The 29th European Crystallographic Meeting
ECM29 ROVINJ 2015 CROATIA
August 23-28, 2015
Rovinj, Croatia

CSD 50 1965 - 2015

CCDC Metadata Initiatives

Suzanna Ward, Ian Bruno, Colin Groom

Workshop on Metadata for raw data from X-ray diffraction and other structural techniques

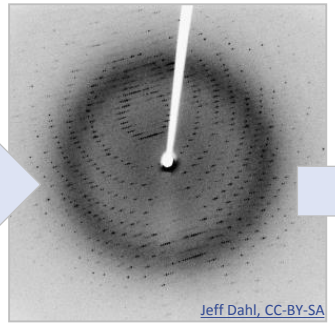
Session IV: Data in the Wider World - From Laboratory to Database

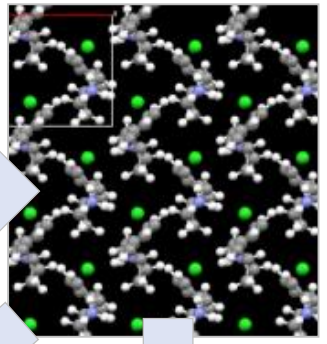
Sunday 23rd August 9:35-10:00

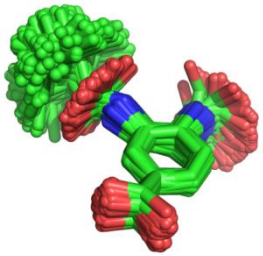
From data to knowledge

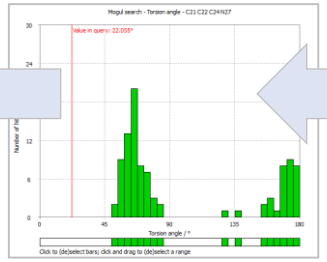
Experimental Data

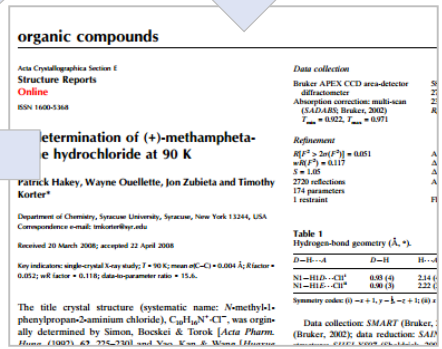
$C_{10}H_{16}N^{+},Cl^{-}$







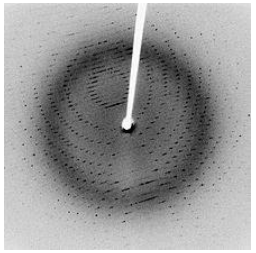




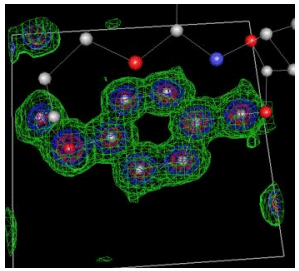
Structural Knowledge

From raw data to model

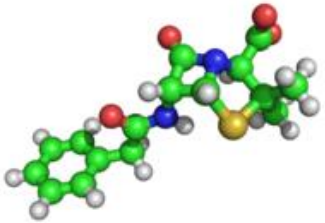
Raw	
Image data	Positions and intensities of diffracted beams



Processed	
Structure factor data	Quantities needed to derive electron density

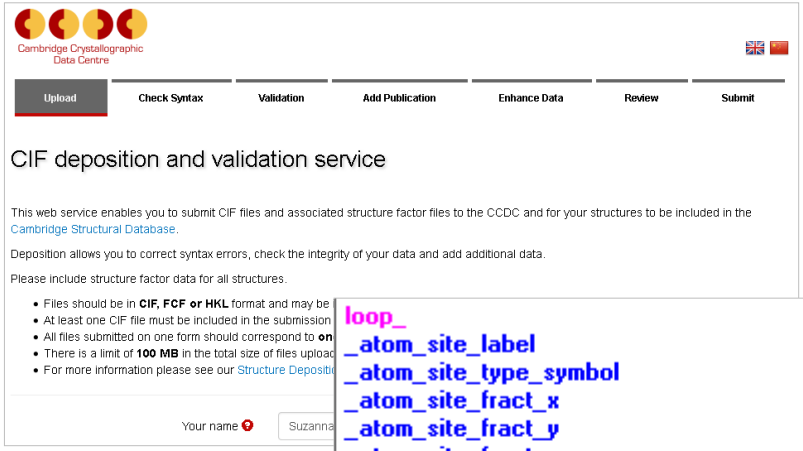


Derived	
Atom coordinates	Modelled and refined from electron density

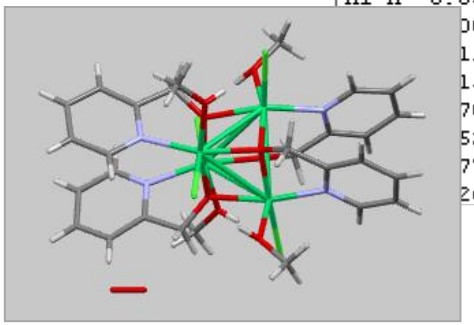


Primary emphasis has typically been on deposition of **derived** data only

The Cambridge Structural Database

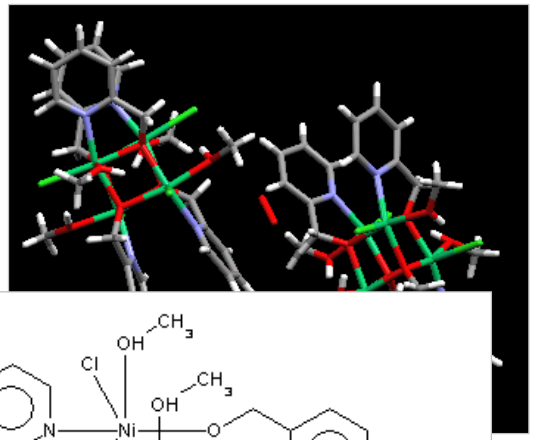


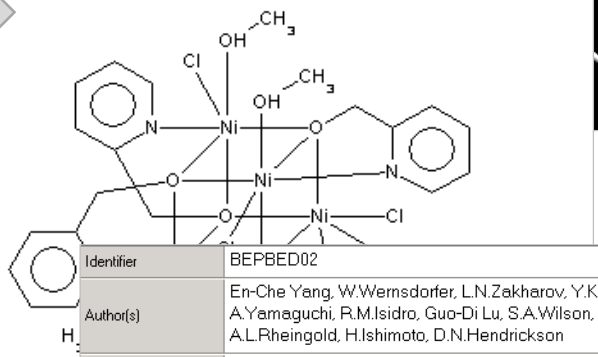

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
C1  C  0.31594(37)  0.75375(20)  0.70189(16)
C2  C  0.19196(41)  0.77066(23)  0.76436(17)
O1  O  0.22259(32)  0.80293(21)  0.83459(13)
C3  C  0.03051(40)  0.74837(26)  0.73926(18)
H1  H -0.05699(40)  0.75910(26)  0.77720(18)
   00500(38)  0.71281(22)  0.66463(17)
   11609(38)  0.69941(22)  0.65036(17)
   1523(28)  0.69499(17)  0.60921(13)
   7057(33)  0.71613(20)  0.62788(16)
   5291(33)  0.70477(20)  0.58724(16)
   7764(37)  0.65275(21)  0.52604(17)
   2644(37)  0.69329(21)  0.48496(17)
```



Deposited CIF

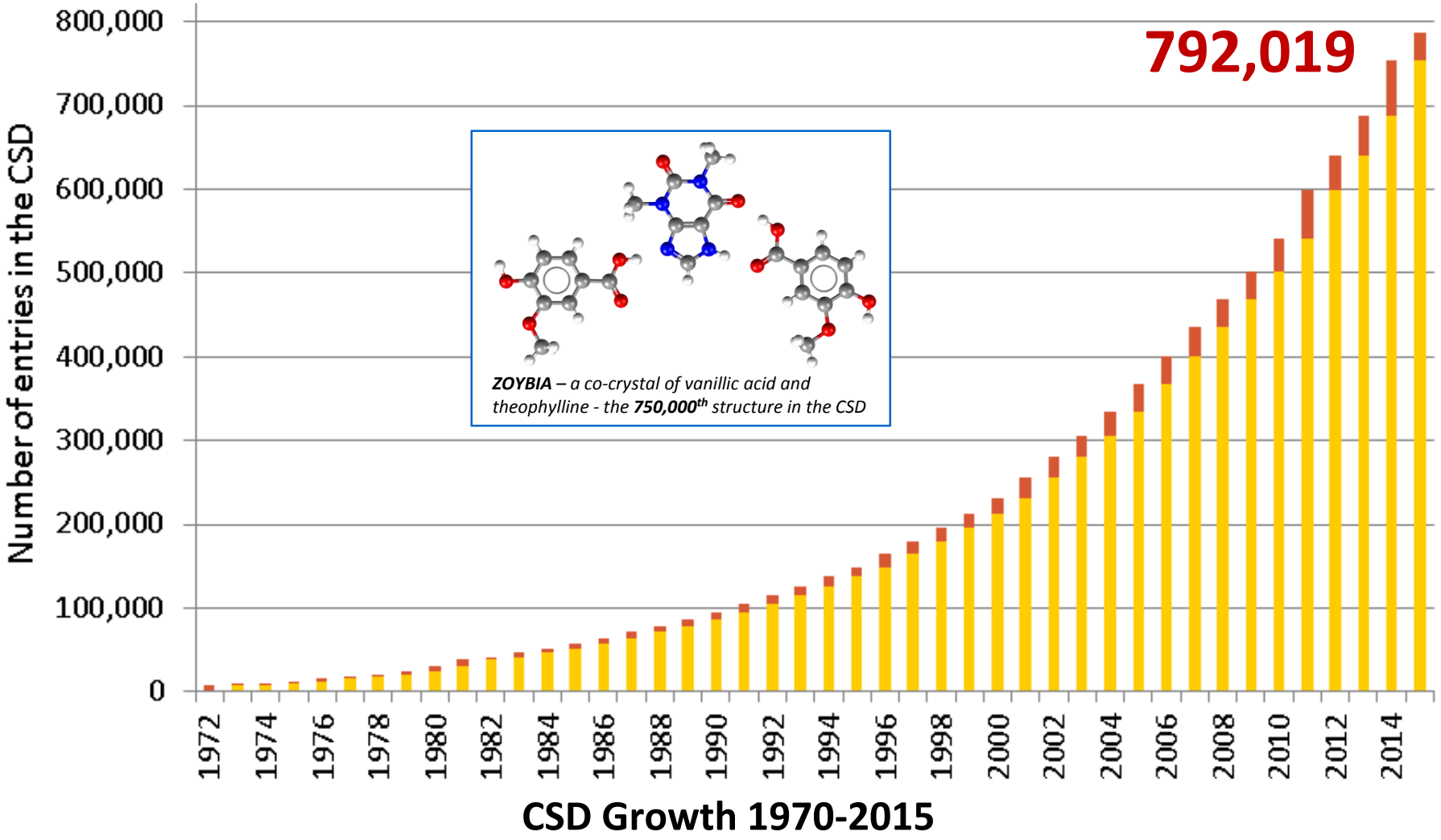
CSD Entry





Identifier	BEPBED02
Author(s)	En-Che Yang, W.Wernsdorfer, L.N.Zakharov, Y.Karaki, A.Yamaguchi, R.M.Isidro, Guo-Di Lu, S.A.Wilson, A.L.Rheingold, H.Ishimoto, D.N.Hendrickson
Literature Reference	*Inorg.Chem.* (2006), **45**, 529, doi:10.1021/ic050093r
Formula	$C_{28}H_{40}Cl_4N_4Ni_4O_8H_2O$
Compound Name	tetrakis(μ3-2-(Oxymethyl)pyridine-N,O,O,O)-tetrachloro-tetraki s(methanol)-tetra-nickel(ii) monohydrate
Synonym	
Space Group	I -4 2 d
Cell Lengths	**a** 16.1421(6) **b** 16.1421(6) **c** 29.4689(14)
Cell Angles	**α** 90 **β** 90 **γ** 90
R-Factor (%)	5.21
Disorder	The water molecule is disordered by symmetry.

The Cambridge Structural Database



CSD Growth 1970-2015

Metadata

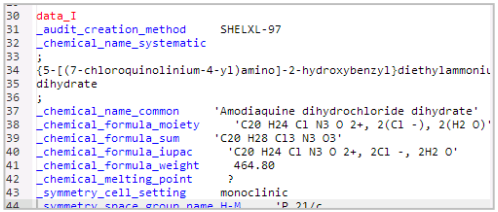
A set of data that describes and gives information about other data

- Data that describes the substance studied
  - Important for discovery, data analysis and mining
- Data that describes the dataset as a whole
  - important for provenance (structure of record), attribution (citation)
  - Data that describes the associated publication
- Data that describes the experiment
  - where was it done, how it was done, who did it, who funded it

Recent metadata developments at CCDC

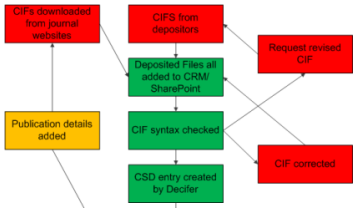
Web Deposit — Launched 2014

- Updated Web Interface with greater interactivity
- Aim to capture as much input from depositor as possible



CSD-Xpedite — Deployed 2013

- Manages depositions and creation of CSD entries
- Designed to adapt to changing internal and external needs



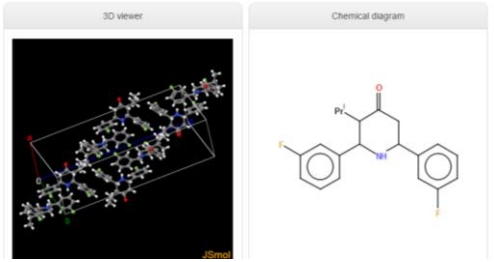
CCDC DOIs — Implemented 2014

- Unambiguous and persistent identification of datasets
- Foundation for formalising data citation and interoperability

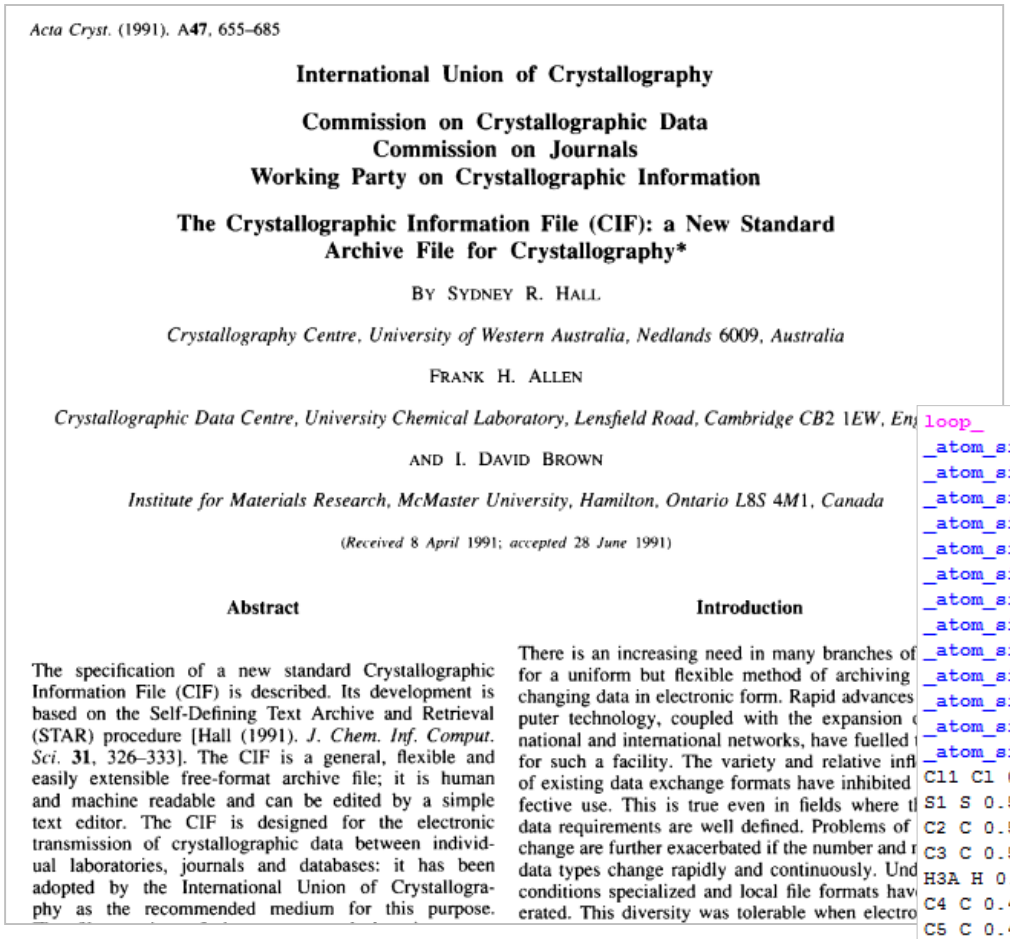
10.5517/CCPHZ37

Get Structures — Launched 2014

- Update to previous “structure request” service
- Enhanced web interface for accessing individual structures



Crystallographic Information Framework (CIF)



Acta Cryst. (1991). A**47**, 655–685

International Union of Crystallography

Commission on Crystallographic Data
Commission on Journals
Working Party on Crystallographic Information

The Crystallographic Information File (CIF): a New Standard Archive File for Crystallography*

BY SYDNEY R. HALL

Crystallography Centre, University of Western Australia, Nedlands 6009, Australia

FRANK H. ALLEN

Crystallographic Data Centre, University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, Eng

AND I. DAVID BROWN

Institute for Materials Research, McMaster University, Hamilton, Ontario L8S 4M1, Canada

(*Received* 8 *April* 1991; *accepted* 28 *June* 1991)

Abstract

The specification of a new standard Crystallographic Information File (CIF) is described. Its development is based on the Self-Defining Text Archive and Retrieval (STAR) procedure [Hall (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333]. The CIF is a general, flexible and easily extensible free-format archive file; it is human and machine readable and can be edited by a simple text editor. The CIF is designed for the electronic transmission of crystallographic data between individual laboratories, journals and databases: it has been adopted by the International Union of Crystallography as the recommended medium for this purpose.

Introduction

There is an increasing need in many branches of
for a uniform but flexible method of archiving
changing data in electronic form. Rapid advances
puter technology, coupled with the expansion
national and international networks, have fuelled
for such a facility. The variety and relative infl
of existing data exchange formats have inhibited
fective use. This is true even in fields where t
data requirements are well defined. Problems of
change are further exacerbated if the number and
data types change rapidly and continuously. Und
conditions specialized and local file formats hav
erated. This diversity was tolerable when electro

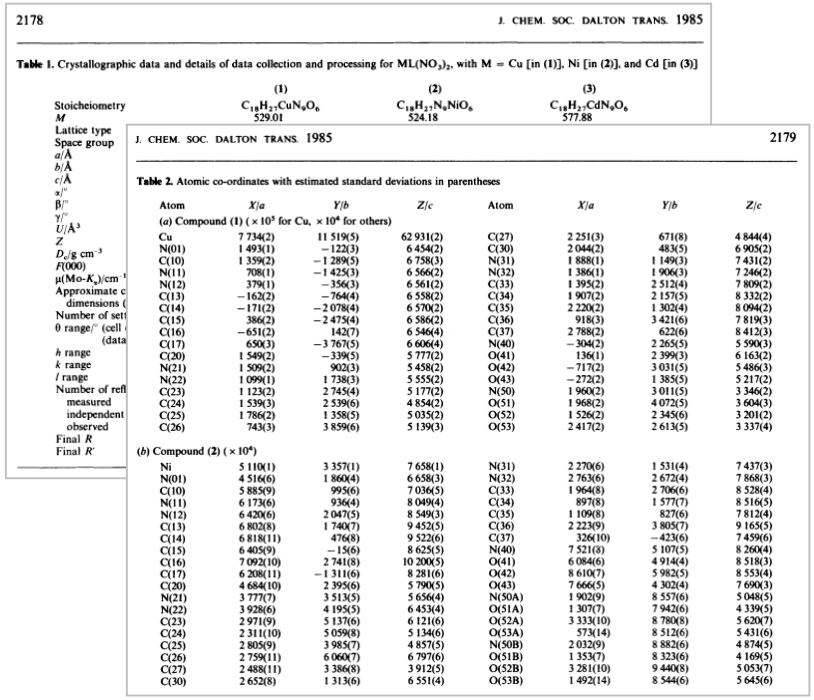
```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disorder_group
C11 C1 0.5993(2) 1.0007(7) 0.8131(17) 0.044(3) Uani 0.50 1 d PDU A 1
S1 S 0.5321(3) 0.8260(6) 0.9322(3) 0.0327(11) Uani 0.50 1 d PDU A 1
C2 C 0.5529(4) 0.8802(9) 0.8184(9) 0.029(4) Uani 0.50 1 d PDU A 1
C3 C 0.5286(7) 0.8174(18) 0.7440(7) 0.031(4) Uani 0.50 1 d PDU A 1
H3A H 0.5350 0.8343 0.6771 0.037 Uiso 0.50 1 calc PR A 1
C4 C 0.4918(8) 0.7220(19) 0.7783(8) 0.027(4) Uani 0.50 1 d PDU A 1
C5 C 0.4900(6) 0.7171(14) 0.8779(9) 0.029(4) Uani 0.50 1 d PDU A 1
C12 C1 0.3202(2) 0.4982(6) 1.0830(5) 0.0586(15) Uani 0.50 1 d PDU A 1
S2 S 0.38755(19) 0.6658(5) 0.9578(5) 0.0400(10) Uani 0.50 1 d PDU A 1
```

- A standard format for crystallographic data
  - derived results
  - raw and processed data
  - experimental conditions
  - publication data

Acta Cryst., **1991**, A47, 655 doi:10.1107/S010876739101067X

The growth of electronic deposition

Hand-typed tables of coordinates



2178 J. CHEM. SOC. DALTON TRANS. 1985

Table 1. Crystallographic data and details of data collection and processing for $ML(NO_3)_2$, with M = Cu [in (1)], Ni [in (2)], and Cd [in (3)]

	(1)	(2)	(3)
Stoicheiometry	$C_{18}H_{27}CuN_9O_6$	$C_{18}H_{27}N_9NiO_6$	$C_{18}H_{27}CdN_9O_6$
M	529.01	524.18	577.88
Lattice type			
Space group			
a/Å			
b/Å			
c/Å			
α/°			
β/°			
γ/°			
U/Å^3			
Z			
D_c/g cm^{-3}			
F(000)			
μ(Mo-K_α)/cm^{-1}			
Approximate c			
dimensions (			
Number of sett			
θ range/° (cell			
(data			
h range			
k range			
l range			
Number of refl			
measured			
independent			
observed			
Final *R*			
Final *R*′			

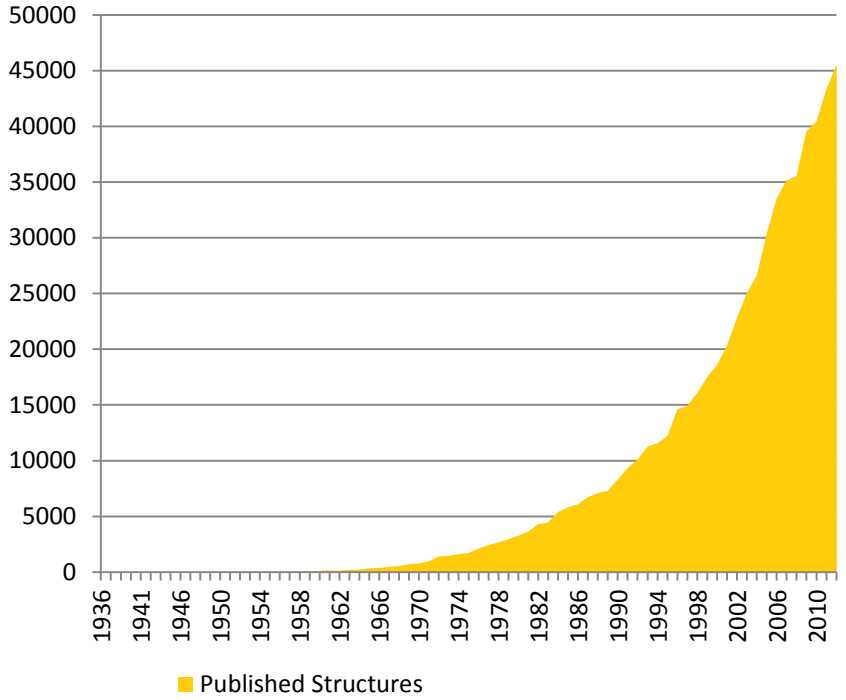
J. CHEM. SOC. DALTON TRANS. 1985 2179

Table 2. Atomic co-ordinates with estimated standard deviations in parentheses

Atom	*X*/*a*	*Y*/*b*	*Z*/*c*	Atom	*X*/*a*	*Y*/*b*	*Z*/*c*
(*a*) Compound (1) (× 10^5 for Cu, × 10^4 for others)							
Cu	7 734(2)	11 519(5)	62 931(2)	C(27)	2 251(3)	671(8)	4 844(4)
N(01)	1 493(1)	−122(3)	6 454(2)	C(30)	2 044(2)	483(5)	6 905(2)
C(10)	1 359(2)	−1 289(5)	6 758(3)	N(31)	1 888(1)	1 149(3)	7 431(2)
N(11)	708(1)	−1 425(3)	6 566(2)	N(32)	1 386(1)	1 906(3)	7 246(2)
N(12)	379(1)	−356(3)	6 561(2)	C(33)	1 395(2)	2 512(4)	7 809(2)
C(13)	−162(2)	−764(4)	6 558(2)	C(34)	1 907(2)	2 157(5)	8 332(2)
C(14)	−171(2)	−2 078(4)	6 570(2)	C(35)	2 220(2)	1 302(4)	8 094(2)
C(15)	386(2)	−2 475(4)	6 586(2)	C(36)	918(3)	3 421(6)	7 819(3)
C(16)	−651(2)	142(7)	6 546(4)	C(37)	2 788(2)	622(6)	8 412(3)
C(17)	650(3)	−3 767(5)	6 606(4)	N(40)	−304(2)	2 265(5)	5 590(3)
C(20)	1 549(2)	−339(5)	5 777(2)	O(41)	136(1)	2 399(3)	6 163(2)
N(21)	1 509(2)	902(3)	5 458(2)	O(42)	−717(2)	3 031(5)	5 486(3)
N(22)	1 099(1)	1 738(3)	5 555(2)	O(43)	−272(2)	1 385(5)	5 217(2)
C(23)	1 123(2)	2 745(4)	5 177(2)	N(50)	1 960(2)	3 011(5)	3 346(2)
C(24)	1 539(3)	2 539(6)	4 854(2)	O(51)	1 968(2)	4 072(5)	3 604(3)
C(25)	1 786(2)	1 358(5)	5 035(2)	O(52)	1 526(2)	2 345(6)	3 201(2)
C(26)	743(3)	3 859(6)	5 139(3)	O(53)	2 417(2)	2 613(5)	3 337(4)
(*b*) Compound (2) (× 10^4)							
Ni	5 110(1)	3 357(1)	7 658(1)	N(31)	2 270(6)	1 531(4)	7 437(3)
N(01)	4 516(6)	1 860(4)	6 658(3)	N(32)	2 763(6)	2 672(4)	7 868(3)
C(10)	5 885(9)	995(6)	7 036(5)	C(33)	1 964(8)	2 706(6)	8 528(4)
N(11)	6 173(6)	936(4)	8 049(4)	C(34)	897(8)	1 577(7)	8 516(5)
N(12)	6 420(6)	2 047(5)	8 549(3)	C(35)	1 109(8)	827(6)	7 812(4)
C(13)	6 802(8)	1 740(7)	9 452(5)	C(36)	2 223(9)	3 805(7)	9 165(5)
C(14)	6 818(11)	476(8)	9 522(6)	C(37)	326(10)	−423(6)	7 459(6)
C(15)	6 405(9)	−15(6)	8 625(5)	N(40)	7 521(3)	5 107(5)	8 260(4)
C(16)	7 092(10)	2 741(8)	10 200(5)	O(41)	6 084(6)	4 914(4)	8 518(3)
C(17)	6 208(11)	−1 311(6)	8 281(6)	O(42)	8 610(7)	5 982(5)	8 553(4)
C(20)	4 684(10)	2 395(6)	5 790(5)	O(43)	7 666(5)	4 302(4)	7 690(3)
N(21)	3 777(7)	3 513(5)	5 656(4)	N(50A)	1 902(9)	8 557(6)	5 048(5)
N(22)	3 928(6)	4 195(5)	6 453(4)	O(51A)	1 307(7)	7 942(6)	4 339(5)
C(23)	2 971(9)	5 137(6)	6 121(6)	O(52A)	3 333(10)	8 780(8)	5 620(7)
C(24)	2 311(10)	5 059(8)	5 134(6)	O(53A)	573(14)	8 512(6)	5 431(6)
C(25)	2 805(9)	3 985(7)	4 857(5)	N(50B)	2 032(9)	8 882(6)	4 874(5)
C(26)	2 759(11)	6 060(7)	6 797(6)	O(51B)	1 353(7)	8 323(6)	4 169(5)
C(27)	2 488(11)	3 386(8)	3 912(5)	O(52B)	3 281(10)	9 440(8)	5 053(7)
C(30)	2 652(8)	1 313(6)	6 551(4)	O(53B)	1 492(14)	8 544(6)	5 645(6)

Kleywegt et al. (1985) *J. Chem. Soc., Dalton Trans*, 2177-2184
doi:10.1039/DT9850002177

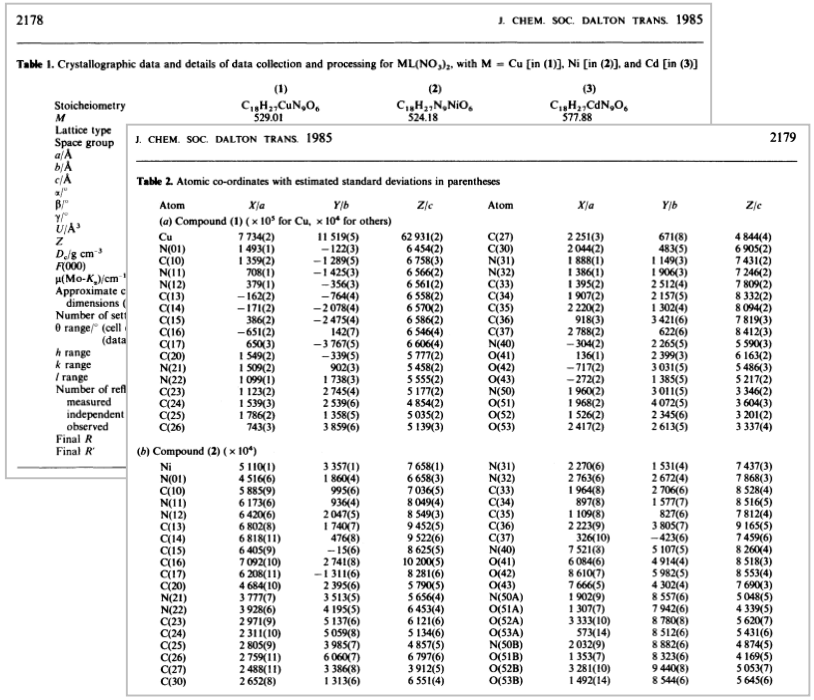
Structures Digitally Deposited cf Published



CSD entries with 3D coordinates (1936-2013)
CCDC Number used as an indication of digital deposit

The growth of electronic deposition

Hand-typed tables of coordinates



2178 J. CHEM. SOC. DALTON TRANS. 1985

Table 1. Crystallographic data and details of data collection and processing for $ML(NO_3)_2$, with M = Cu [in (1)], Ni [in (2)], and Cd [in (3)]

	(1)	(2)	(3)
Stoicheiometry	$C_{18}H_{27}CuN_9O_6$	$C_{18}H_{27}N_9NiO_6$	$C_{18}H_{27}CdN_9O_6$
M	529.01	524.18	577.88
Lattice type			
Space group			
a/Å			
b/Å			
c/Å			
α/°			
β/°			
γ/°			
U/Å^3			
Z			
D_c/g cm^{-3}			
F(000)			
μ(Mo-K_α)/cm^{-1}			
Approximate c			
dimensions (			
Number of sett			
θ range/° (cell			
(data			
h range			
k range			
l range			
Number of refl			
measured			
independent			
observed			
Final *R*			
Final *R*′			

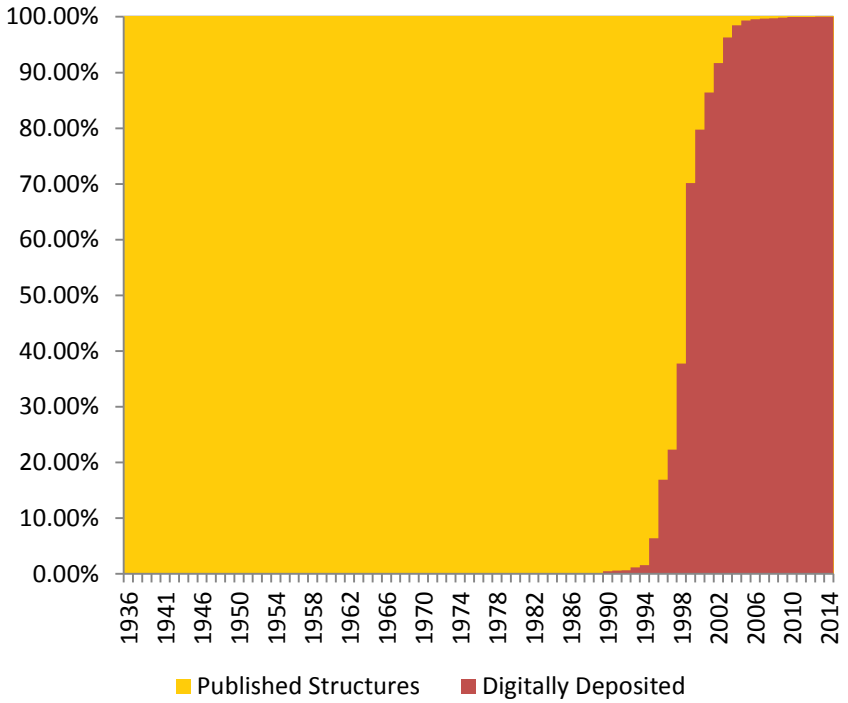
J. CHEM. SOC. DALTON TRANS. 1985 2179

Table 2. Atomic co-ordinates with estimated standard deviations in parentheses

Atom	*X*/*a*	*Y*/*b*	*Z*/*c*	Atom	*X*/*a*	*Y*/*b*	*Z*/*c*
(*a*) Compound (1) (× 10^5 for Cu, × 10^4 for others)							
Cu	7 734(2)	11 519(5)	62 931(2)	C(27)	2 251(3)	671(8)	4 844(4)
N(01)	1 493(1)	−122(3)	6 454(2)	C(30)	2 044(2)	483(5)	6 905(2)
C(10)	1 359(2)	−1 289(5)	6 758(3)	N(31)	1 888(1)	1 149(3)	7 431(2)
N(11)	708(1)	−1 425(3)	6 566(2)	N(32)	1 386(1)	1 906(3)	7 246(2)
N(12)	379(1)	−356(3)	6 561(2)	C(33)	1 395(2)	2 512(4)	7 809(2)
C(13)	−162(2)	−764(4)	6 558(2)	C(34)	1 907(2)	2 157(5)	8 332(2)
C(14)	−171(2)	−2 078(4)	6 570(2)	C(35)	2 220(2)	1 302(4)	8 094(2)
C(15)	386(2)	−2 475(4)	6 586(2)	C(36)	918(3)	3 421(6)	7 819(3)
C(16)	−651(2)	142(7)	6 546(4)	C(37)	2 788(2)	622(6)	8 412(3)
C(17)	650(3)	−3 767(5)	6 606(4)	N(40)	−304(2)	2 265(5)	5 590(3)
C(20)	1 549(2)	−339(5)	5 777(2)	O(41)	136(1)	2 399(3)	6 163(2)
N(21)	1 509(2)	902(3)	5 458(2)	O(42)	−717(2)	3 031(5)	5 486(3)
N(22)	1 099(1)	1 738(3)	5 555(2)	O(43)	−272(2)	1 385(5)	5 217(2)
C(23)	1 123(2)	2 745(4)	5 177(2)	N(50)	1 960(2)	3 011(5)	3 346(2)
C(24)	1 539(3)	2 539(6)	4 854(2)	O(51)	1 968(2)	4 072(5)	3 604(3)
C(25)	1 786(2)	1 358(5)	5 035(2)	O(52)	1 526(2)	2 345(6)	3 201(2)
C(26)	743(3)	3 859(6)	5 139(3)	O(53)	2 417(2)	2 613(5)	3 337(4)
(*b*) Compound (2) (× 10^4)							
Ni	5 110(1)	3 357(1)	7 658(1)	N(31)	2 270(6)	1 531(4)	7 437(3)
N(01)	4 516(6)	1 860(4)	6 658(3)	N(32)	2 763(6)	2 672(4)	7 868(3)
C(10)	5 885(9)	995(6)	7 036(5)	C(33)	1 964(8)	2 706(6)	8 528(4)
N(11)	6 173(6)	936(4)	8 049(4)	C(34)	897(8)	1 577(7)	8 516(5)
N(12)	6 420(6)	2 047(5)	8 549(3)	C(35)	1 109(8)	827(6)	7 812(4)
C(13)	6 802(8)	1 740(7)	9 452(5)	C(36)	2 223(9)	3 805(7)	9 165(5)
C(14)	6 818(11)	476(8)	9 522(6)	C(37)	326(10)	−423(6)	7 459(6)
C(15)	6 405(9)	−15(6)	8 625(5)	N(40)	7 521(3)	5 107(5)	8 260(4)
C(16)	7 092(10)	2 741(8)	10 200(5)	O(41)	6 084(6)	4 914(4)	8 518(3)
C(17)	6 208(11)	−1 311(6)	8 281(6)	O(42)	8 610(7)	5 982(5)	8 553(4)
C(20)	4 684(10)	2 395(6)	5 790(5)	O(43)	7 666(5)	4 302(4)	7 690(3)
N(21)	3 777(7)	3 513(5)	5 656(4)	N(50A)	1 902(9)	8 557(6)	5 048(5)
N(22)	3 928(6)	4 195(5)	6 453(4)	O(51A)	1 307(7)	7 942(6)	4 339(5)
C(23)	2 971(9)	5 137(6)	6 121(6)	O(52A)	3 333(10)	8 780(8)	5 620(7)
C(24)	2 311(10)	5 059(8)	5 134(6)	O(53A)	573(14)	8 512(6)	5 431(6)
C(25)	2 805(9)	3 985(7)	4 857(5)	N(50B)	2 032(9)	8 882(6)	4 874(5)
C(26)	2 759(11)	6 060(7)	6 797(6)	O(51B)	1 353(7)	8 323(6)	4 169(5)
C(27)	2 488(11)	3 386(8)	3 912(5)	O(52B)	3 281(10)	9 440(8)	5 053(7)
C(30)	2 652(8)	1 313(6)	6 551(4)	O(53B)	1 492(14)	8 544(6)	5 645(6)

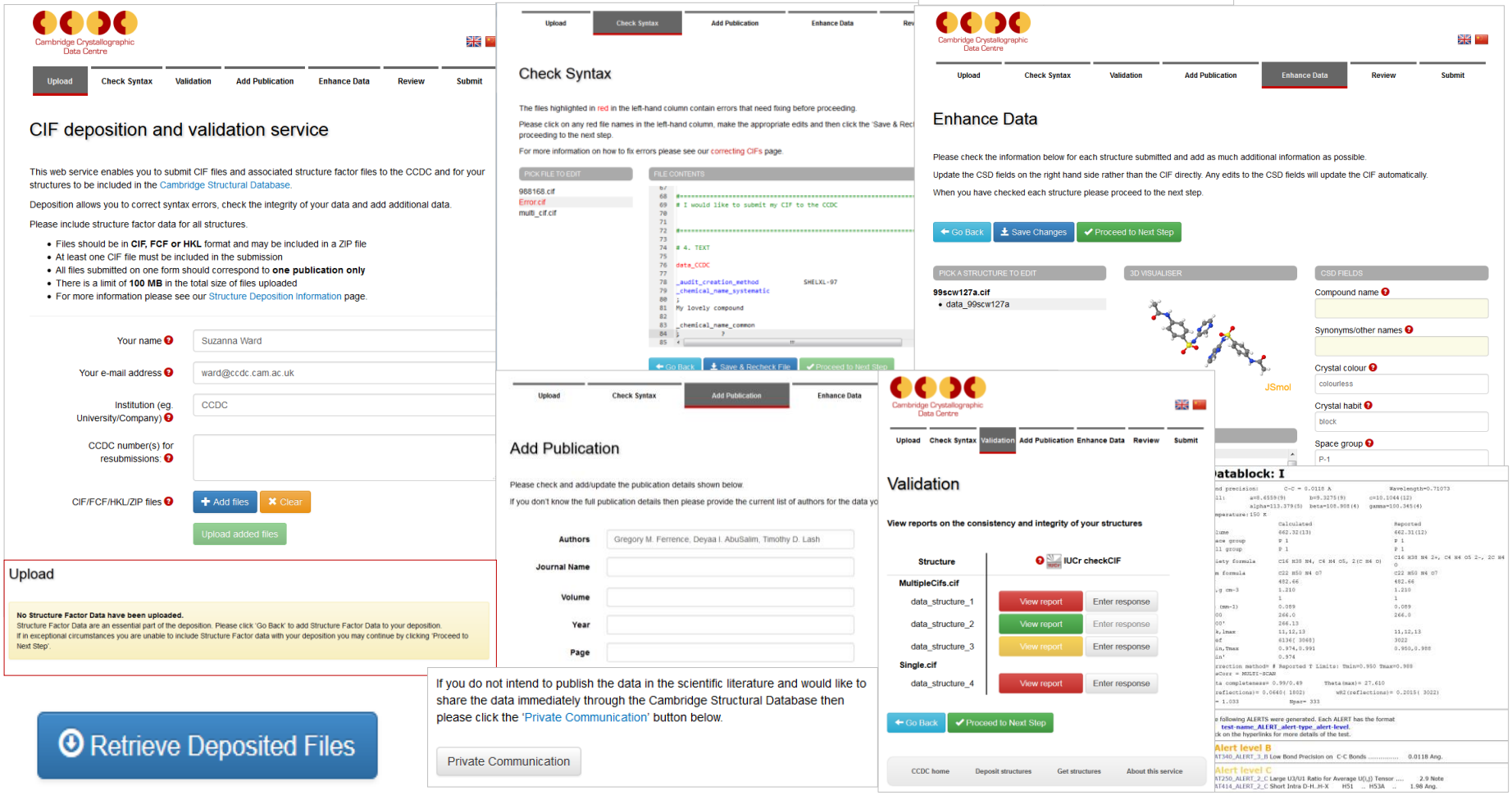
Kleywegt et al. (1985) *J. Chem. Soc., Dalton Trans*, 2177-2184
doi:10.1039/DT9850002177

Structures Digitally Deposited cf Published



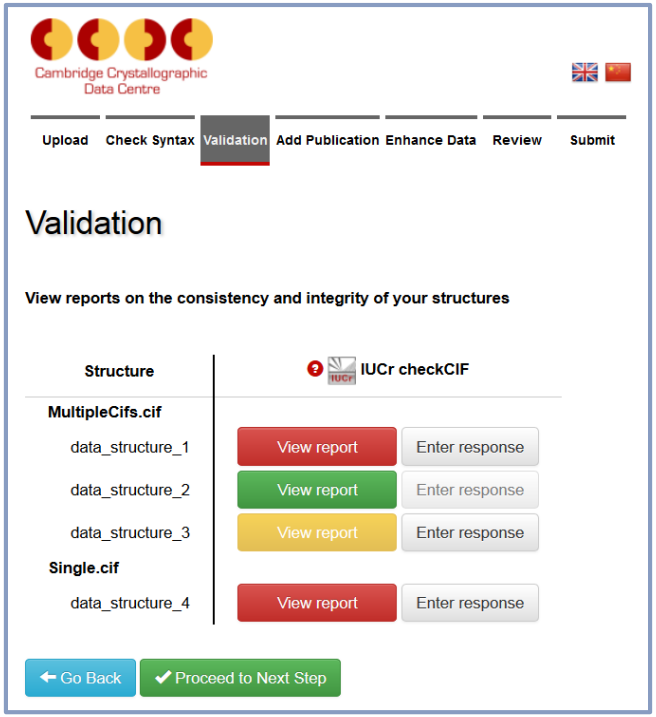
CSD entries with 3D coordinates (1936-2013)
CCDC Number used as an indication of digital deposit

CSD-Deposit


Deposit Structures
Upload your data to the CCDC for inclusion in the Cambridge Structural Database
Cambridge Crystallographic Data Centre
Upload
Check Syntax
Validation
Add Publication
Enhance Data
Review
Submit
CIF deposition and validation service
This web service enables you to submit CIF files and associated structure factor files to the CCDC and for your structures to be included in the Cambridge Structural Database.
Deposition allows you to correct syntax errors, check the integrity of your data and add additional data.
Please include structure factor data for all structures.
Files should be in CIF, FCF or HKL format and may be included in a ZIP file
At least one CIF file must be included in the submission
All files submitted on one form should correspond to one publication only
There is a limit of 100 MB in the total size of files uploaded
For more information please see our Structure Deposition Information page.
Your name
Suzanna Ward
Your e-mail address
ward@ccdc.cam.ac.uk
Institution (eg. University/Company)
CCDC
CCDC number(s) for resubmissions:
CIF/FCF/HKL/ZIP files
Add files
Clear
Upload added files
Upload
No Structure Factor Data have been uploaded.
Structure Factor Data are an essential part of the deposition. Please click 'Go Back' to add Structure Factor Data to your deposition.
If in exceptional circumstances you are unable to include Structure Factor data with your deposition you may continue by clicking 'Proceed to Next Step'.
Check Syntax
The files highlighted in red in the left-hand column contain errors that need fixing before proceeding.
Please click on any red file names in the left-hand column, make the appropriate edits and then click the 'Save & Rec
proceeding to the next step.
For more information on how to fix errors please see our correcting CIFs page.
PICK FILE TO EDIT
FILE CONTENTS
968168.cif
Error.cif
multi_cif.cif
Go Back
Save & Recheck File
Proceed to Next Step
Enhance Data
Please check the information below for each structure submitted and add as much additional information as possible.
Update the CSD fields on the right hand side rather than the CIF directly. Any edits to the CSD fields will update the CIF automatically.
When you have checked each structure please proceed to the next step.
Go Back
Save Changes
Proceed to Next Step
PICK A STRUCTURE TO EDIT
99scw127a.cif
data_99scw127a
3D VISUALISER
JSmol
CSD FIELDS
Compound name
Synonyms/other names
Crystal colour
colourless
Crystal habit
block
Space group
P-1
Add Publication
Please check and add/update the publication details shown below.
If you don't know the full publication details then please provide the current list of authors for the data yo
Authors
Gregory M. Ferrence, Deyaa I. AbuSalim, Timothy D. Lash
Journal Name
Volume
Year
Page
Validation
View reports on the consistency and integrity of your structures
Structure
IUCr checkCIF
MultipleCifs.cif
data_structure_1
data_structure_2
data_structure_3
Single.cif
data_structure_4
View report
Enter response
Go Back
Proceed to Next Step
CCDC home
Deposit structures
Get structures
About this service
Datablock: I
If you do not intend to publish the data in the scientific literature and would like to share the data immediately through the Cambridge Structural Database then please click the 'Private Communication' button below.
Private Communication
Retrieve Deposited Files

IUCr checkCIF and CSD-Deposit

- Deposition to CCDC now includes:
  - Generation of checkCIF reports
    - Available to depositor during submission
    - Stored at CCDC at submission
    - Soon to be available to publishers and referees during peer review process
  - An easy way to embed validation responses into CIF
  - Retrieval of edited CIFs and checkCIF reports
- **Launched July 2015**



Level A	Most likely a serious problem - resolve or explain
Level B	A potentially serious problem, consider carefully
Level C	Check. Ensure it is not caused by an omission or oversight
Level G	General information/check it is not something unexpected

http://www.ccdc.cam.ac.uk/NewsandEvents/News/Pages/NewsItem.aspx?newsid=40

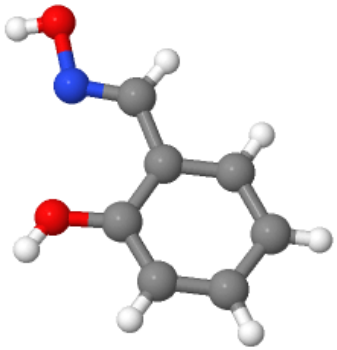
Capturing additional metadata

PICK A STRUCTURE TO EDIT

Example_1.cif
- data_5_93GPa

Example_2.cif
- data_tBu10kbar

3D VISUALISER



JSmol

CIF

```
1  data_5_93GPa
2
3  _audit_creation_method           CRYSTALS_ver_12.80
4  _chemical_name_systematic        '2-hydroxybenzaldehyde oxime'
5  _chemical_name_common            salicylaldoxime
6  _chemical_formula_moiety         'C7 H7 N1 O2'
7  _chemical_formula_sum            'C7 H7 N1 O2'
8  _chemical_formula_weight         137.14
9  _chemical_melting_point          59-61
10 _symmetry_cell_setting           Monoclinic
11 _symmetry_space_group_name_H-M   'P 1 21/n 1 '
12 _symmetry_space_group_name_Hall  '-P 2yn'
13 loop_
```

CSD FIELDS

Compound name
2-hydroxybenzaldehyde oxime

Synonyms/other names
salicylaldoxime

Crystal colour

Crystal habit
block

Space group
P 1 21/n 1

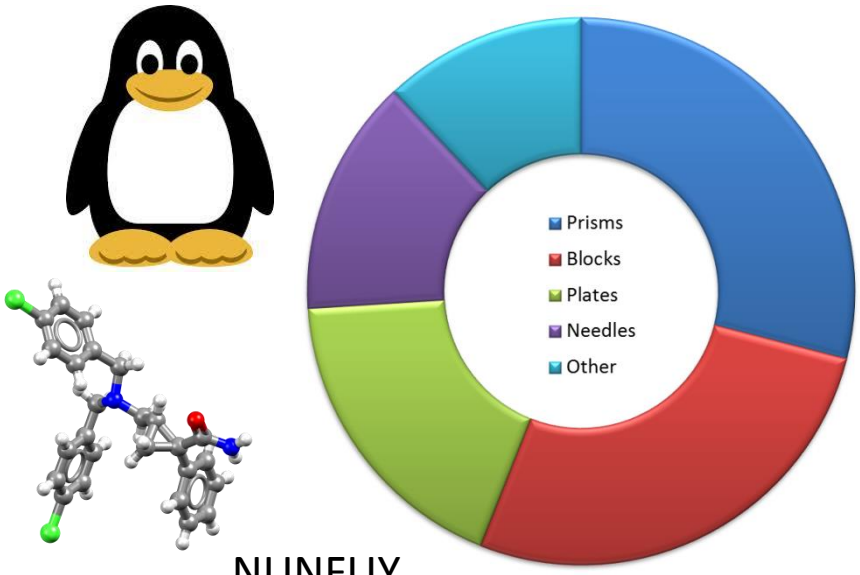
Study temperature (K)
293

Formula moiety
C7 H7 N1 O2

Formula sum
C7 H7 N1 O2

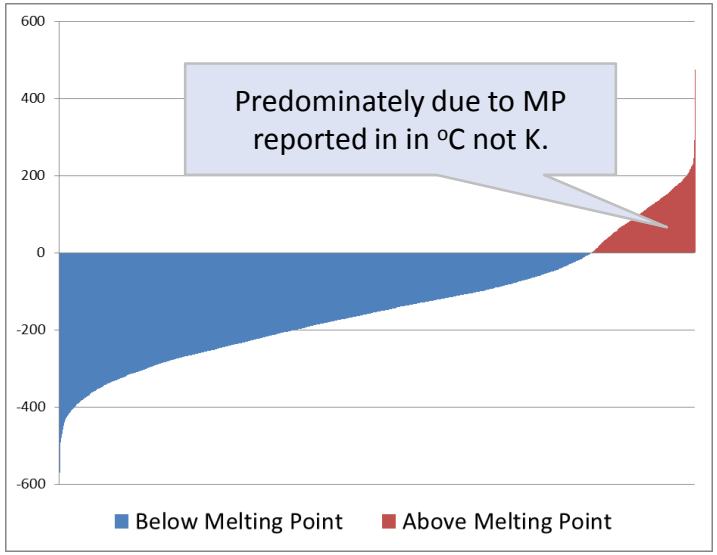
How you capture metadata is important

- Important to capture metadata in a semantic or defined way.
- Free text is not necessarily a good thing
- There can also be confusion over meaning for example the morphology as grown or as cut?



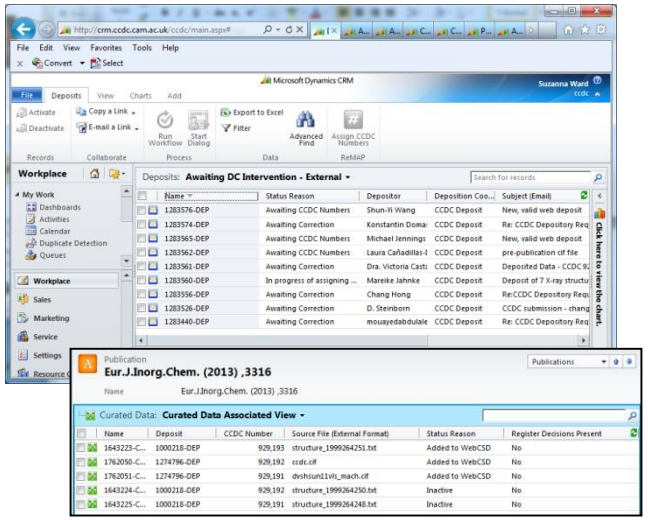
NUNFUY

Study Temperature relative to Melting Point



Plot based on data as initially deposited and before validation at CCDC

CSD-Xpedite



- Syntax errors
- Inaccuracies
- Revisions
- Republications
- Timely release

Over 70,000 structures deposited annually

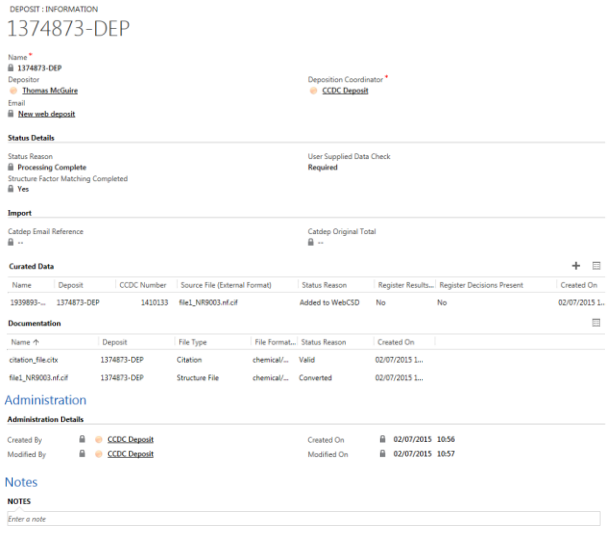
The CCDC's internal Systems automate most steps involved in processing deposited CIFs but manual intervention sometimes needed.

CSD-Xpedite- Metadata used internally

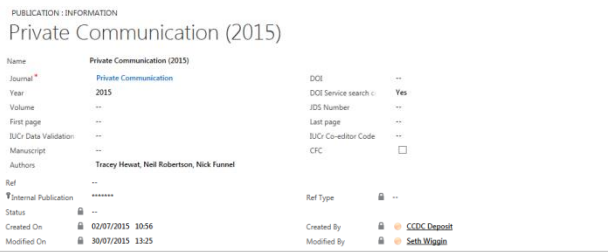
CSD-Xpedite

- Deployed 2013
- Based on CRM
- Used to manage all the transactions that go on behind the scenes
- Number of different entities/record types
- Each entity has metadata associated with it
- Entities include contact, deposit, CSD entry, journal, publication, documentation

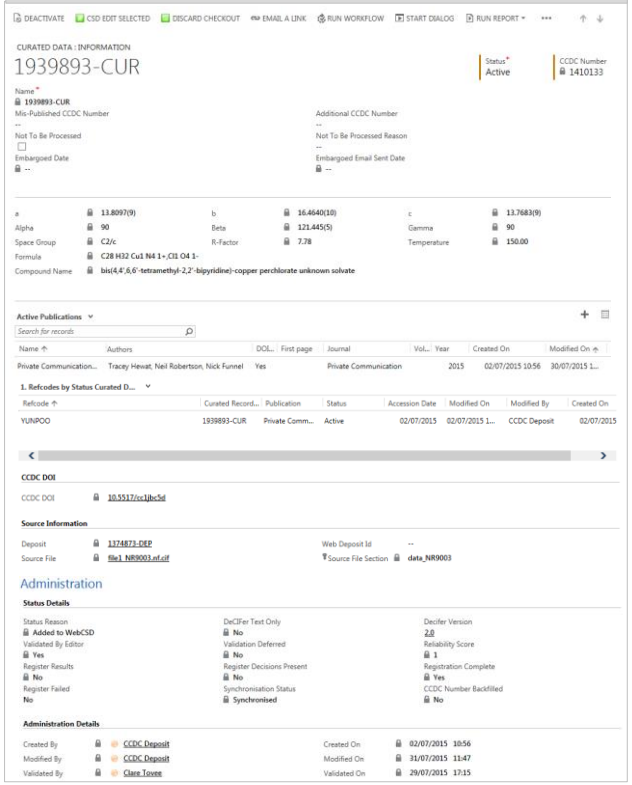
Deposition metadata



Publication metadata



CSD entry metadata

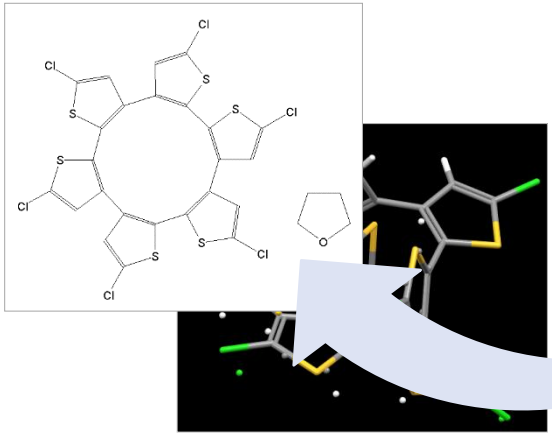


Assignment of chemistry

- Automated processes based on structural knowledge
- Manual validation by expert editorial staff

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$

```
loop_
_atom_site_label
_atom_site_type_symbol
_atom_site_fract_x
_atom_site_fract_y
_atom_site_fract_z
_atom_site_U_iso_or_equiv
_atom_site_adp_type
_atom_site_occupancy
_atom_site_symmetry_multiplicity
_atom_site_calc_flag
_atom_site_refinement_flags
_atom_site_disorder_assembly
_atom_site_disorder_group
C11 Cl 0.5993(2) 1.0007(7) 0.8131(17) 0.044(3) Uani 0.50 1 d PDU A 1
S1 S 0.5321(3) 0.8260(6) 0.9322(3) 0.0327(11) Uani 0.50 1 d PDU A 1
C2 C 0.5529(4) 0.8802(9) 0.8184(9) 0.029(4) Uani 0.50 1 d PDU A 1
C3 C 0.5286(7) 0.8174(18) 0.7440(7) 0.031(4) Uani 0.50 1 d PDU A 1
H3A H 0.5350 0.8343 0.6771 0.037 Uiso 0.50 1 calc PR A 1
C4 C 0.4918(8) 0.7220(19) 0.7783(8) 0.027(4) Uani 0.50 1 d PDU A 1
C5 C 0.4900(6) 0.7171(14) 0.8779(9) 0.029(4) Uani 0.50 1 d PDU A 1
C12 Cl 0.3202(2) 0.4982(6) 1.0830(5) 0.0586(15) Uani 0.50 1 d PDU A 1
S2 S 0.38755(19) 0.6658(5) 0.9578(5) 0.0400(10) Uani 0.50 1 d PDU A 1
```

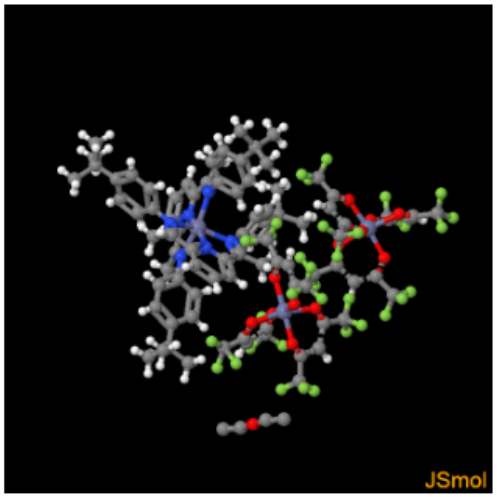


Assignment of chemistry is important for discovery, data mining , analysis and interoperability

Data that describes the substance

ACEXUB : mer-(bis(2,6-bis(1-(4-t-Butylphenylimino)ethyl)pyridine)-zinc(ii))
bis(tris(hexafluoroacetylacetonato)zincate(ii)) diethyl ether solvate
Space Group: C2/c, Cell: *a* 42.138(2)Å *b* 19.4444(10)Å *c* 26.255(2)Å, *α* 90° *β* 110.969(10)° *γ* 90°

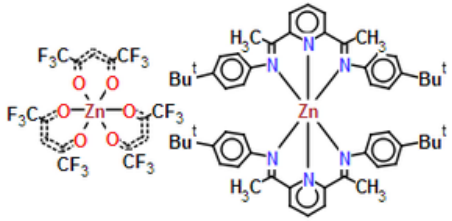
3D viewer



H | Disorder | Menu | Open

Style	Labels	Packing	Measure
Ball and Stick	No Labels	None	None

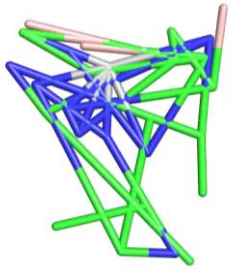
Chemical diagram



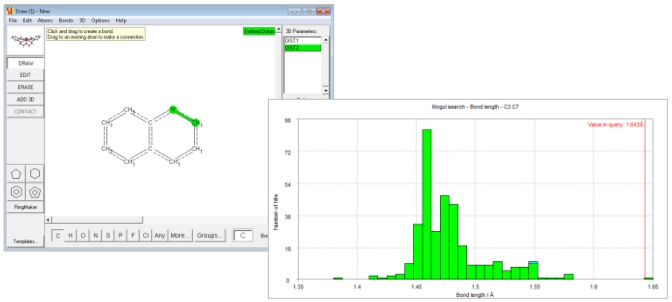
View group symbols key

ACD/Labs

Compound names based on IUPAC conventions using ACDname and manual curation

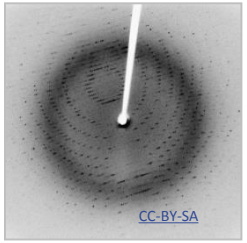


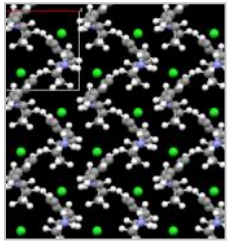
2D chemical diagram generated from existing CSD entries, CCDC diagram generation software and manual curation

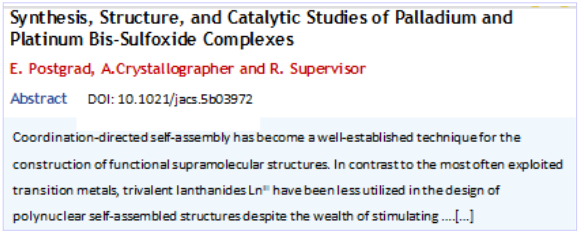


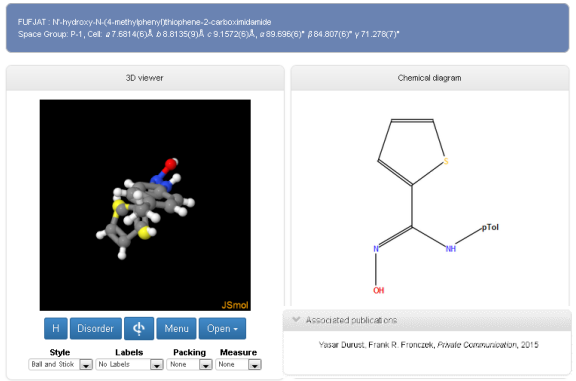
Data that describes the dataset as a whole

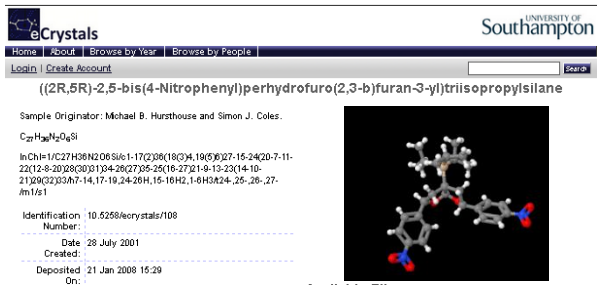
- Important for provenance and attribution
- Includes:
  - Publication data
    - Could be in the scientific literature and/or repository
  - Authorship data
    - Creation of the sample, data collection, data analysis











Metadata - Associating CSD data to publications

Metadata exchanges

- Automated workflows with:

  - Publishers - Third Parties

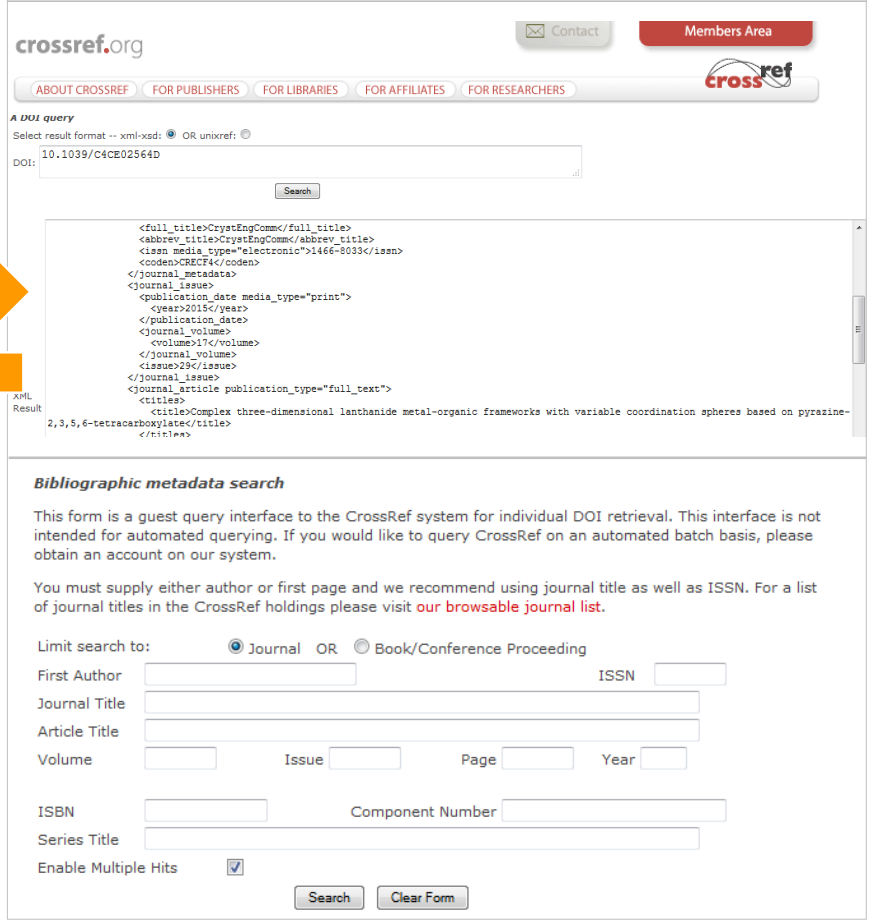
Publication metadata at CCDC

Journal *	CrystEngComm	DOI	10.1039/C4CE02564D
Year	2015	DOI Service search c	Yes
Volume	17	JDS Number	--
First page	5377	Last page	--
IUCr Data Validation	--	IUCr Co-editor Code	--
Manuscript	--	CFC	☐
Authors	Conrad W. Ingram, Geoffrey Kibakaya, John Bacsa, Stephan R. Mathis,II, Alvin A. Holder, Varma H. Rambaran, Brandon Dennis, Esmeralda Castaneda, Julianne S. Robbins, Z. John Zhang		

Metadata from publishers

```
<doi>doi:10.1107/S2056989015014383</doi>
– <title>
    Crystal structure of 6,6,12,12-tetrachlorotricyclo[8.2.0.04,7]dodecane-5,11-dione
  </title>
  <journal>urn:issn:2056-9890</journal>
  <cnor>xu5862</cnor>
  <author>Turan Akın, E.</author>
  <author>Hökelek, T.</author>
  <volume>71</volume>
  <issue>9</issue>
  <firstpage>1000</firstpage>
  <lastpage>1002</lastpage>
  <published>2015-08-06</published>
  <ccdc>1415865</ccdc>
```

Publication metadata from third parties



Metadata enabling referee access to data

From: 418e7b.0b9f8125@editorialmanager.com
Sent: 04 March 2015 03:45
Subject: Manuscript [redacted] for review

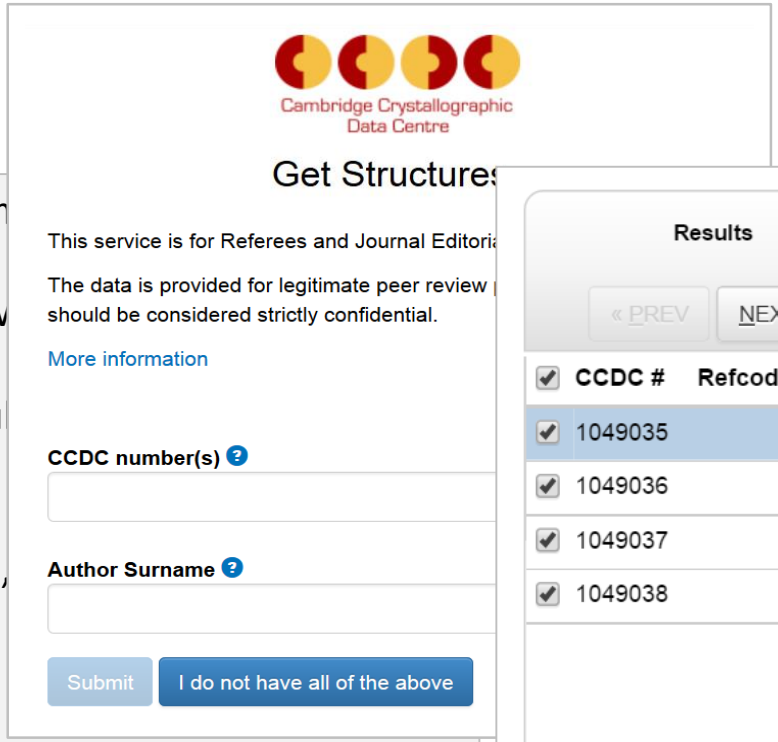
In view of your expertise I would be very grateful
following manuscript …

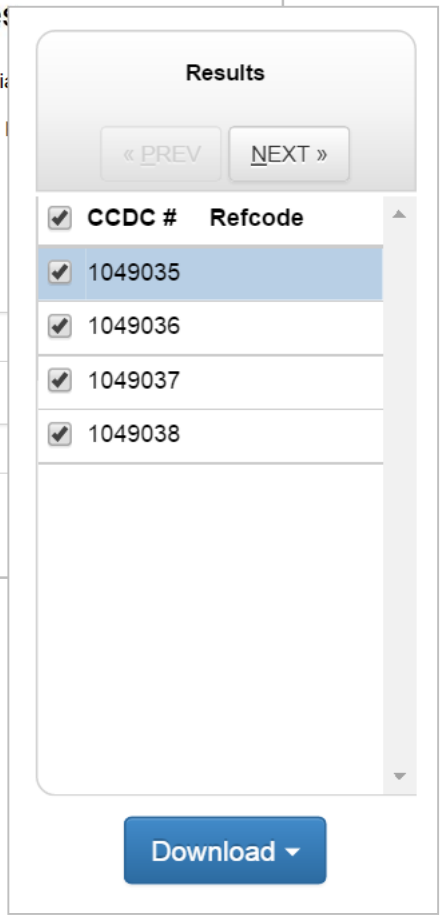
If this paper includes new structural information,
available from one of the following databases:

CCDC (organic and metal-organic compounds):
http://www.ccdc.cam.ac.uk/data_request/referee.access

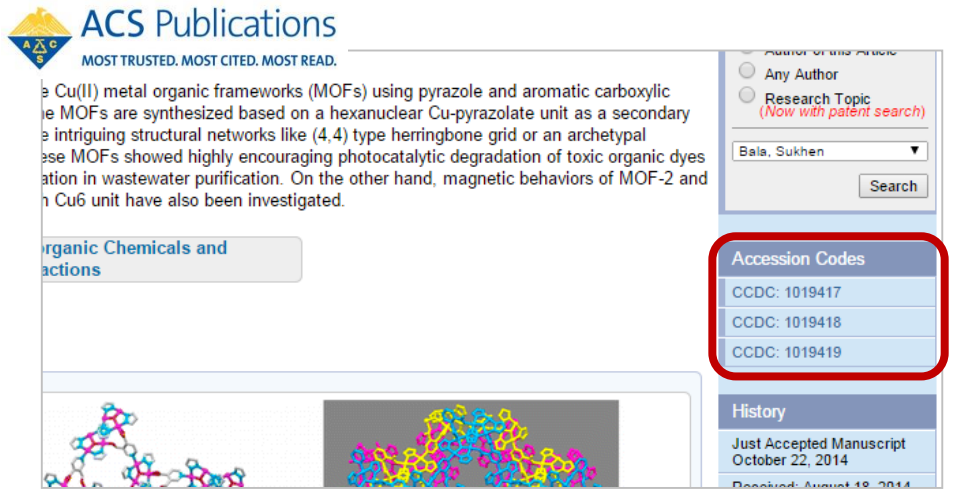
Fiz Karlsruhe (inorganic crystal structures): http://www.fiz-informationsdienste.de/en/DB/icsd/depot.html

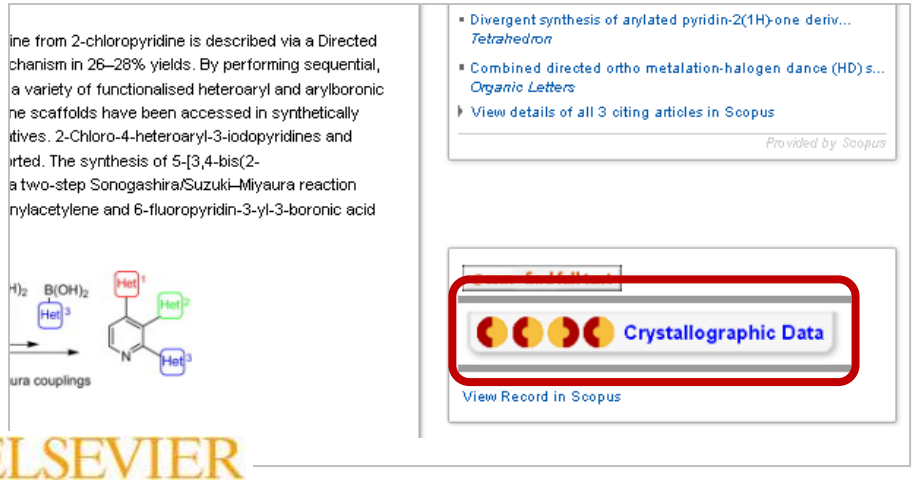
PDB (proteins): http://www.rcsb.org/pdb/





Metadata used to link articles and data



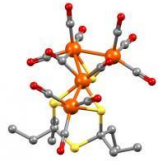


Links are in place for ACS, RSC, Elsevier, Wiley and IUCr journals

CCDC crystal structure DOIs

- Over 560,000 dataset DOIs registered since April 2014
- DataCite metadata includes related article DOIs
- Foundation for formalising data citation and interoperability
- Data Citation Principles - Data should be considered legitimate, citable products of research…

DataCite



DC1 Data Citation Principles

10.5517/CCPHZ37

https://www.force11.org/datacitation

Synthesis, Structure, and Catalytic Studies of Palladium and Platinum Bis-Sulfoxide Complexes

E. Postgrad, A.Crystallographer and R. Supervisor

Abstract DOI: 10.1021/jacs.5b03972

Coordination-directed self-assembly has become a well-established technique for the construction of functional supramolecular structures. In contrast to the most often exploited transition metals, trivalent lanthanides Ln^{III} have been less utilized in the design of polynuclear self-assembled structures despite the wealth of stimulating ….[…]

Supporting Information

Crystallographic Information File PDF

» ja5b03972_si_001.cif (2.93 MB) » ja5b03972_si_002.pdf (2.48 MB)

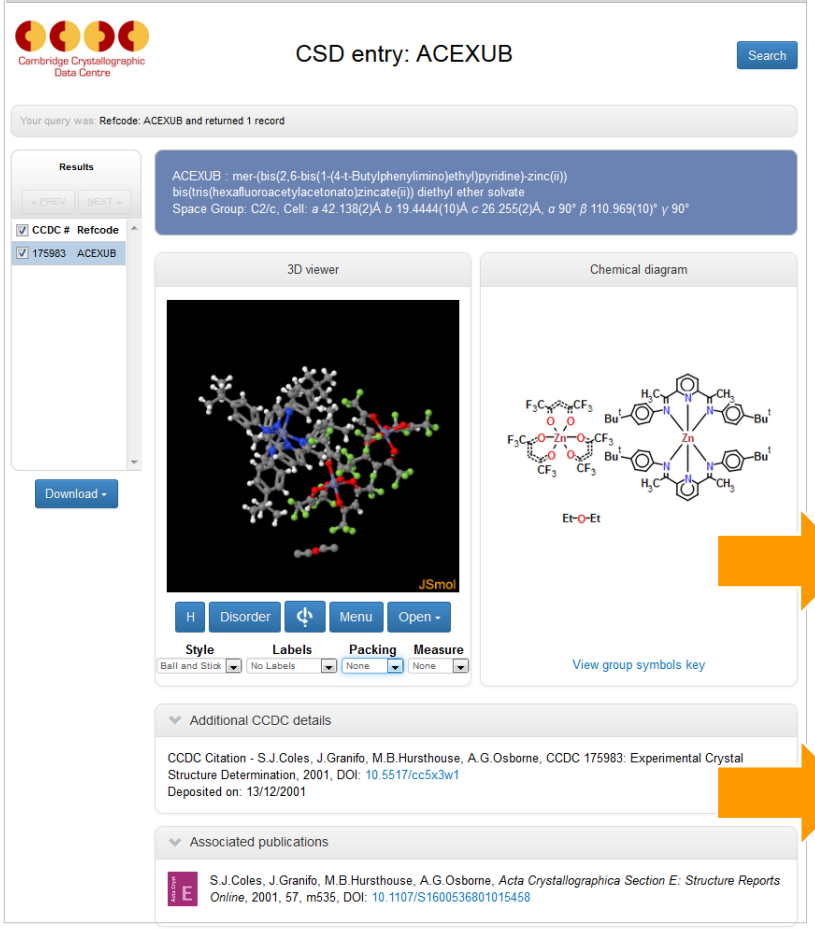
Crystallographer doesn't always get credited for their contribution.

Article DOI primarily identifies the paper and not the data.

Crystal structure data often buried amongst the supporting information.

Provenance and attribution

CSD DOI – Access to CSD entry & data



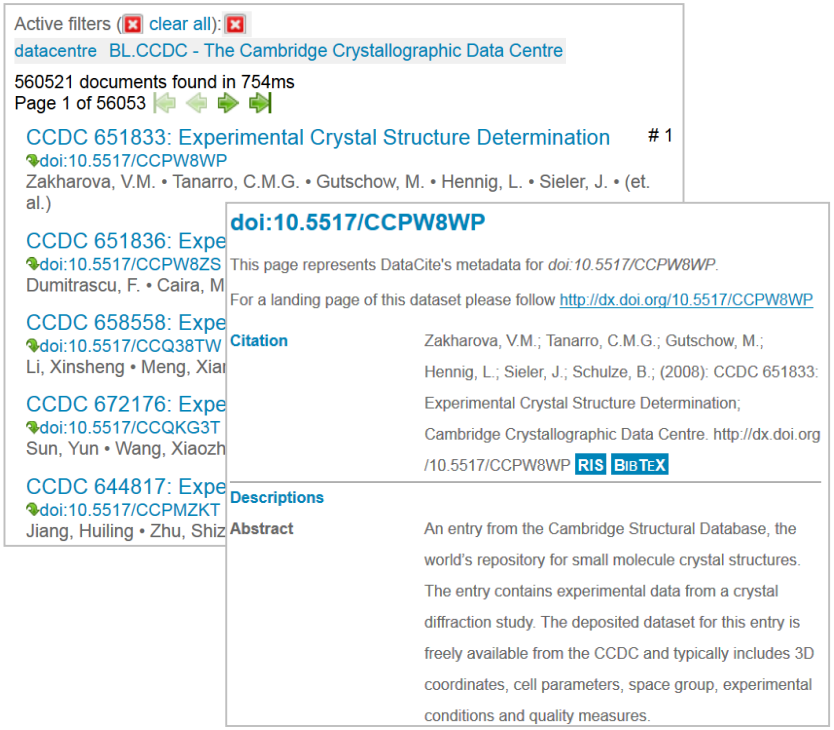
Information added to structure of record

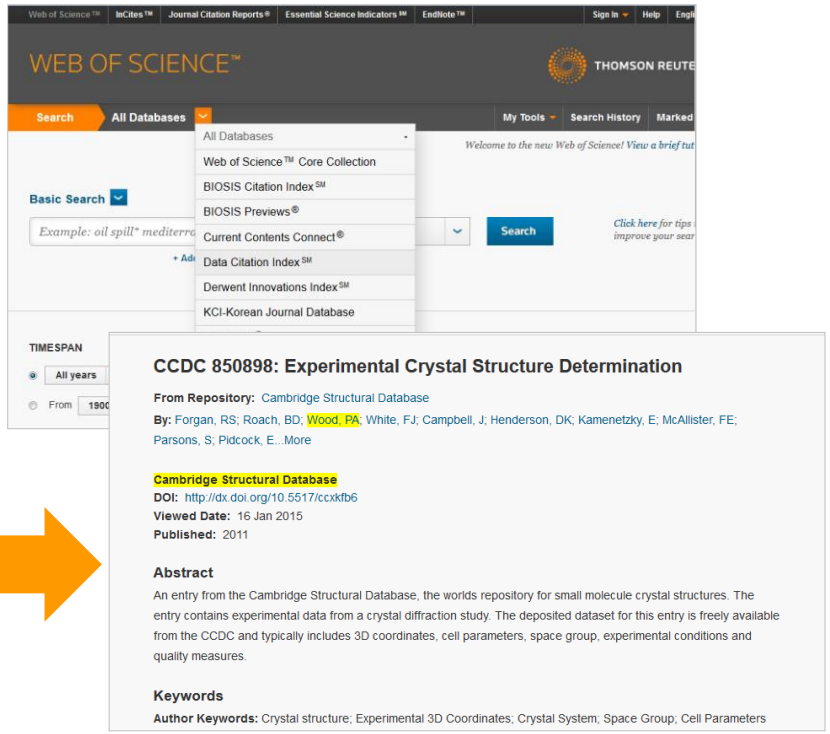
```
##############################################################################
#
# This file contains crystal structure data downloaded from the
# Cambridge Structural Database (CSD) hosted by the Cambridge
# Crystallographic Data Centre (CCDC).
#
# Full information about CCDC data access policies and citation
# guidelines are available at http://www.ccdc.cam.ac.uk/access/V1
#
# Audit and citation data items may have been added by the CCDC.
# Please retain this information to preserve the provenance of
# this file and to allow appropriate attribution of the data.
#
##############################################################################

data_I
_audit_block_doi                 10.5517/cc5x3w1
_database_code_depnum_ccdc_archive 'CCDC 175983'
loop_
_citation_id
_citation_doi
_citation_year
1 10.1107/S1600536801015458 2001
_audit_update_record
;
2001-12-13 deposited with the CCDC.
2015-08-18 downloaded from the CCDC.
```

Dataset publication

CCDC 175983: Experimental Crystal Structure Determination. **A. Crystallographer**, *Cambridge Crystallographic Data Centre* (2001) http://dx.doi.org/10.5517/cc5x3w1

DOIs – facilitating interoperability





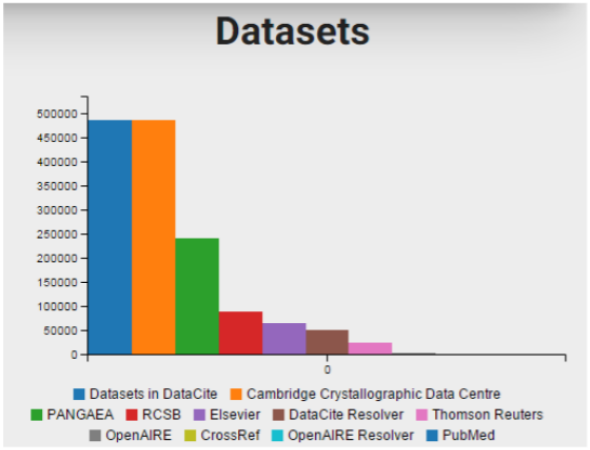
DataCite

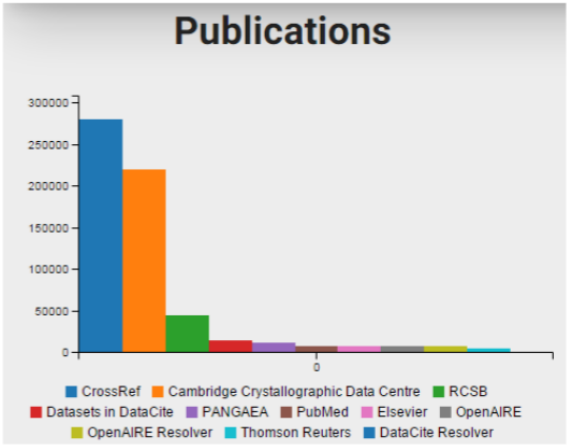
Facilitates formal citation AND interoperability. Currently adding indication of full citation and resourceType to metadata

COVERED BY DATA CITATION INDEX THOMSON REUTERS

Coverage of CCDC data by the Thomson Reuters Data Citation Index achieved via the DataCite metadata store.

DOIs – facilitating interoperability



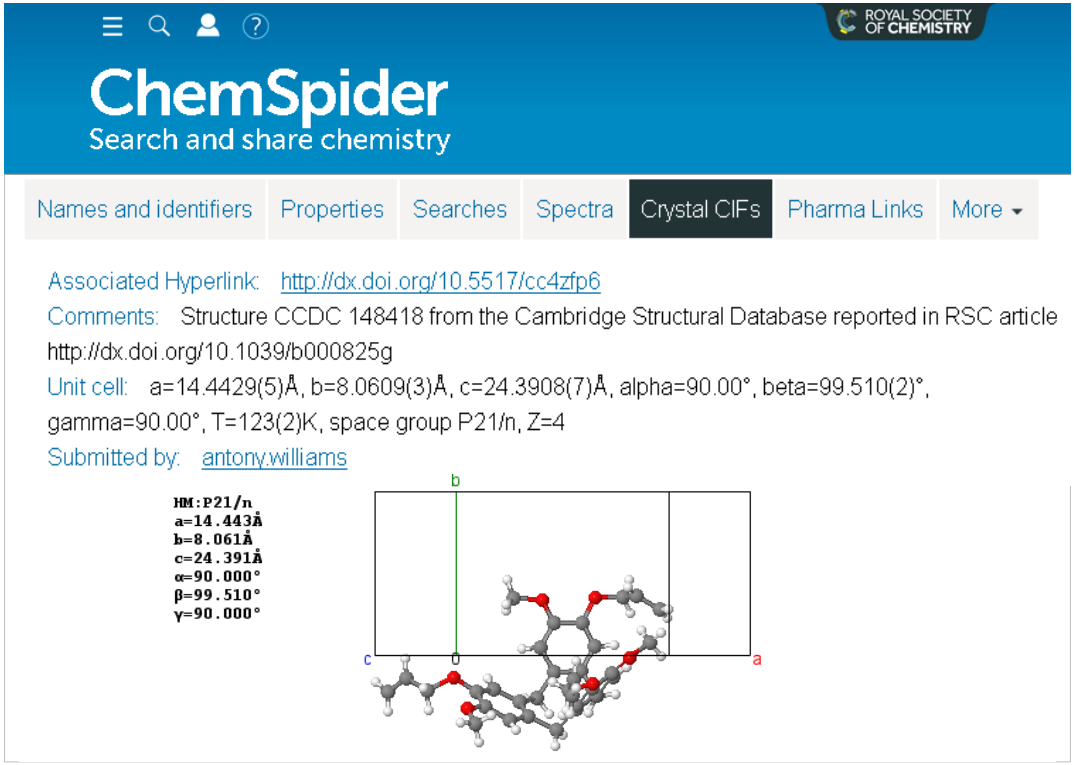


UK Research Data Discovery Service: Aim is to aggregate metadata for research data held within UK universities and national, discipline specific data centres.

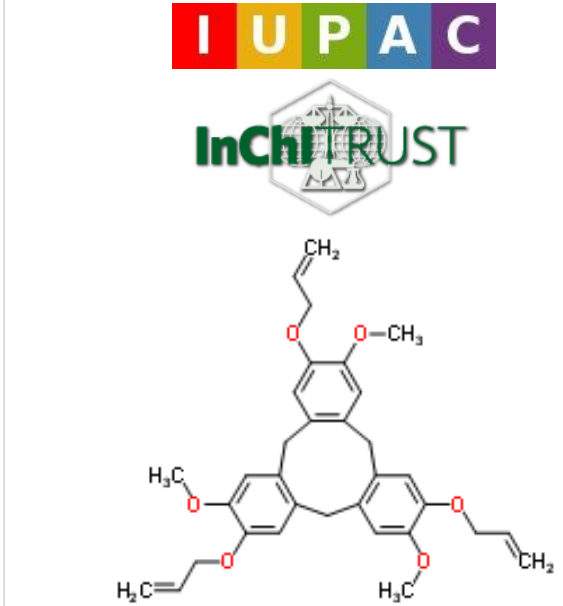
RDA-WDS Publishing Data Services prototype taking a feed of CCDC article-dataset links from DataCite.

Data literature inter-inking service (beta) that ingests data article -inks. Inspection portal : http://dliservice.research-infrastructures.eu/

Linking based on chemistry



Bulk identification of links between ChemSpider molecules and CSD Crystal Structures (~50,000) facilitated by InChIs.



Standard InChI:
`InChI=1S/C33H36O6/c1-7-10-37-31-19-25-13-23-17-29(35-5)33(39-12-9-3)21-27(23)15-24-18-30(36-6)32(38-11-8-2)20-26(24)14-22(25)16-28(31)34-4/h7-9,16-21H,1-3,10-15H2,4-6H3`

Standard InChIKey:
`IZHKSTHBLQRIOW-UHFFFAOYSA-N`

Other crystal data repositories

- PDB
  - match PDB ligands to best representative CSD molecules
  - working group on ligand validation

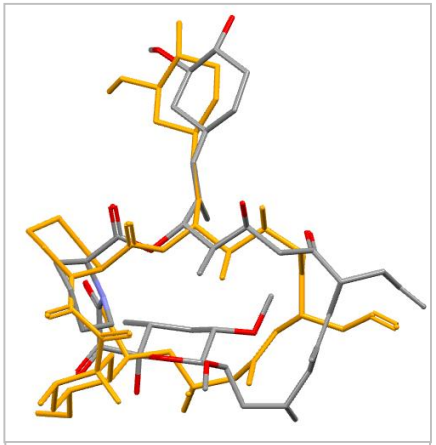
29 July, 2015

Data correspondences between the PDB and CSD archives now available

The Worldwide Protein Data Bank and the Cambridge Crystallographic Data Centre (CCDC; **http://www.ccdc.cam.ac.uk**) are pleased to announce the availability of a new data resource containing correspondences between the biopolymer components and ligand molecules found in the PDB archive that exactly match small-molecule X-ray structures in the Cambridge Structural Database (CSD) archive.

The chemical structure of every unique molecule in the Protein Data Bank is described in the **PDB Chemical Component Dictionary**. The new PDB Chemical Component Model data file complements information in the PDB by providing the following CSD information for matching molecular entries: accession code correspondences, Cartesian coordinates and R-value, data-collection temperature and a disorder flag, SMILES and InChI descriptors, and a Digital Object Identifier (DOI) for the citation associated with the CSD entry.

At present, there are 20,077 chemical components in the PDB Chemical Component Dictionary, and for 1,418 of these exact match structures have been identified in the CSD. The new PDB Chemical Component Model file is available from the PDB FTP archive via:



CSD FINWEE10/PDB FK5

Linking experiment to data

ISIS: Neutron and muon instruments at the UK Science and Technology Facilities Council Rutherford Appleton Laboratory.

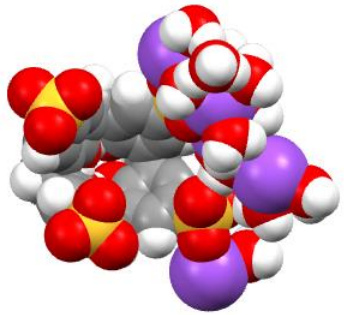
ICAT

Experiment Data

The ISIS data catalogue

All data produced at ISIS is catalogued into ICAT – the ISIS data catalogue. This catalogue is made available as a searchable website, via Mantid and as an API.

ISIS will issue citable DOIs for all experiments carried out at ISIS. You are encouraged to cite these in publications relating to ISIS experiments. For more information see DOIs



CCDC doi:10.5517/ccxns8l

```
_diffrn_radiation_type          neutron
_diffrn_radiation_source        'ISIS spallation source'
```

CCDC and STFC investigating the potential for linking between:

- raw data stored at STFC
- modelled structure stored at CCDC

Challenge: reliably associating results with original experiment

Metadata about people and institutions

ORCID

Assigning DOIs through DataCite allows researchers to add their data to researcher IDs such as ORCiD

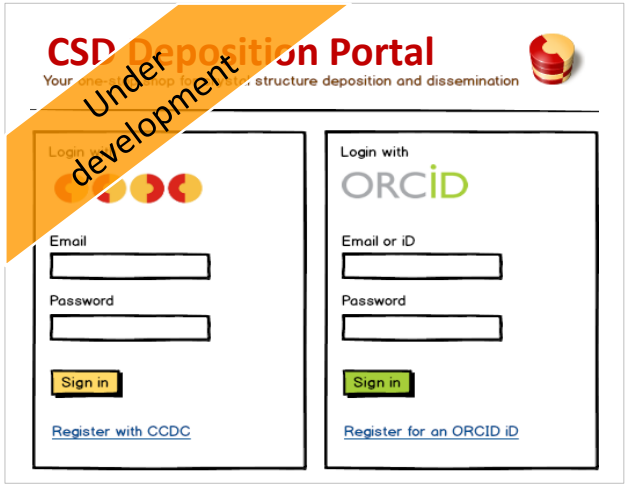
fundref

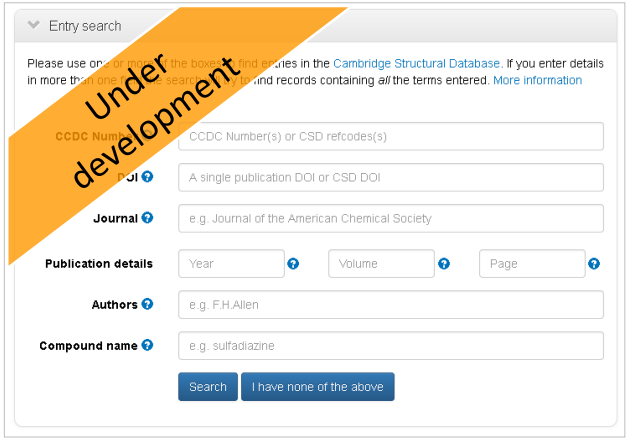
Funders are wanting to know the output of their funded research. This will require linking between people, publications, data and funders

isni

Could be used to find institutions research output. Ringgold IDs now in our internal systems and we want to work towards reliably identifying institutions at point of deposition

What next?





CSD Deposition Portal

- Integration of Researcher and Institutional IDs
- Exposure of CCDC services during deposition
  - Ability to edit CSD entry metadata such as 2D chemical diagram
  - Extend validation and integrity checks
- Capturing other metadata
  - Topological descriptions, funders, experimental data, links to associated data, etc
- Integration with third party software

Interoperability, links and programmatic access

- Extension of links….
- Extend CSD Python API to access deposited data

CSD-Community Access

- Extension of search functionality
- Addition of InChI lookups

Thank you