INTERNATIONAL DATA WEEK 2016

The Cambridge Structural Database

Ian Bruno, Suzanna Ward

The Cambridge Crystallographic Data Centre

@ijbruno @ccdc_cambridge

SciDataCon 2016: Crystallography and Structural Databases, 12 September 2016

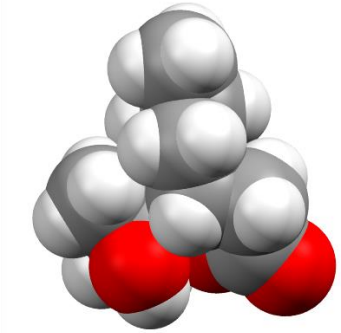
The Cambridge Structural Database

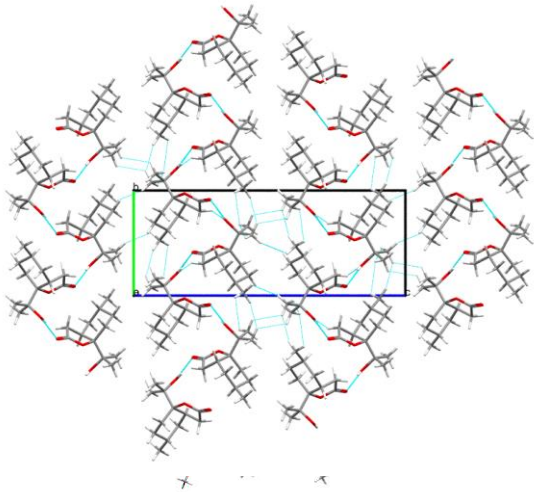
A Database of Crystal Structures

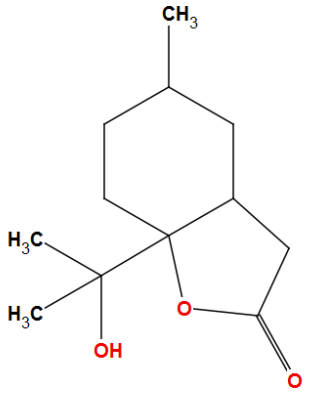
KOCXUX: $C_{24}H_{27}NS_2Sn$: (5-Isopropyl-1,3,5-dithiazinan-2-yl)-triphenyl-tin
R.Colorado-Peralta et al., J.Organomet.Chem. (2014)

GEQMIA: $C_{26}H_{23}NO_4$: 1,2,6-tris(4-methoxyphenyl)pyridin-4(1H)-one
K.M.Oberg, CSD Communication (2012) Contribution from Colorado State University

JASJAP: (1R,4R,6R)-(-)-1-(1'-Hydroxy-1'-methylethyl)-4-methyl-9-oxabicyclo(4.3.0)nonan-8-one
I.Dams, A.Bialonska, Z.Ciunik, C.Wawrzenczyk, Tetrahedron:Asymm. (2005), 16, 2087



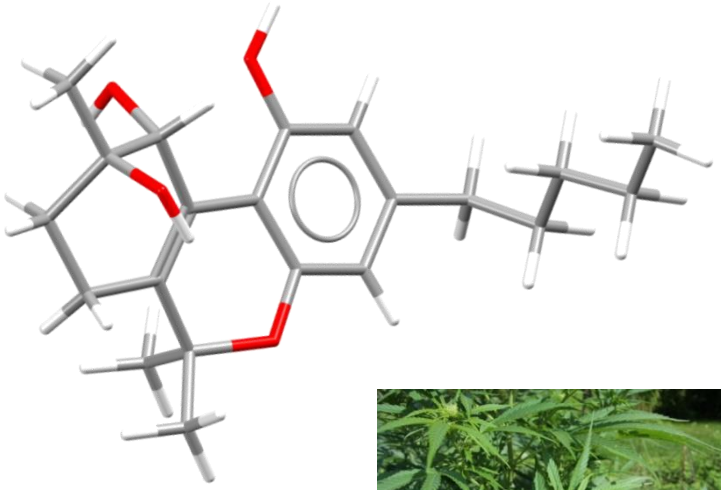




Moderate antifeedant activity against Sitophilis granarius L., Trogoderma granarium Ev., Tribolium confusum Duv., the lesser mealworm and the Colorado potato beetle

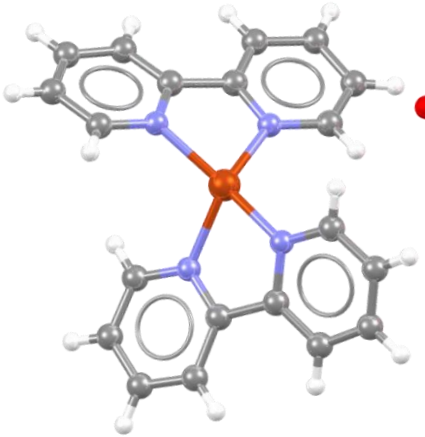
The Cambridge Structural Database

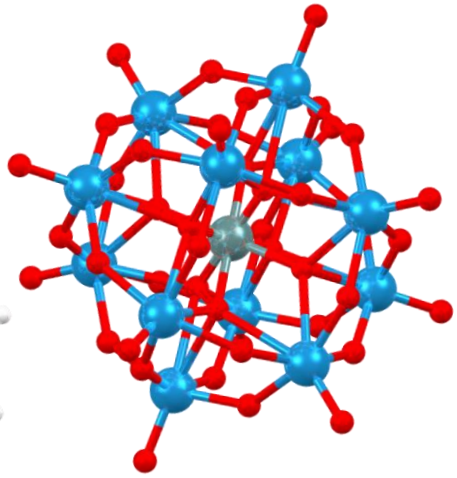
A Database of Crystal Structures



CIXFOE: (+-)-trans-Cannabitriol
an extract from Cannabis sativa

PEPSIO





tetrakis(bis(2,2'-Bipyridine)-copper(i)) (μ_{12}-germanato)-tetracosakis(μ_2-oxo)-dodeca-oxo-dodeca-tungsten

The Cambridge Structural Database

A Repository of Research Datasets

CCDC 252295: doi:10.5517/cc8gjks

```
_cell_length_b                    8.3463(7)
```

CCDC 917625: doi:10.5517/cczsvtd

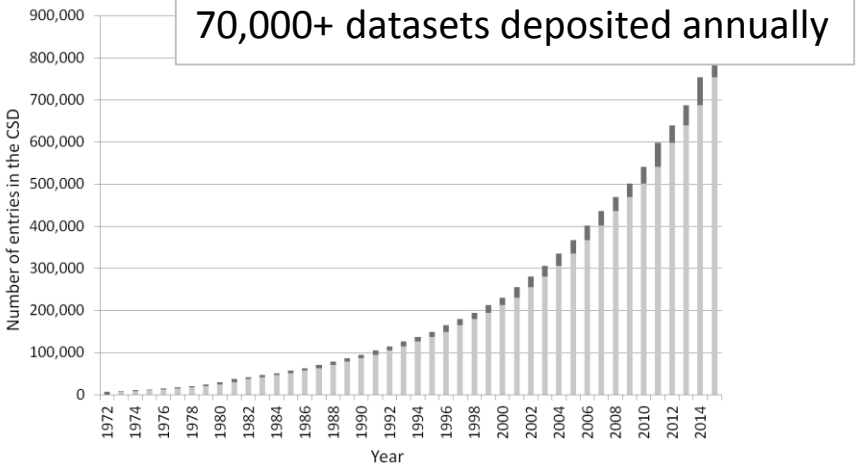
```
_refine_ls_wR_factor_ref          0.1665
_refine_ls_wR_factor_gt           0.1240
```

CCDC 933841: doi:10.5517/cc10bqxz

```
_diffrn_source                    'Enraf Nonius FR590'
_diffrn_ambient_temperature       293(2)
_diffrn_radiation_wavelength      0.71073
_diffrn_radiation_type            MoK\a
_diffrn_radiation_monochromator   graphite
_diffrn_radiation_probe           x-ray
_diffrn_detector                  'CCD plate'

_refine_ls_structure_factor_coef  Fsqd
_refine_ls_matrix_type            full
_refine_ls_weighting_scheme       calc
_refine_ls_hydrogen_treatment     mixed

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
C2 C 0.25054(14) 0.43179(18) 0.0354(4) 0.0479(7) Uani 1 1 d . . .
H2 H 0.2161 0.3958 -0.0319 0.057 Uiso 1 1 calc R . .
C4 C 0.31582(17) 0.4034(2) -0.2628(5) 0.0667(10) Uani 1 1 d . . .
H4A H 0.2787 0.3685 -0.313 0.08 Uiso 1 1 calc R . .
H4B H 0.3337 0.4248 -0.3705 0.08 Uiso 1 1 calc R . .
C6 C 0.3379(2) 0.3078(2) -0.0106(6) 0.0682(10) Uani 1 1 d . . .
C7 C 0.42890(17) 0.3945(2) -0.0854(6) 0.0714(11) Uani 1 1 d . . .
H7 H 0.4213 0.4391 0.0007 0.086 Uiso 1 1 calc R . .
C8 C 0.4788(2) 0.3353(4) 0.0224(9) 0.143(2) Uani 1 1 d . . .
H8A H 0.4655 0.321 0.141 0.215 Uiso 1 1 calc R . .
H8B H 0.5216 0.3607 0.0465 0.215 Uiso 1 1 calc R . .
H8C H 0.4808 0.2867 -0.0523 0.215 Uiso 1 1 calc R . .
C9 C 0.4558(2) 0.4305(4) -0.2510(9) 0.129(2) Uani 1 1 d . . .
H9A H 0.4591 0.3883 -0.3428 0.193 Uiso 1 1 calc R . .
H9B H 0.4988 0.4532 -0.207 0.193 Uiso 1 1 calc R . .
H9C H 0.4267 0.4728 -0.3088 0.193 Uiso 1 1 calc R . .
C10 C 0.27717(13) 0.59110(18) 0.3750(4) 0.0411(7) Uani 1 1 d . . .
```

850,000+ crystal structure datasets

70,000+ datasets deposited annually

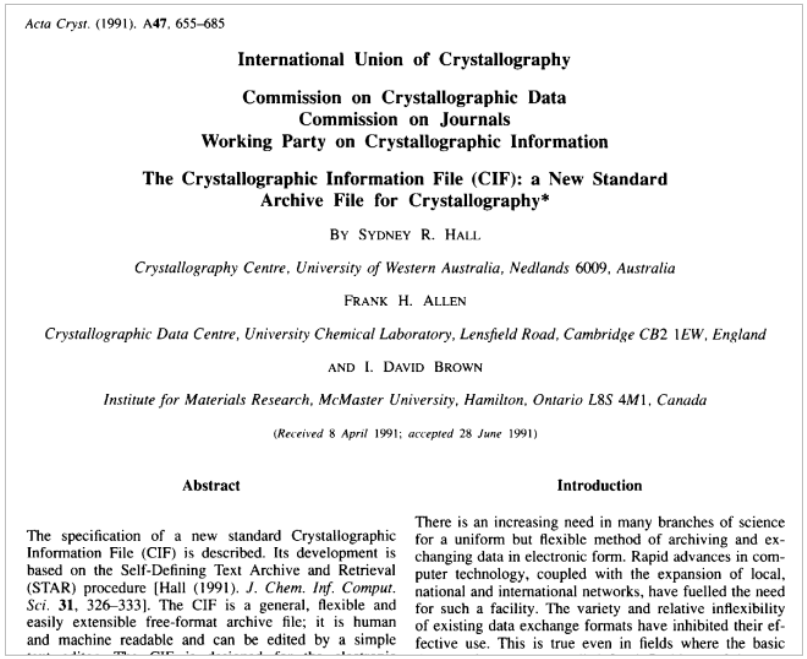


General areas of interest and opportunity:

- ❑ Identifiers
- ❑ Validation
- ❑ Linking

The Crystallographic Information File (CIF)

Community Standards: CIF format



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, England

AND I. DAVID BROWN

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

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

Abstract

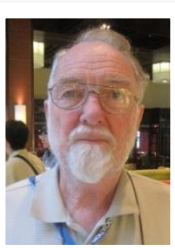
The specification of a new standard Crystallographic Information File (CIF) is described. Its development is based on the Self-Defining Text Archive and Retrieval (STAR) procedure [Hall (1991). *J. Chem. Inf. Comput. Sci.* **31**, 326–333]. The CIF is a general, flexible and easily extensible free-format archive file; it is human and machine readable and can be edited by a simple

Introduction

There is an increasing need in many branches of science for a uniform but flexible method of archiving and exchanging data in electronic form. Rapid advances in computer technology, coupled with the expansion of local, national and international networks, have fuelled the need for such a facility. The variety and relative inflexibility of existing data exchange formats have inhibited their effective use. This is true even in fields where the basic

Syd Hall

Frank Allen



David Brown

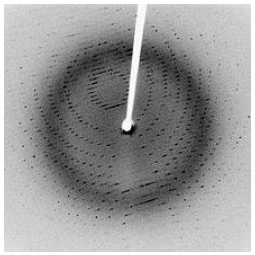
- A standard format for archive and exchange of crystallographic data
  - derived results
  - raw and processed data
  - experimental conditions etc.

```
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
Cl1 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
Cl2 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
```

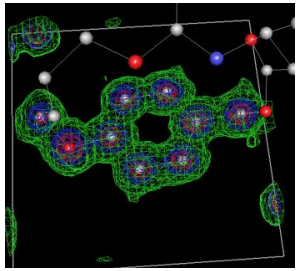

Acta Cryst., **1991**, A47, 655 doi:10.1107/S010876739101067X

From Raw Data to Model

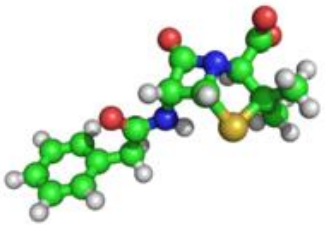
Raw	
Image data	Positions and intensities of diffracted beams



Processed	
Structure factor data	Quantities needed to derive electron density



Derived	
Atom coordinates	Modelled and refined from electron density

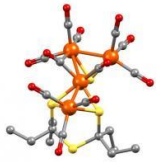


DOI:10.5517/ccdc.csd.cc1kx2j2

DOIs: Data Citation and Interoperability

DataCite

- Over 500,000 dataset DOIs registered since April 2014
- DataCite metadata includes related article DOIs
- **Foundation for formalising data citation and interoperability**



10.5517/CCPHZ37

DC1 Data Citation Principles

Data should be considered legitimate, citable products of research…

https://www.force11.org/datacitation

Dataset Publication

CCDC 892348: Experimental Crystal Structure Determination. **A. Crystallographer**, *Cambridge Crystallographic Data Centre* (2013) http://dx.doi.org/10.5517/CCYYKFV

RDA RESEARCH DATA ALLIANCE

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

COVERED BY DATA CITATION INDEX THOMSON REUTERS

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

Jisc

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

Raw Data: Small Molecules

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

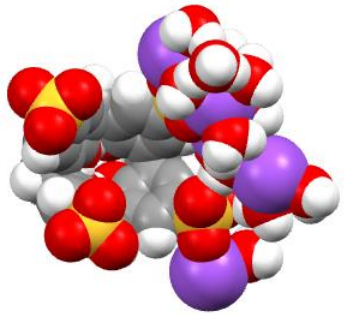
ICAT

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



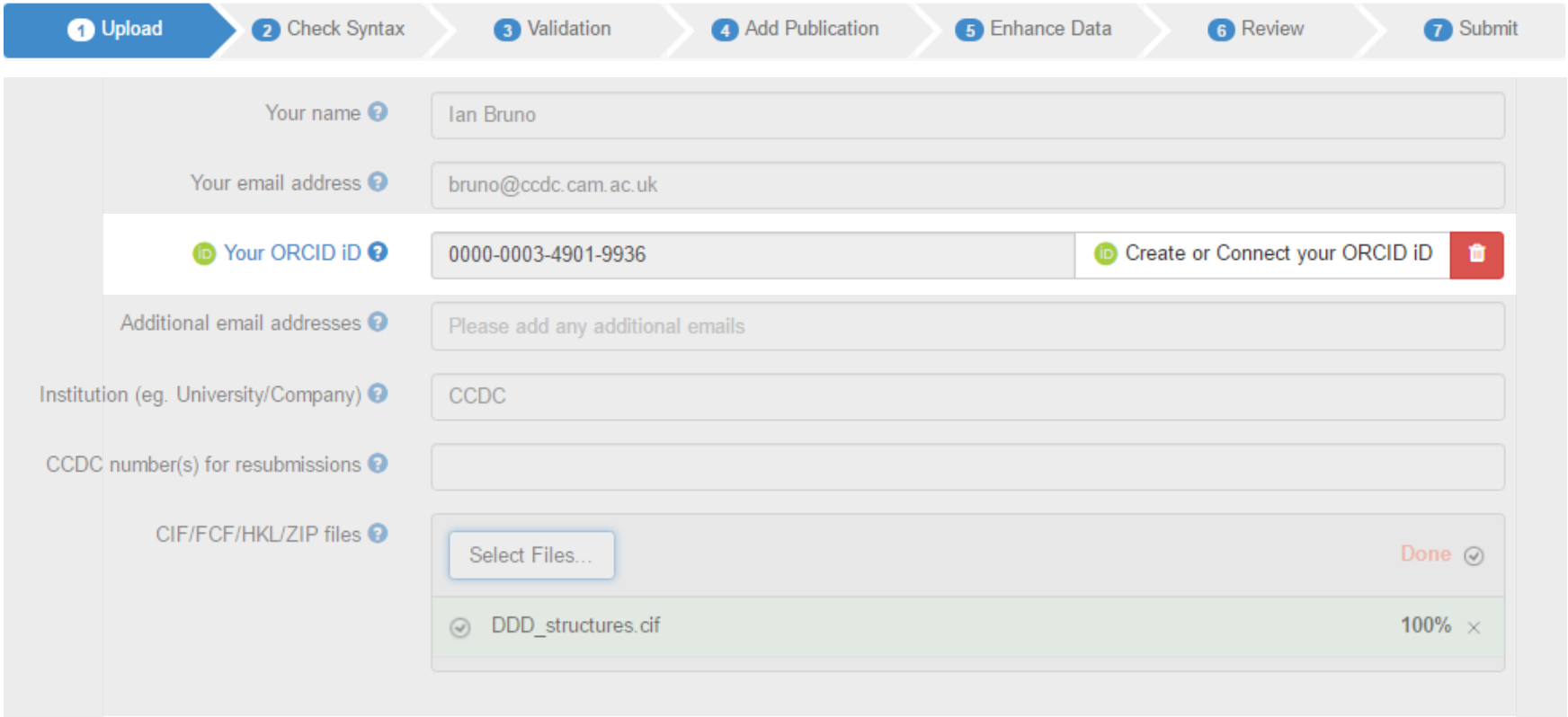
CCDC doi:10.5517/ccxns8l

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

Both the raw data stored at STFC and the modelled structure stored at CCDC have DOIs. **Can we link these?**

Challenge: reliably associating results with original experiment

CCDC Data Deposition Service



ORCID
Connecting Research and Researchers

CCDC deposition service now features an initial implementation of an ORCID integration that will enable researchers to be unambiguously connected with their research output

Publication Standards

International Union of CRYSTALLOGRAPHY

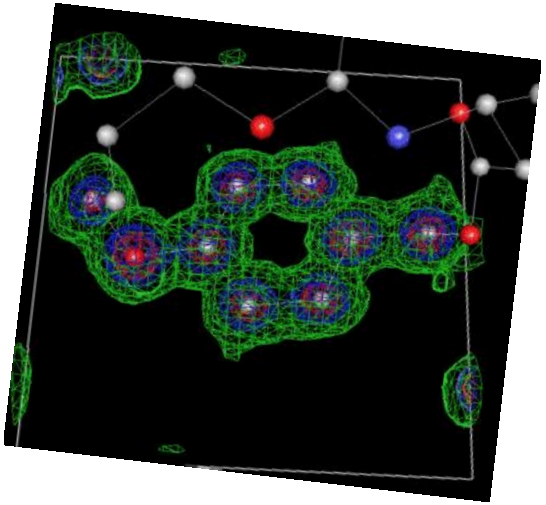
Home > home page > leading article > 2011 > 2011-06-02 http://www.iucr.org/home/leading-article/2011/2011-06-02

Publication standards for crystal structures

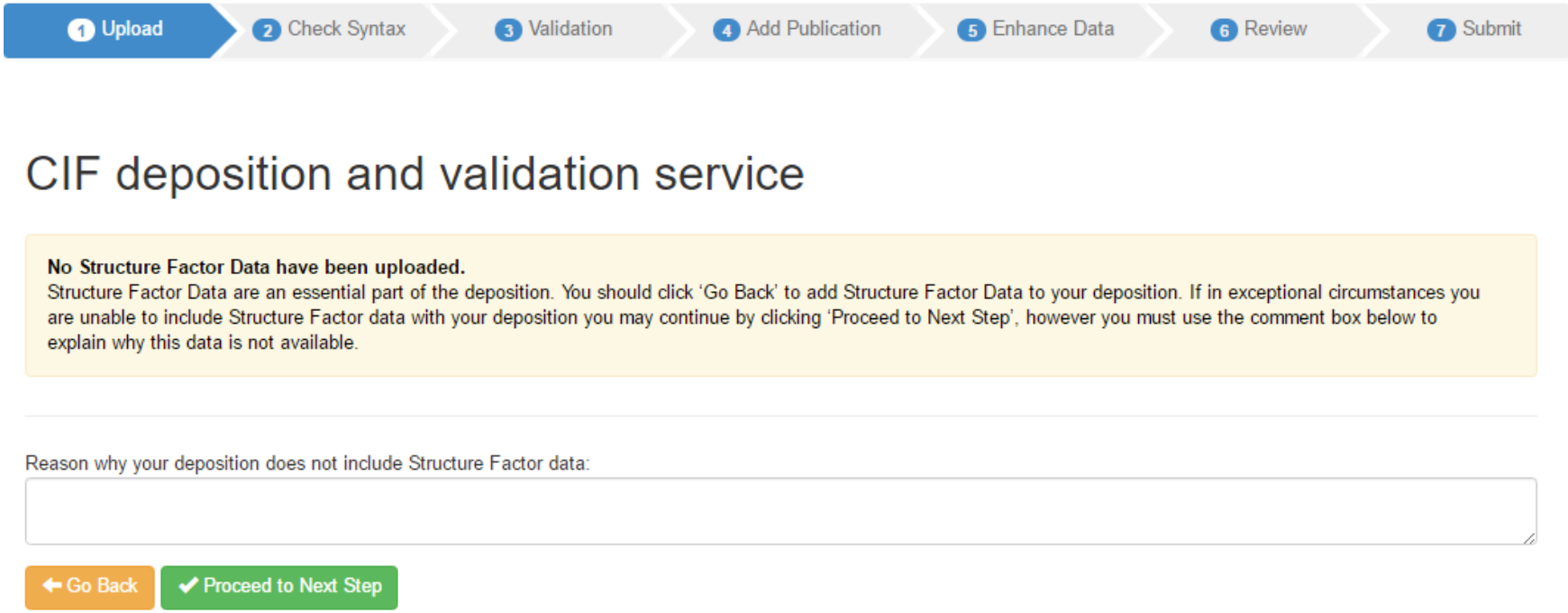
- Encourage authors to provide machine-readable structure factors
- Authors, referees and editors should take advantage of checkCIF

Structure factors: Processed data that enable more rigorous assessment of the derived atomic coordinates

checkCIF: a service of the IUCr that assesses completeness, consistency and integrity of the data

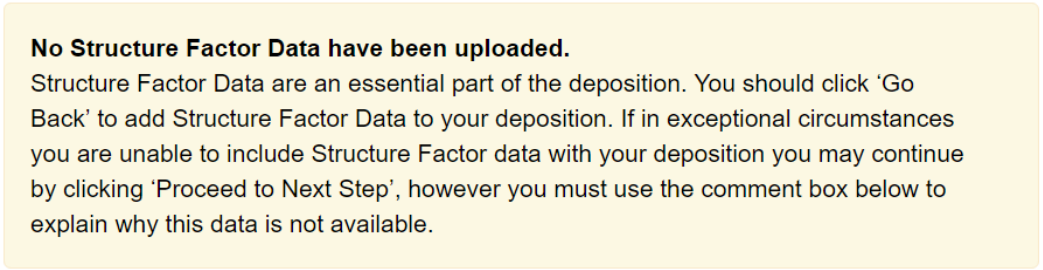


Deposition of Processed Data

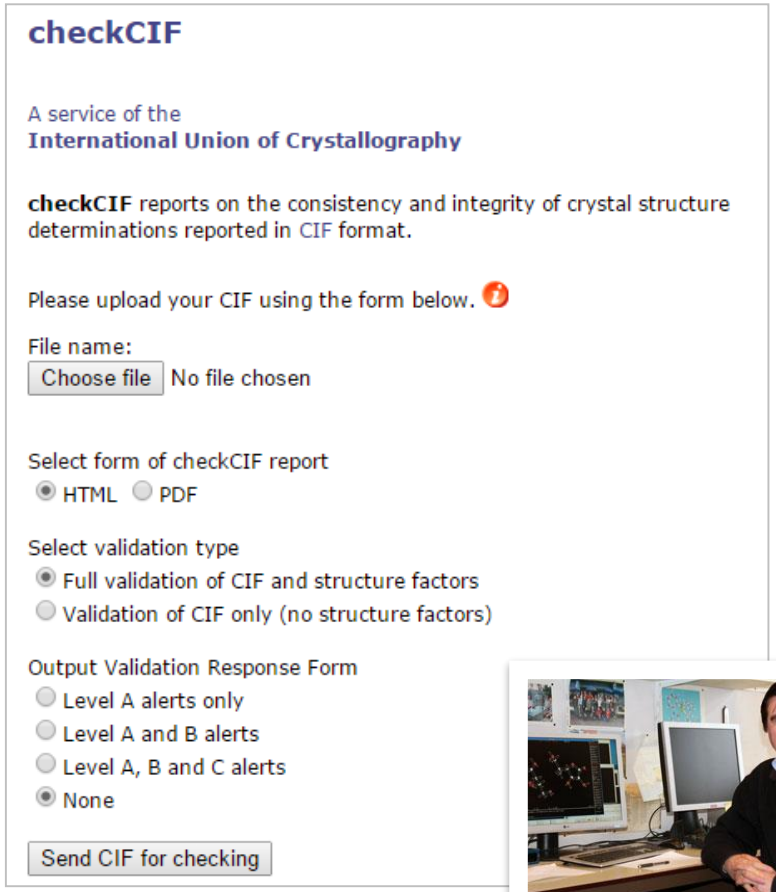


Deposition of Processed Data

- **7% of 800 recently deposited structures did not have structure factors**
- **Reasons given for why deposition structure factor data is not included:**
  - *The structures were measured in 2003 and shelx97 was used*
  - *The new software is not work[ing] well on my computer now*
  - *In present stage, I can not provide structure factor data*
  - *The CIF file was generated without SF*
  - *Not available at time of deposition*
  - *Submission to a journal*
  - *Don't have them*
  - *Convenience*
  - *NO REASON*
  - *No need*
  - *Yes*
  - *No*



Publication Standards: checkCIF



- Checks consistency and integrity of the data
- Generates alerts that should either be corrected or explained
- Structure factors enable more rigorous checks
- Many journals require or request checkCIF reports

Much of checkCIF based on components of PLATON developed by Ton Spek, Utrecht University

checkCIF is sponsored by

Crystallography Journals Online

ELSEVIER

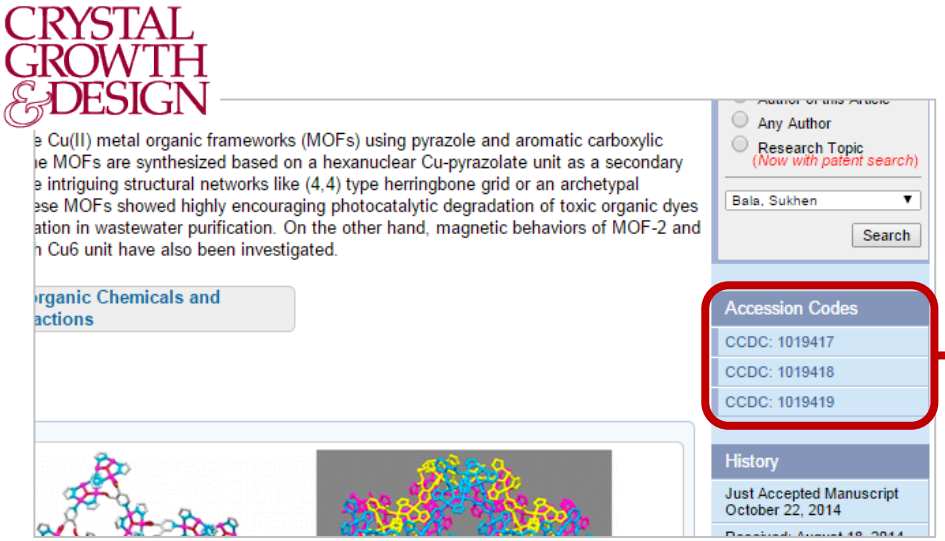
WILEY

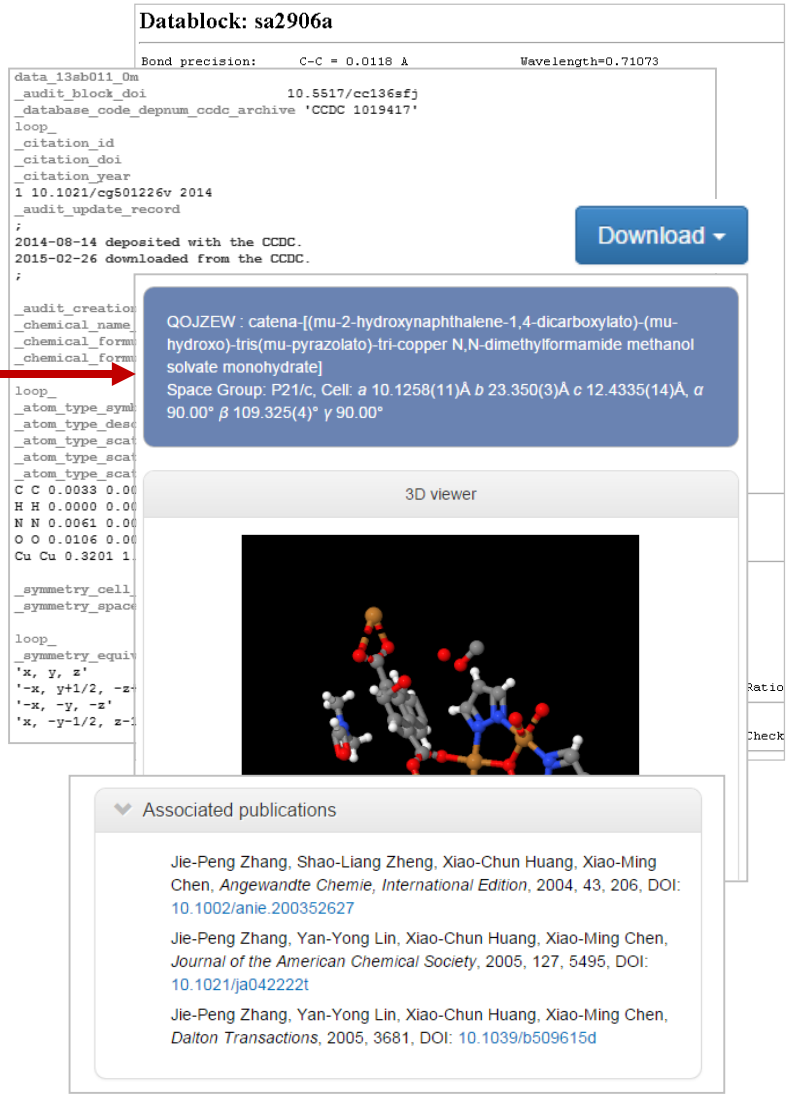
ROYAL SOCIETY OF CHEMISTRY

Reducing Technical Barriers

Validation reports available to reviewers pre-publication and researchers post-publication alongside the data to which they pertain

Linking Articles and Data



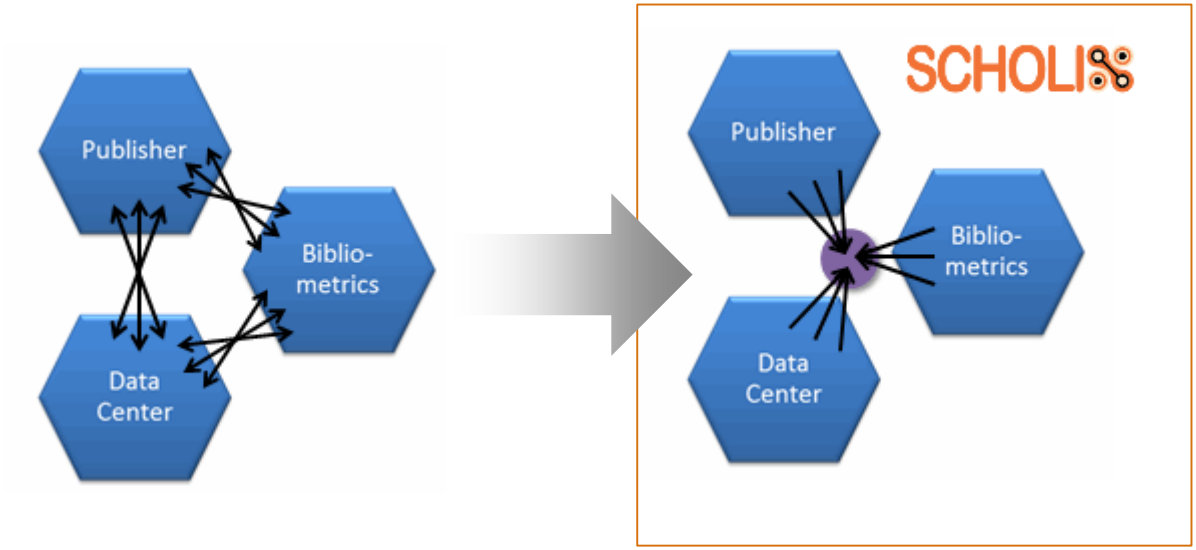


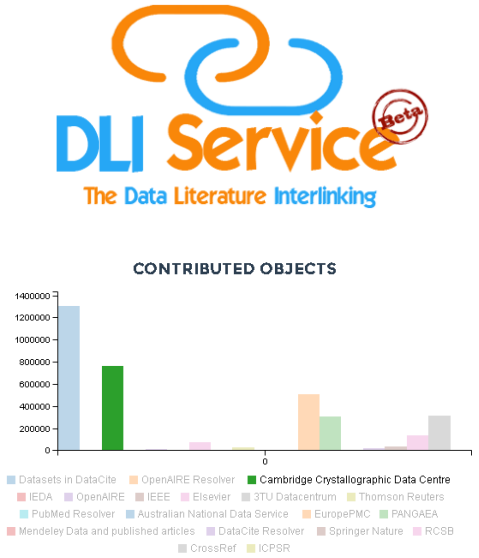
Links between articles and data stored at CCDC are in place for ACS, RSC, Elsevier, Wiley and IUCr journals

Current linking mechanism rely on bespoke services and feeds developed and maintained by CCDC and journal publishers

Public Structure Summary Page

- **SCHOLIX: A high-level interoperability framework for exchanging information about the links between scholarly literature and data**
- A product of the ICSU-WDS / RDA Data Publication Services Working Group
- Prototype / demonstrator service developed by OpenAire with support from PANGEA



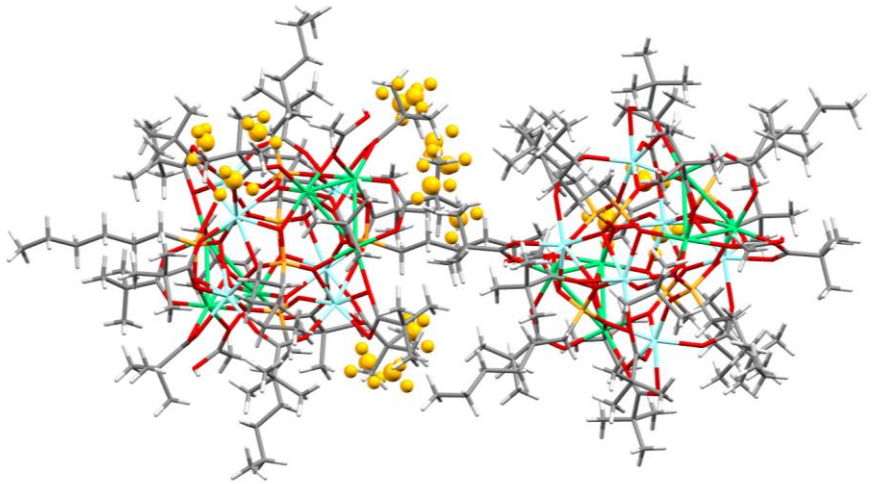


See http://www.scholix.org/about & http://bit.ly/29tdGNU & http://dliservice.research-infrastructures.eu

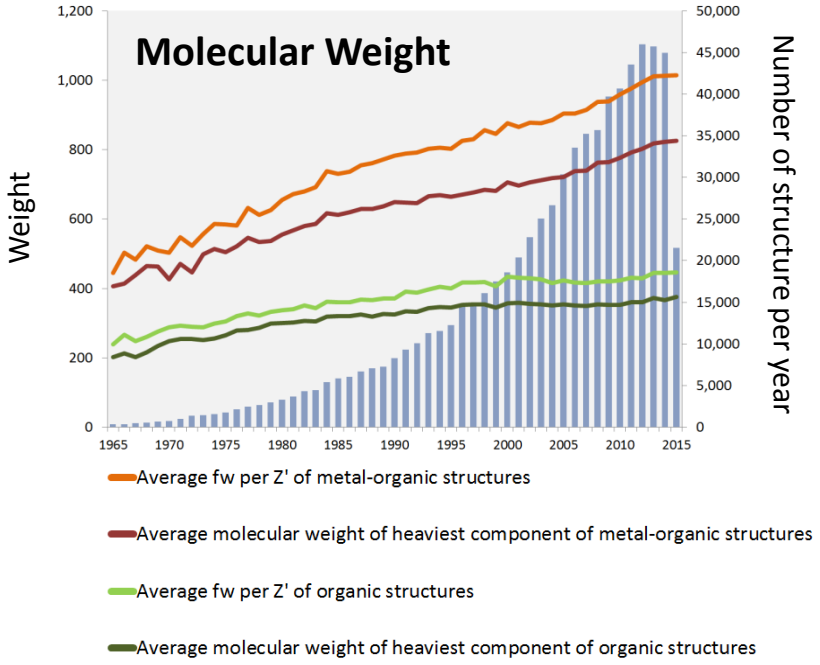
CSD Challenges

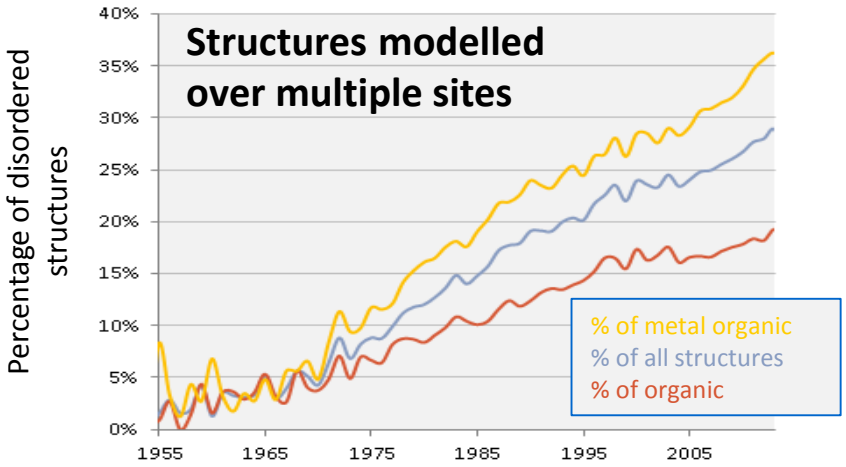
Increasing throughput

- 408,899 different publications
- 1,864 literature sources
- 451 different publishers
- 18 different languages
- 11,113 different depositors in 2015
- 85,913 datasets deposited in 2015

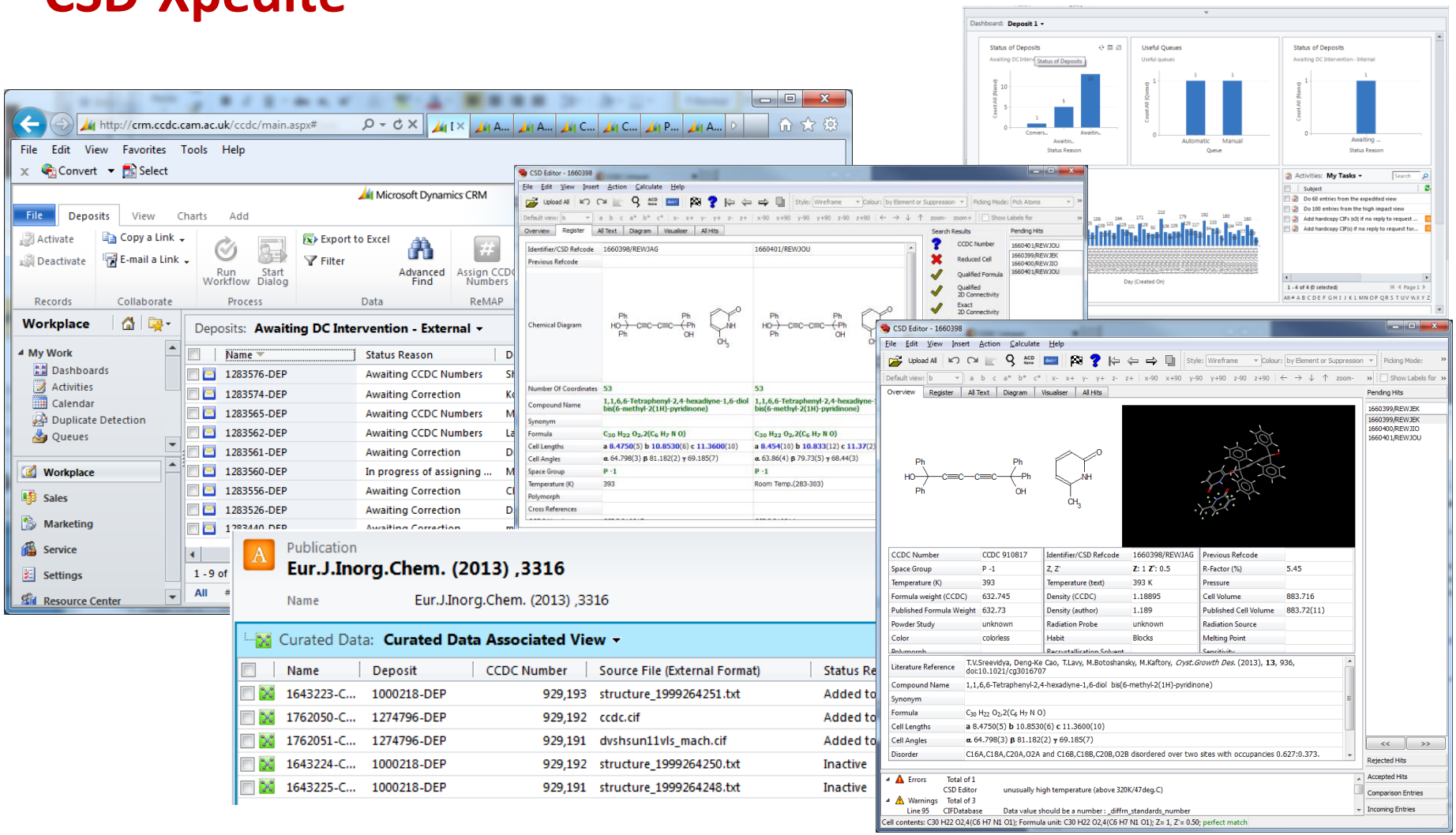


Increasing scientific complexity



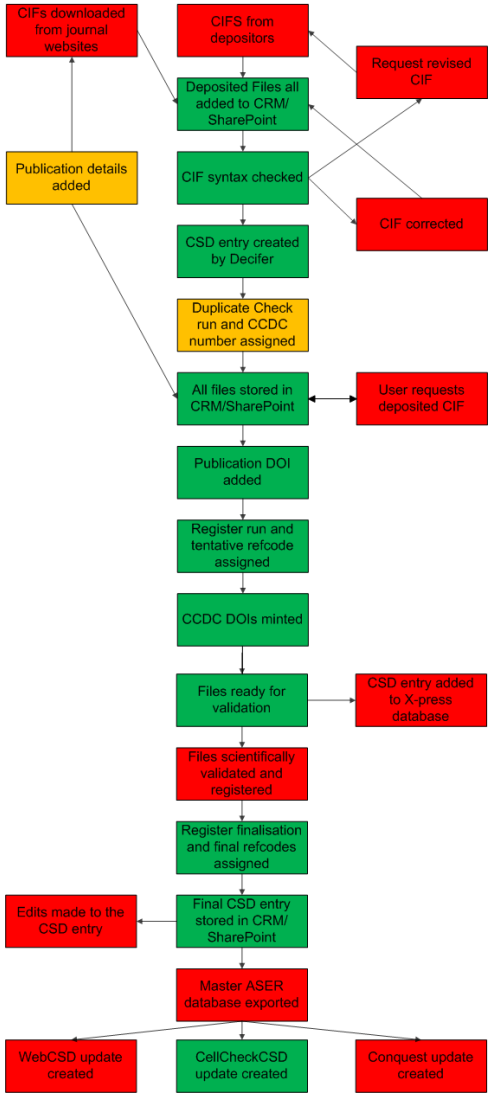


CSD-Xpedite


http://crm.ccdc.cam.ac.uk/ccdc/main.aspx#
Microsoft Dynamics CRM
Workplace
Deposits: Awaiting DC Intervention - External
Publication
Eur.J.Inorg.Chem. (2013) ,3316
Curated Data: Curated Data Associated View
Name
Deposit
CCDC Number
Source File (External Format)
1643223-C... 1000218-DEP 929,193 structure_1999264251.txt
1762050-C... 1274796-DEP 929,192 ccdc.cif
1762051-C... 1274796-DEP 929,191 dvshsun11vls_mach.cif
1643224-C... 1000218-DEP 929,192 structure_1999264250.txt
1643225-C... 1000218-DEP 929,191 structure_1999264248.txt

CSD-Xpedite

- Internal informatics system designed to cope with increasing demands
- Foundation for improved deposition services and interfacing with external services
- Enables timely and appropriate release of structures
- Automate as much as can be reliably automated

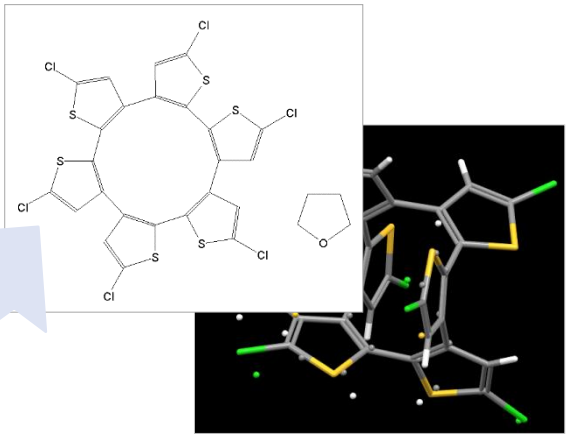


Assignment of Chemistry

```
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
Cl1 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
Cl2 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
```

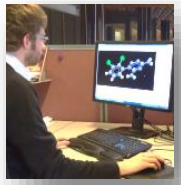
- Initial automatic assignment based on knowledge in the CSD
- Reliability score from automatic assignment enables triage
- All structures ultimately reviewed by expert editorial staff

Assignment of chemistry is required to make data findable, interoperable and reusable

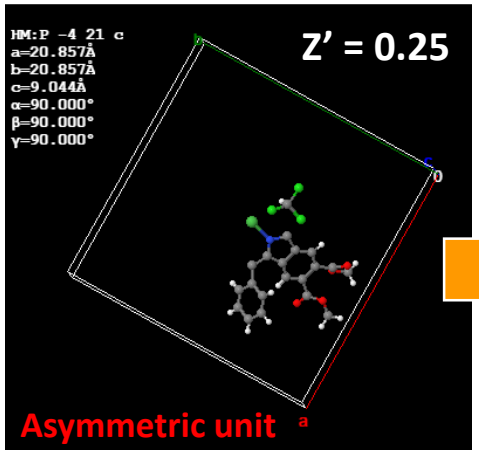


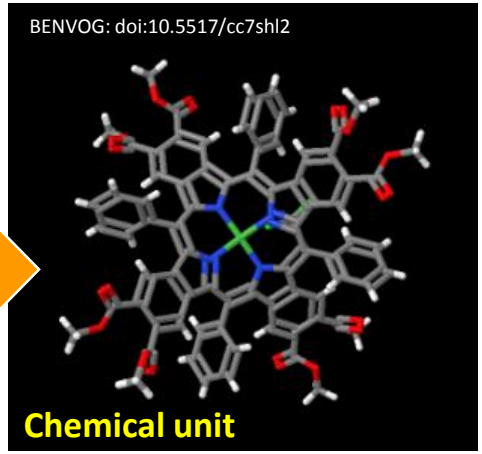
ABOWOD: doi:10.5517/cc5jk3b

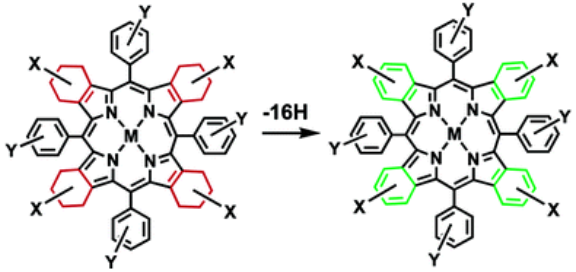
$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}.$$



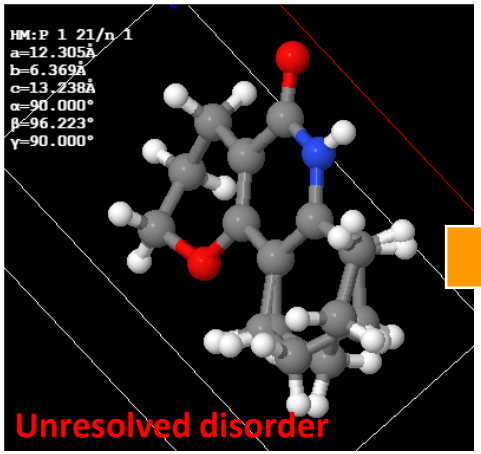
Chemical interpretation of crystallographic data

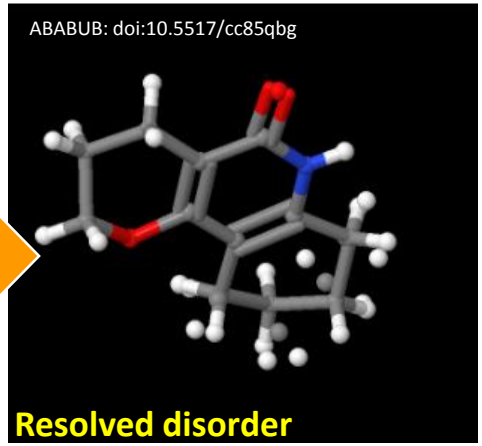


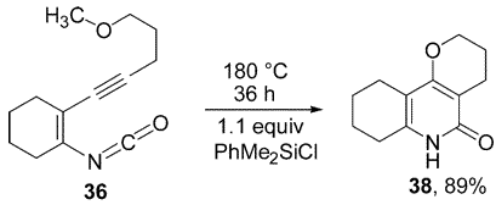




Novel Versatile Synthesis of Substituted Tetrabenzoporphyrins
doi:10.1021/jo0350054



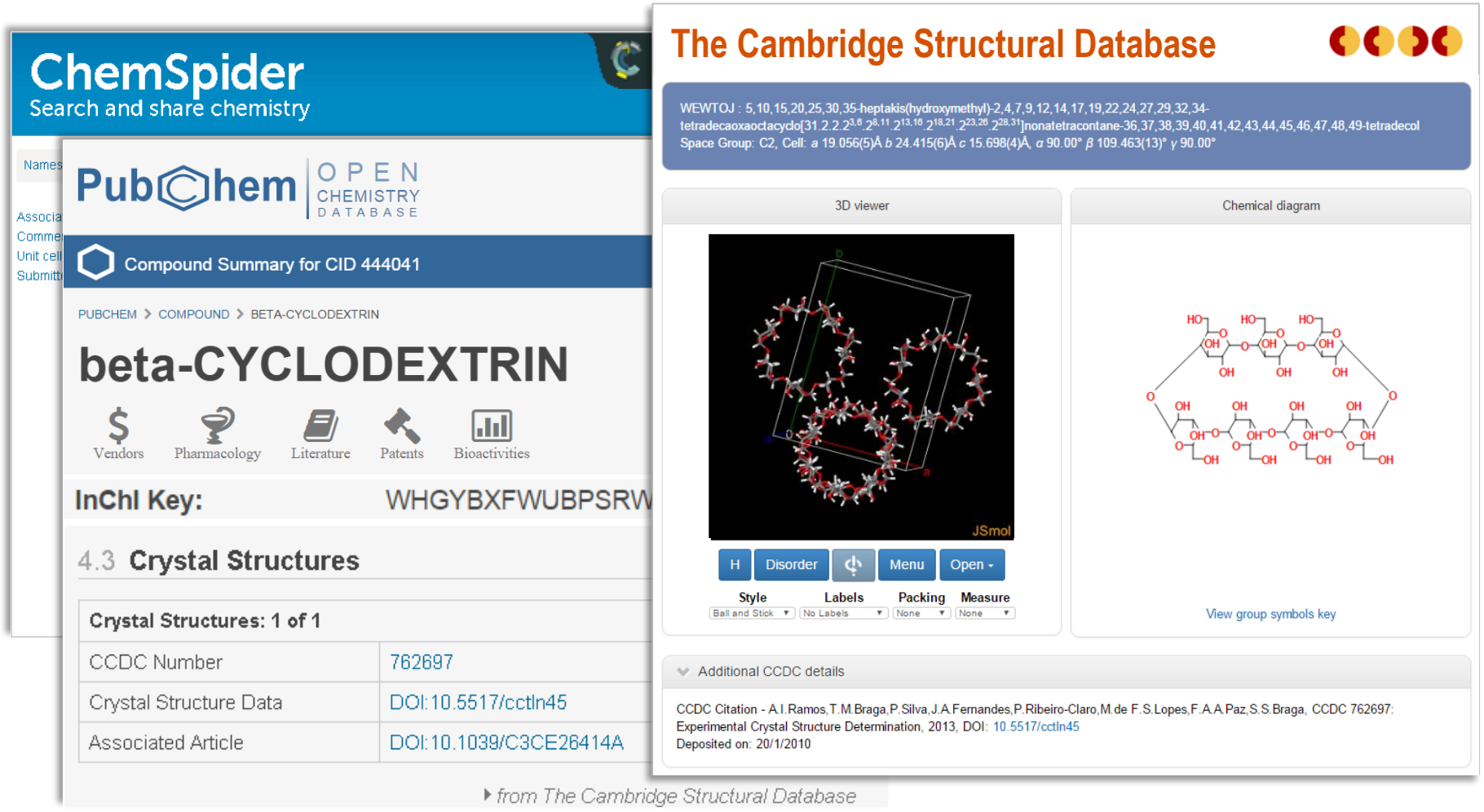




Biradicals/Zwitterions from Thermolysis of Enyne−Isocyanates
doi:10.1021/jo049716t

The CSD provides a chemically meaningful representation of the structure studied

Discoverability



Standard Chemical Representations

- Linking enabled by InChI - the IUPAC International Chemical Identifier
- InChI works well for organic molecules
- Doesn't work so well for metal-organics, polymers, mixtures - i.e. at least 50 % of the CSD
- Representation challenges in chemistry are not specific to InChI

Coordination of Data Management Policy and Practice across ICSU Unions/Disciplines in an Open Data World

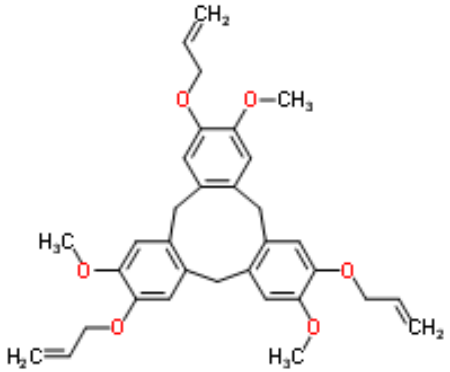
Tuesday 13 September, 16:00

SciDataCon 2016

- The role of IUPAC in the curation of chemical data in the digital age
- The experience of the IUCr

IUPAC

InChI TRUST

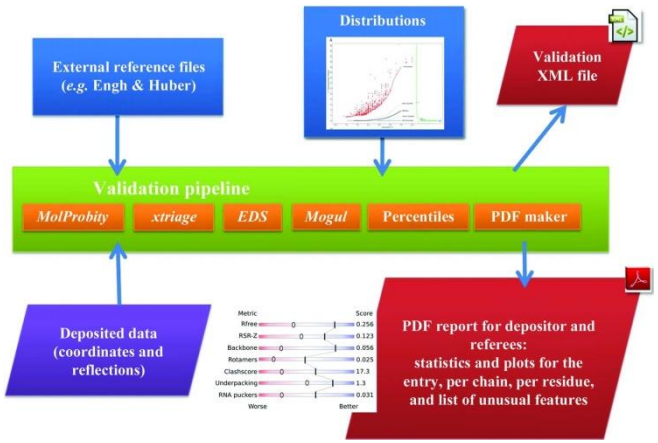


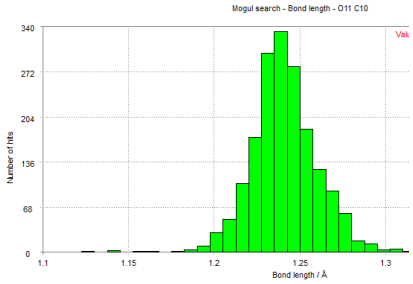
Standard InChI:
`InChI=1S/C33H36O6/c1-7-10-37-31-19-25-13-23-17-29(35-5)33(39-12-9-3)21-27(23)15-24-18-30(36-6)32(38-11-8-2)20-26(24)14-22(25)16-28(31)34-4/h7-9,16-21H,1-3,10-15H2,4-6H3`

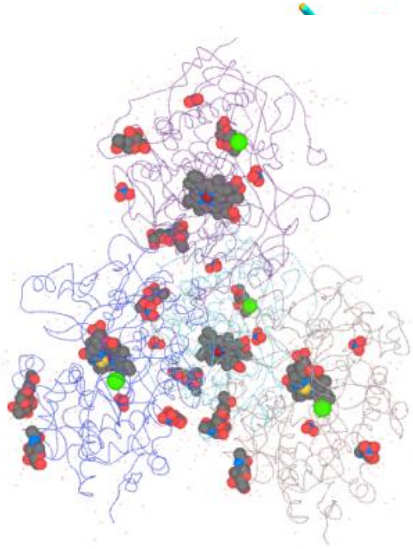
Standard InChIKey:
`IZHKSTHBLQRIOW-UHFFFAOYSA-N`

- **A CSD-based PDB Chemical Component Model Data File is now available**
  - best representative CSD coordinates for PDB chemical components
  - matching molecules identified using InChI and CCDC Web/Python API

- **PDB Validation Pipeline applies CSD-based geometr**
  - uses Mogul - a knowledge base of molecular geo



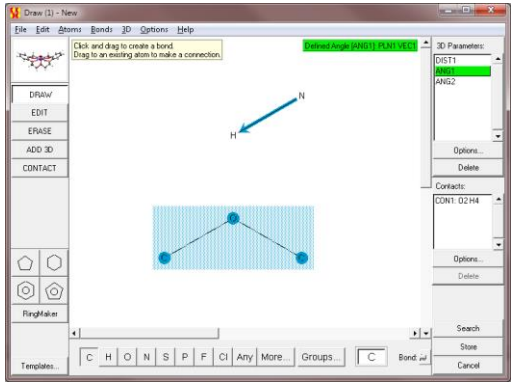




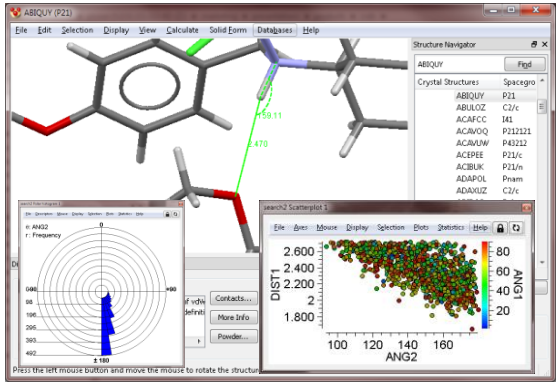
Mol	Type	Chain	Res	Link	Bond lengths			Bond angles		
					Counts	RMSZ	#|Z| > 2	Counts	RMSZ	#|Z| > 2
3	NAG	A	603	1,3	12,14,15	1.75	2 (16%)	15,19,21	2.05	2 (13%)
3	NAG	A	604	3	12,14,15	1.32	1 (8%)	15,19,21	1.78	4 (26%)
3	NAG	B	603	1,3	12,14,15	1.37	1 (8%)	15,19,21	2.89	7 (46%)
3	NAG	B	604	3	12,14,15	0.95	1 (8%)	15,19,21	1.96	5 (33%)

CSD-System

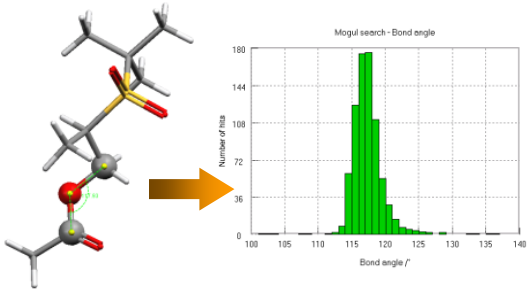
ConQuest: Advanced 3D searching



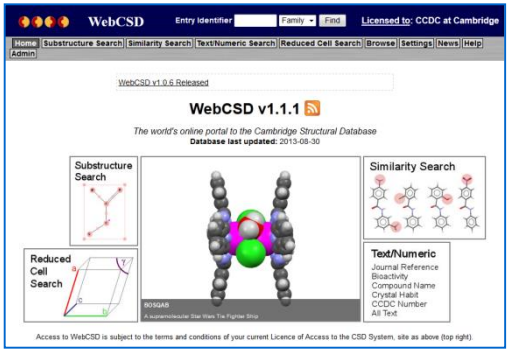
Mercury: Visualisation & data analysis



Mogul: Molecular geometry analysis



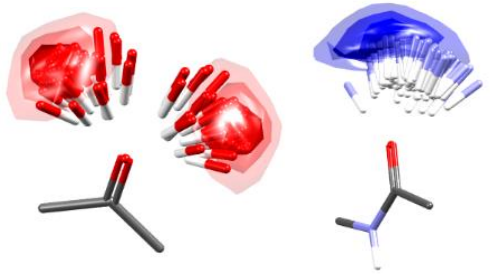
WebCSD: On-line portal to the CSD



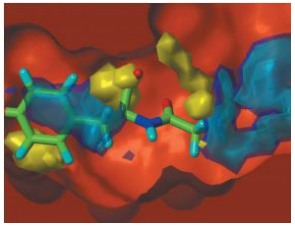
CSD Python API: Custom search and analysis

```python
import sys

from ccdc import conformer
from ccdc import io

args = parser.parse_args()

mol_reader = io.MoleculeReader(args.inmolfn)
engine = conformer.GeometryAnalyser()

molecules = []
min_unusual_torsions = sys.maxint
for (idx, molecule) in enumerate(mol_reader):
    molecule.standardise_aromatic_bonds()
    molecule.standardise_delocalised_bonds()

    # Do the analysis
    geometry_analysed_molecule = engine.analyse_molecule(molecule)

    # Count number of unusual torsions
    molecule.unusual_torsions = [
        t for t in geometry_analysed_molecule.analysed_torsions
            if t.unusual and t.enough_hits]
    num_unusual_torsions = len(molecule.unusual_torsions)
    molecule.num_unusual_torsions = num_unusual_torsions
    molecules.append(molecule)

    if num_unusual_torsions < min_unusual_torsions:
        min_unusual_torsions = num_unusual_torsions
```

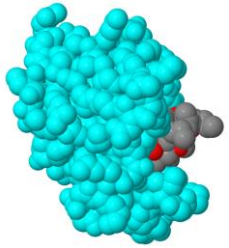
IsoStar: Molecular interaction analysis



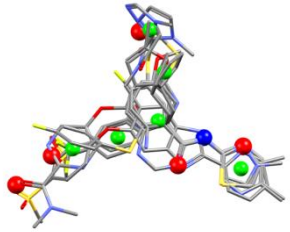
CSD-Discovery: Selected Applications



SuperStar: Knowledge-based prediction of intermolecular interactions based on data from small molecule crystal structures



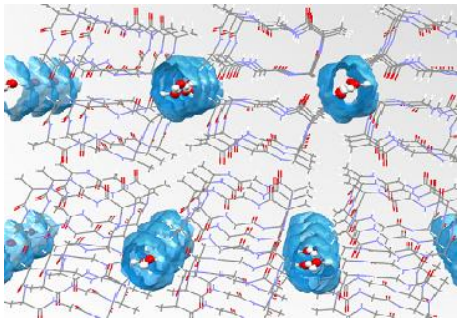
GOLD: Protein-ligand docking - virtual screening, lead optimisation and binding mode prediction



Ligand Overlay: applying the CSD to identify common binding modes, interactions and geometries of structurally diverse ligands

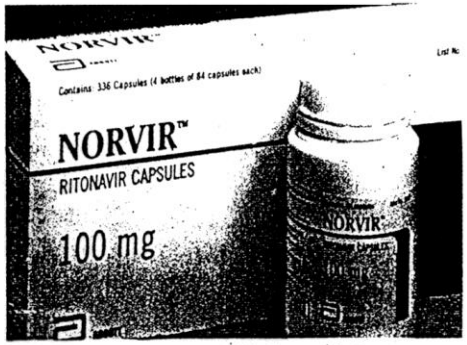
CSD-Materials

- Informatics-based solutions that aid in the understanding and prediction of solid form stability and properties



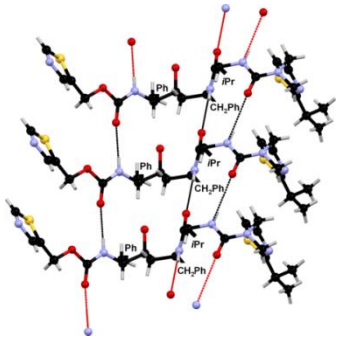
Manufacturing problems hit Abbott's HIV drug ritonavir

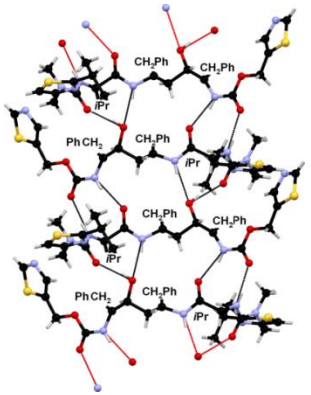
Capsules of Abbott Laboratories' protease inhibitor Norvir (ritonavir) are likely to become unavailable by the middle of August. The company has a problem with the manufacture of the anti-HIV capsules which it cannot resolve at present.



Capsules unlikely to be available from mid-August

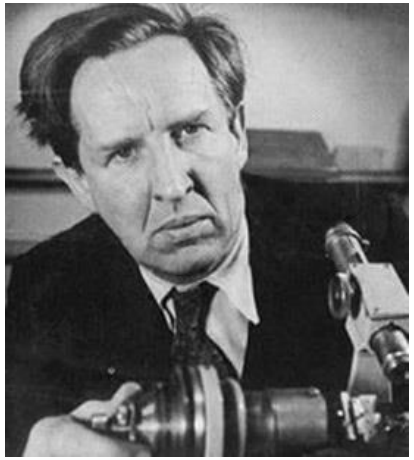
The problem relates to "undesirable" crystal formation. Abbott says that a series ples from a number of marketed batches of capsules were examined and there was no





Different crystal forms, different interactions, different solubility, different stability.

Early Vision



J. D. Bernal

Bernal and group at Stonehenge, 1948

Olga Kennard

"The database was established in 1965 to fulfil a dream of myself and a great scientist, the polymath J.D. Bernal. We had a passionate belief that the **collective use of data would lead to the discovery of new knowledge** which transcends the results of individual experiments."

Kennard, O. "From Private Data to Public Knowledge." The Impact of Electronic Publishing on the Academic Community. Ed. I Butterworth. Portland Press Ltd, 1997. 159–166.

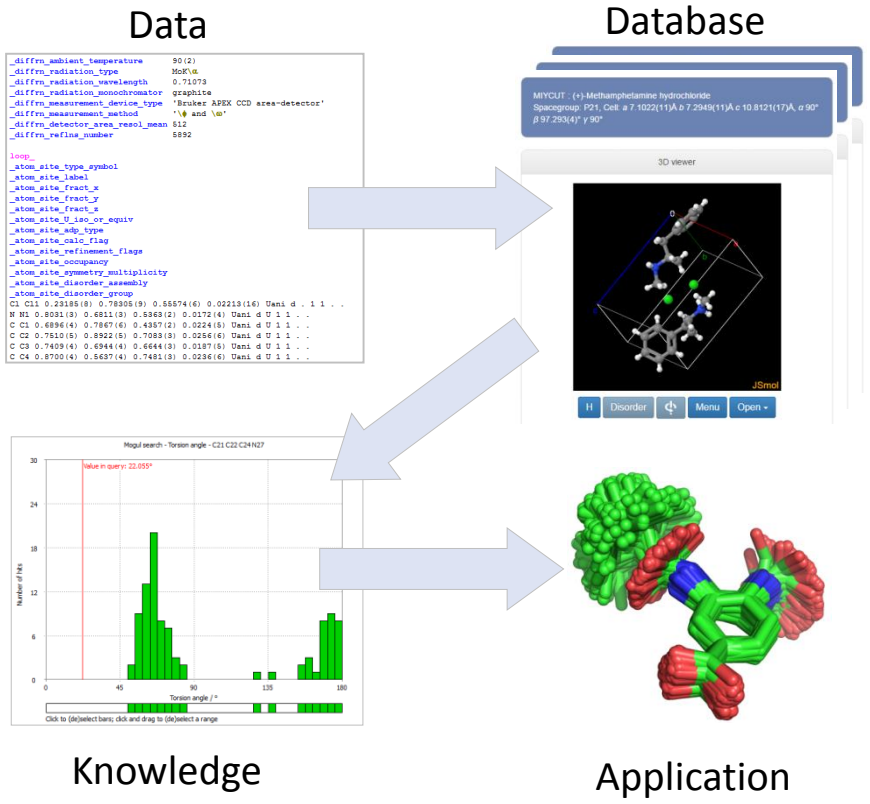
Unlocking Knowledge

The Cambridge Crystallographic Data Centre

International Data Repository
Archive of crystal structure data
High quality scientific database

Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research



Chemistry plus crystallography enables discovery and application of new knowledge

The Cambridge Crystallographic Data Centre

International Data Repository
Archive of crystal structure data
High quality scientific database

Scientific Software Provider
Search/analysis/visualisation tools
Scientific applications

Collaborative Research Organisation
New methodologies
Fundamental research

@ccdc_cambridge

ccdc.cambridge

http://www.ccdc.cam.ac.uk/

SciDataCon 2016

12 Sep 2016 (pm): Getting the incentives right

Reducing barriers to sharing data: a domain repository perspective

13 Sep 2016 (am): Sustainable Business Models for Data Repositories

Sustaining access to research data through value-added services and software