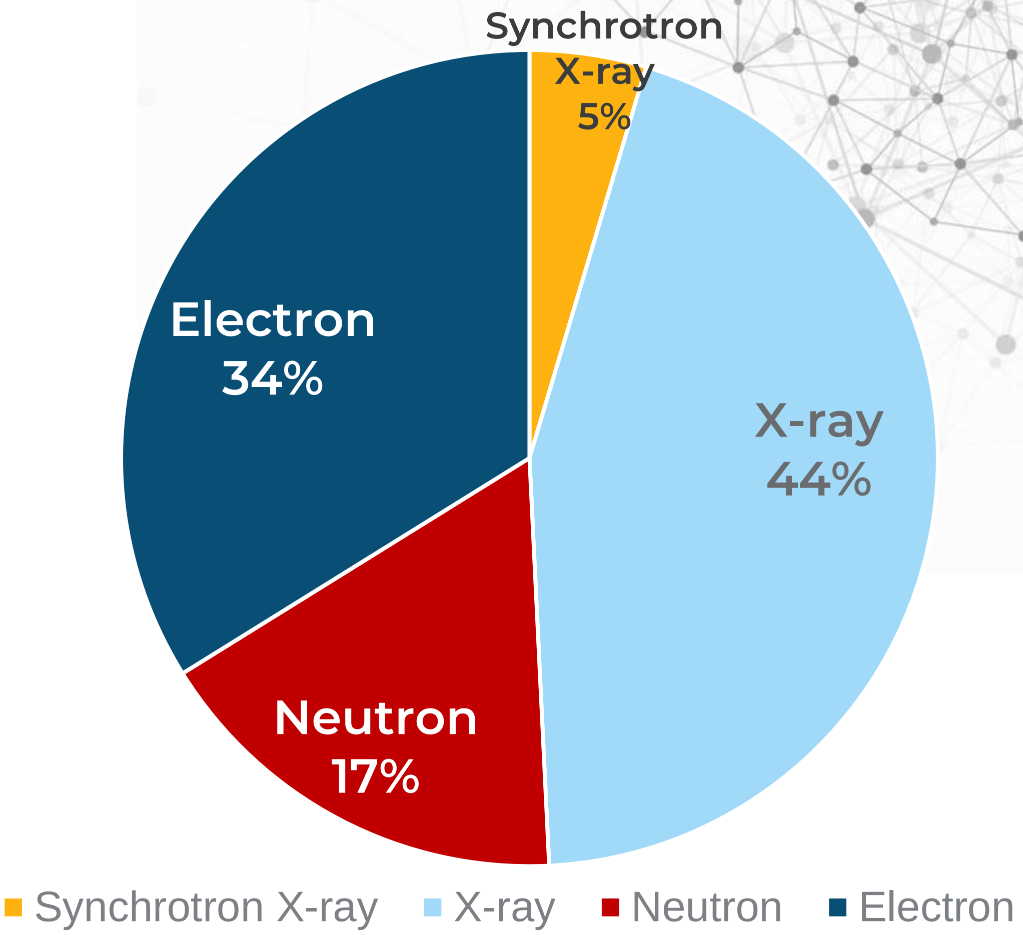
Raw diffraction data

- Currently only a small number of structures have an associated Raw Data DOI.
- As of July 2021: 65 structures (0.005% of CSD).

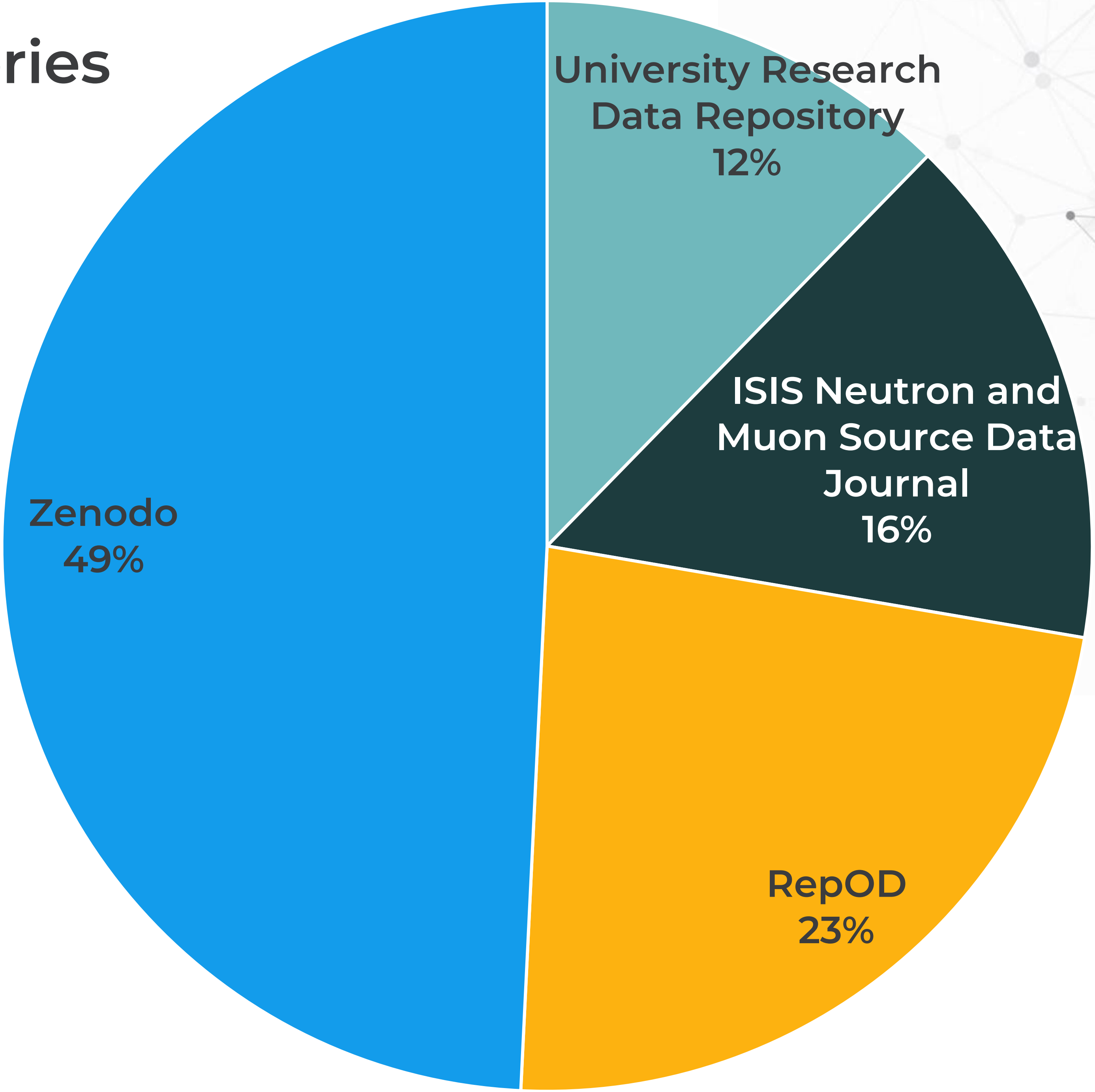


All Raw Data DOIs in Access Structures

Entries with Raw Data DOIs

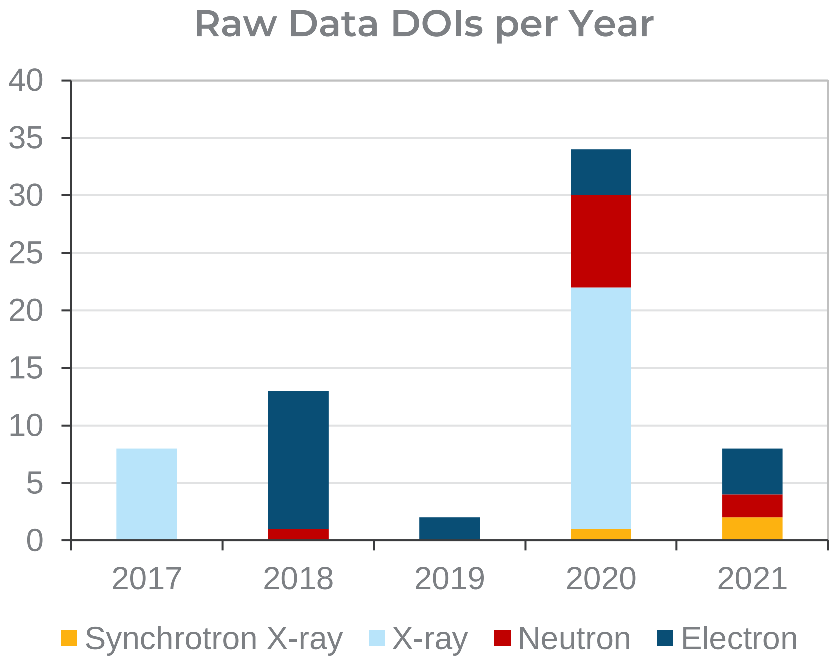


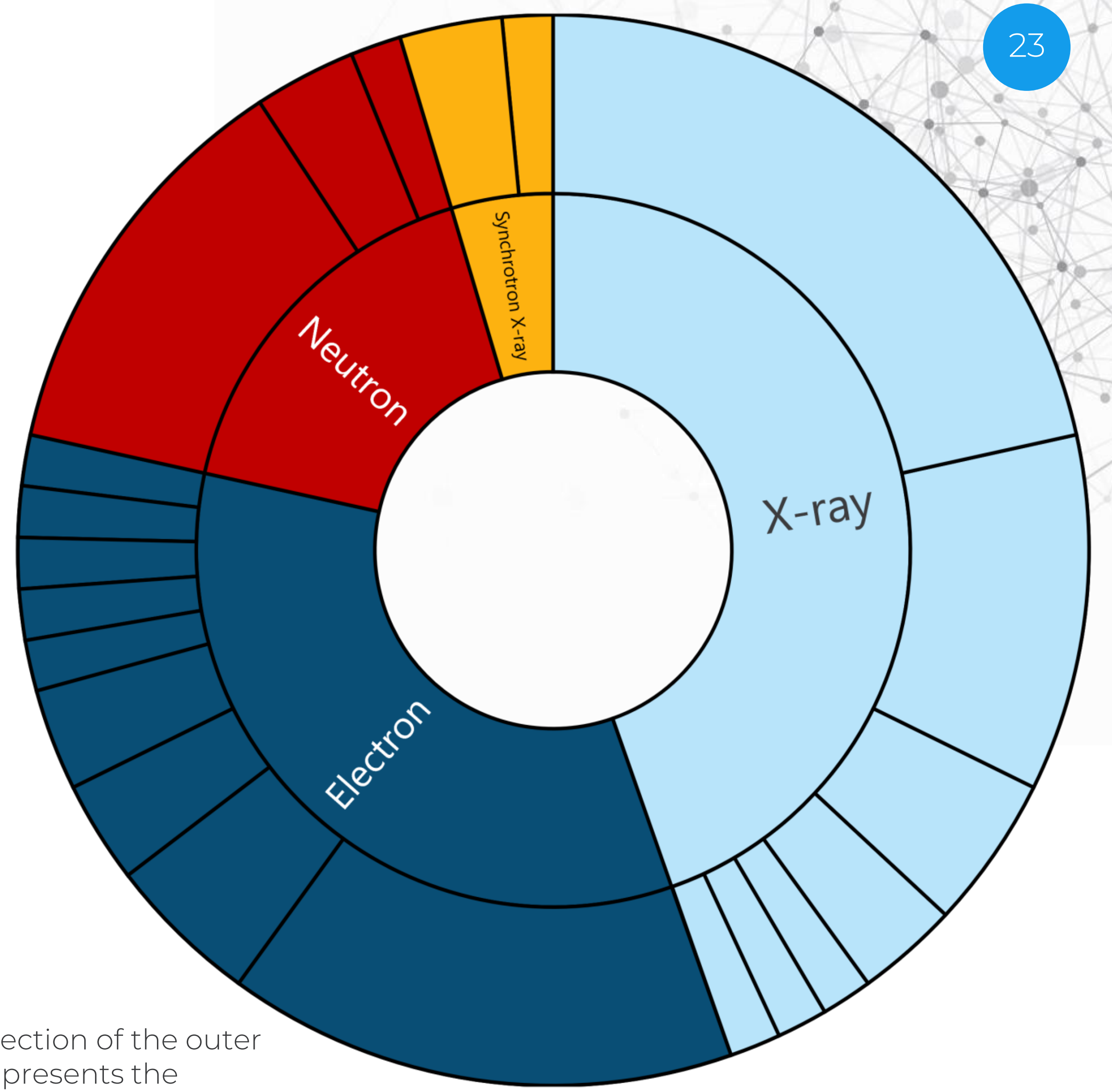
Data repositories



Publications

- All data types are mostly dominated by one publication.
    - One third of X-ray structures with Raw Data DOI come from one paper.





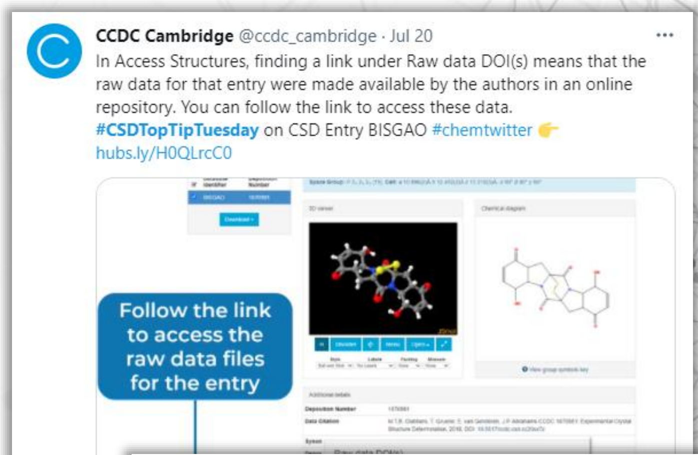
Each section of the outer ring represents the number of structures in a individual publication.

Raw Data DOI Issues

- Information in Raw Data DOI field is checked by hand by Data Team.
    - More often a paper DOI is given in this field.
    - Files aren’t checked in repository.
- Time consuming to search for additional structures in repositories
    - Requires key phrases in metadata/enough information to match back to a CSD Entry (Publication DOI, CCDC number).
    - Some archives can store DOI links to publications or data (CCDC Data DOI).

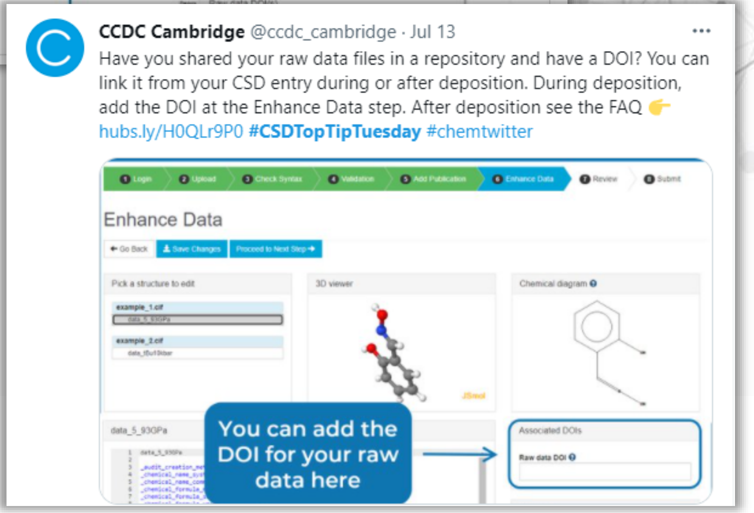
Promoting raw data sharing

- Top Tip Tuesdays on Twitter
- Highlighting Raw Data DOI in trainings
- FAQs on our website
- Emphasising in blogs and presentations



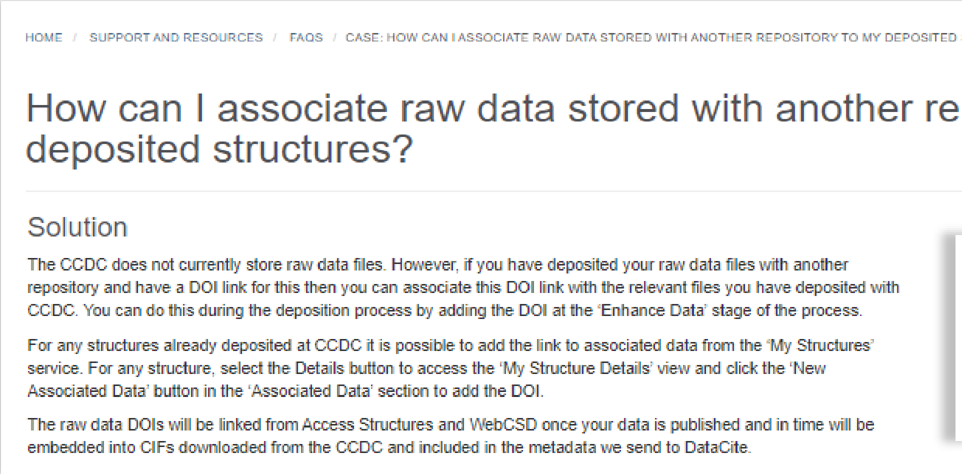
CCDC Cambridge @ccdc_cambridge · Jul 20

In Access Structures, finding a link under Raw data DOI(s) means that the raw data for that entry were made available by the authors in an online repository. You can follow the link to access these data. #CSDTopTipTuesday on CSD Entry BISGAO #chemtwitter 👉 hubs.ly/H0QLrcC0



CCDC Cambridge @ccdc_cambridge · Jul 13

Have you shared your raw data files in a repository and have a DOI? You can link it from your CSD entry during or after deposition. During deposition, add the DOI at the Enhance Data step. After deposition see the FAQ 👉 hubs.ly/H0QLr9P0 #CSDTopTipTuesday #chemtwitter



HOME / SUPPORT AND RESOURCES / FAQS / CASE: HOW CAN I ASSOCIATE RAW DATA STORED WITH ANOTHER REPOSITORY TO MY DEPOSITED STRUCTURES?

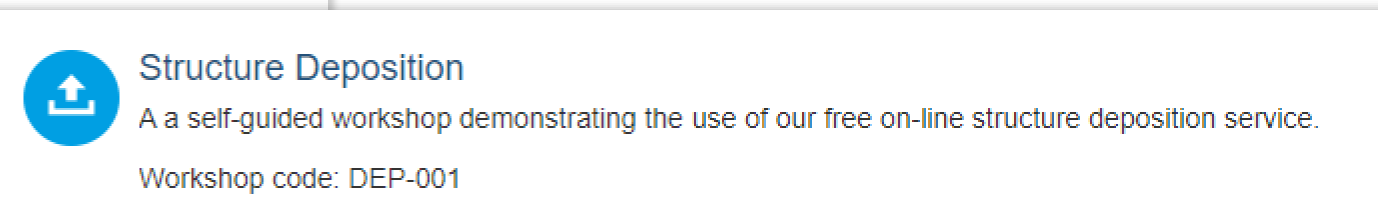
How can I associate raw data stored with another repository to my deposited structures?

Solution

The CCDC does not currently store raw data files. However, if you have deposited your raw data files with another repository and have a DOI link for this then you can associate this DOI link with the relevant files you have deposited with CCDC. You can do this during the deposition process by adding the DOI at the 'Enhance Data' stage of the process.

For any structures already deposited at CCDC it is possible to add the link to associated data from the 'My Structures' service. For any structure, select the Details button to access the 'My Structure Details' view and click the 'New Associated Data' button in the 'Associated Data' section to add the DOI.

The raw data DOIs will be linked from Access Structures and WebCSD once your data is published and in time will be embedded into CIFs downloaded from the CCDC and included in the metadata we send to DataCite.



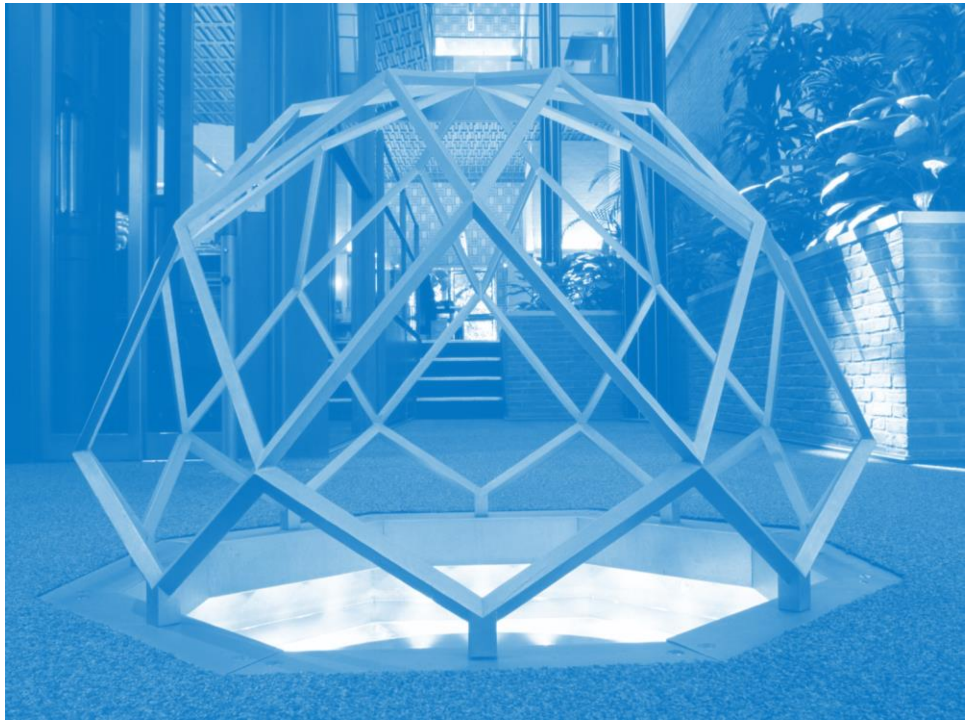
Structure Deposition

A a self-guided workshop demonstrating the use of our free on-line structure deposition service.

Workshop code: DEP-001

https://www.ccdc.cam.ac.uk/support-and-resources/support/case/?caseid=e1fc6d58-e3b7-e711-b787-005056977c87

Thank you for listening



Acknowledgements

Stephanie Boer, **Jason Price** (Australian Synchrotron), **Mike Hoyland**, **Brian McMahon**, **Peter Strickland** (IUCr), **Vasily Bunakov**, **Brian Matthews** (STFC), **Suzanna Ward**, **Seth Wiggin** (CCDC), **Simon Coles**, **John Helliwell**, **Mark Warren**, and **Amy Sarjeant**