Accessibility vs Sustainability
A Balancing Act

Ian Bruno

Cambridge Crystallographic Data Centre, UK

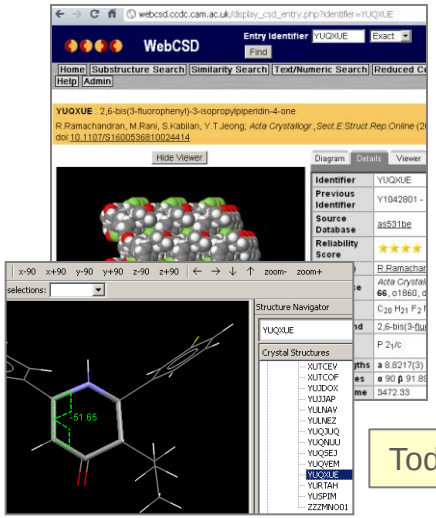
Repositories in Science & Technology: Preserving Access to the Record of Science

CENDI/NFAIS/FLICC Workshop, November 30, 2011

The Cambridge Crystallographic Data Centre

- Acquisition, evaluation, dissemination and use of the world's output of **small molecule crystal structures**
- Compilation of the **Cambridge Structural Database** (CSD) – almost 600,000 entries
- Development of **software** to enable search, analysis and use of crystal structure data
- Engaging in scientific **research**

1971-1984

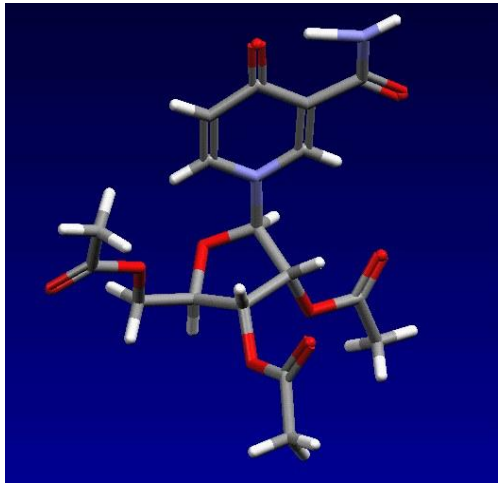


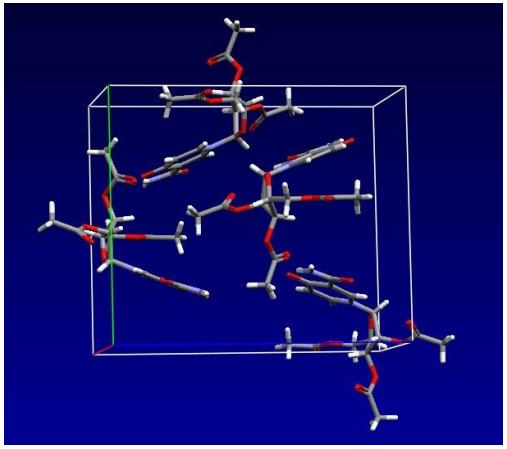
Today

Use of Crystal Structure Data

- CSD provides insights into
  - molecular dimensions and shape
  - molecular interactions

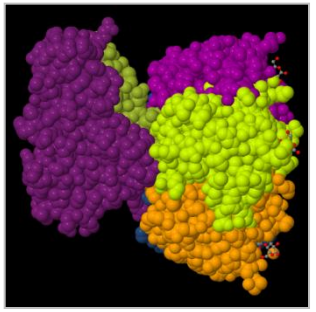
- Widely used for
  - drug design and development
  - design of new materials
  - crystal engineering
  - structure validation
  - education

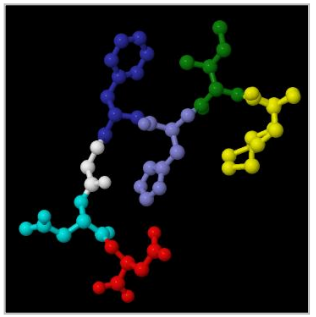


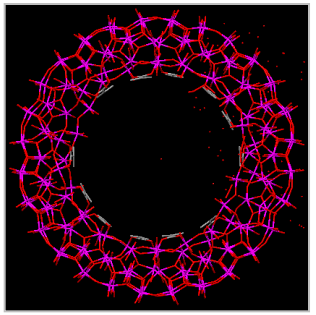


Crystallographic Databases

- Biological macromolecules
  - Protein Data Bank (PDB)
  - grant-funded, 16 or so agencies worldwide
- Organic and metal-organic structures
  - Cambridge Structural Database (CSD)
  - self-supporting, not-for-profit, registered charity
- Inorganic structures
  - ICSD: partnership between FIZ Karlsruhe and NIST
  - CRYSTMET: privately owned (Toth Information Systems)





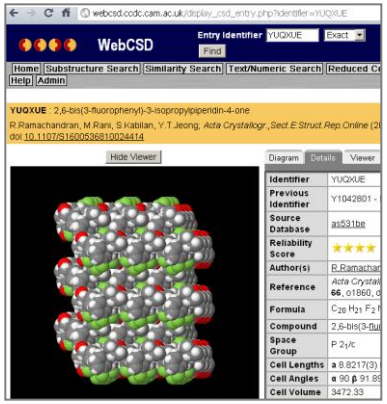


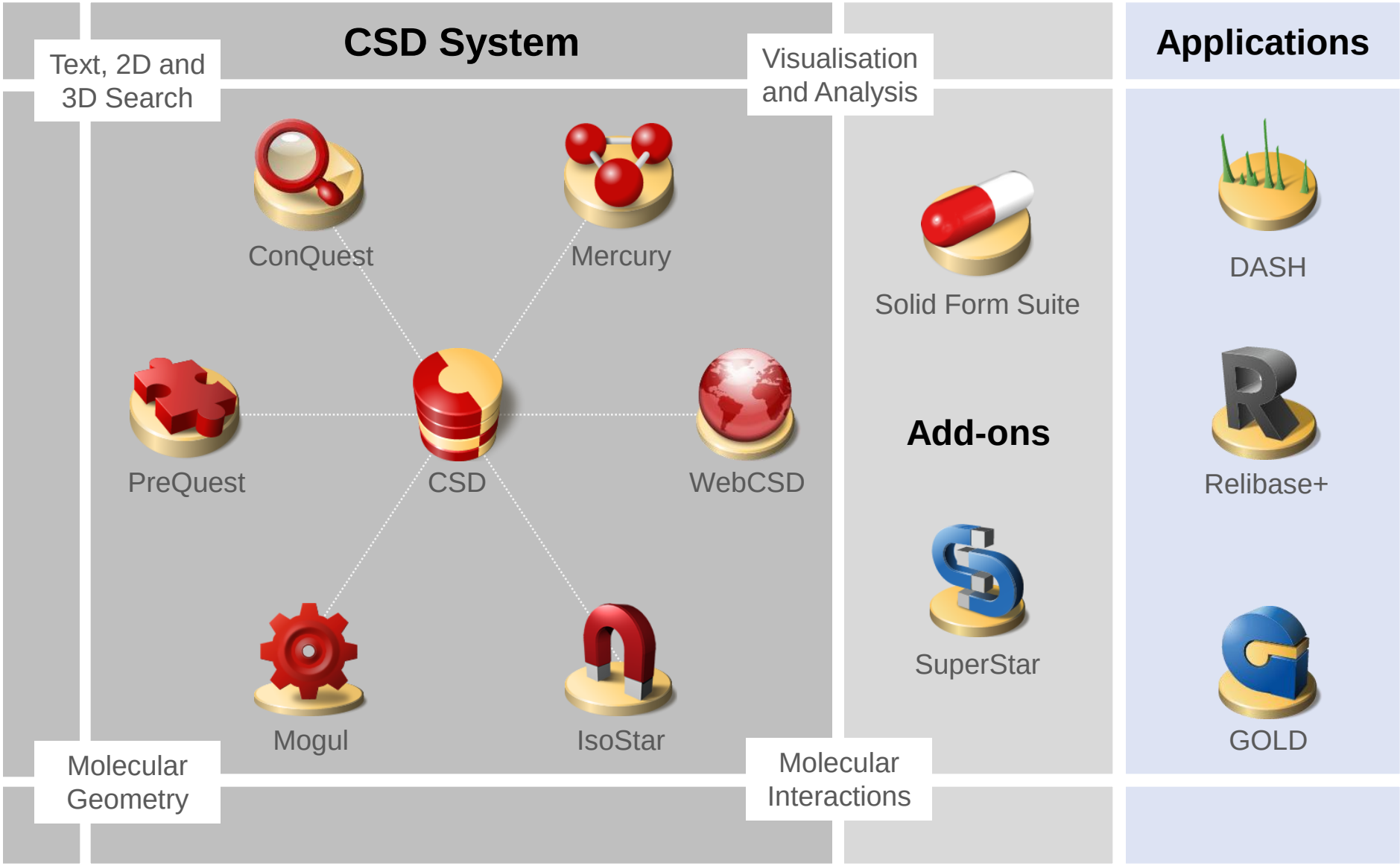
CCDC: A Brief History

- Founded 1965 within the University of Cambridge
  - initially supported by research council funding
- Independent not-for-profit organisation (registered charity) since 1989
  - around 140 industrial sites subscribe to the CSD System
  - around 1140 academic sites in 80 countries
  - financially self supporting with ~70% of income from industrial subscribers
- CCDC Software Ltd established 1998
  - sales of commercial scientific software
  - income subsidises the cost of the CSD System
- Currently ~45 permanent staff
  - editorial, software development, support & marketing, admin, research

Academic Access

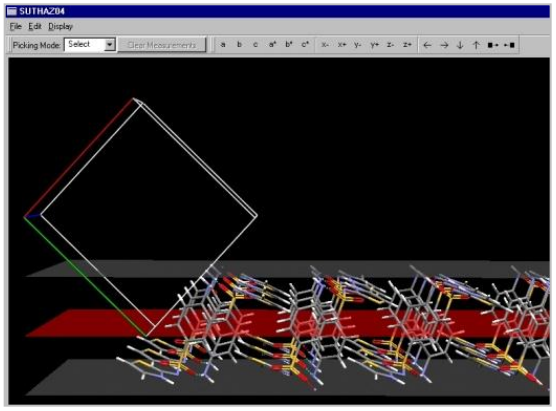
- Cost to academics heavily subsidised by income from other sources
  - in the US by between 90% and 95%
- Cost varies by region
  - in some countries supported by government grants
  - in others cost is subsidised by CCDC as much as 100%
- WebCSD allows for easy access across an institution
  - increase in uptake of academic site-wide licences



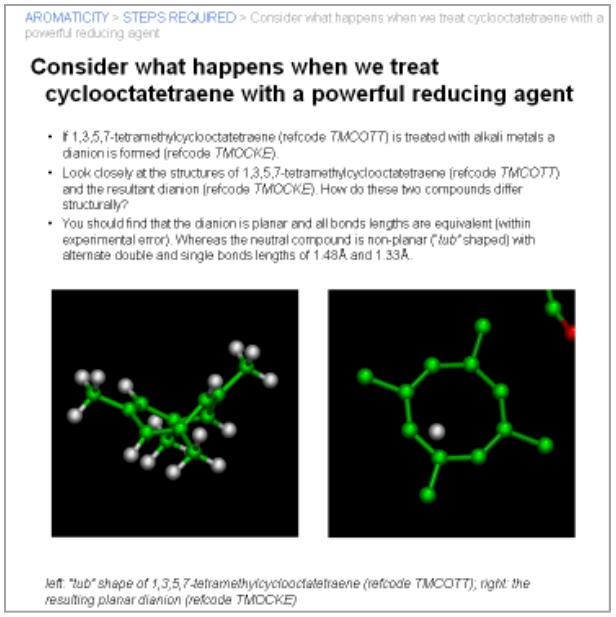

CSD System
Text, 2D and 3D Search
ConQuest
Mercury
PreQuest
CSD
WebCSD
Mogul
IsoStar
Molecular Geometry
Visualisation and Analysis
Solid Form Suite
Add-ons
SuperStar
Molecular Interactions
Applications
DASH
Relibase+
GOLD

Free Software and Services

- Software tools
  - enCIFer (validation of CIFs)
  - Mercury (crystal structure visualisation)



- Targeted subsets of curated data
  - e.g. teaching subset



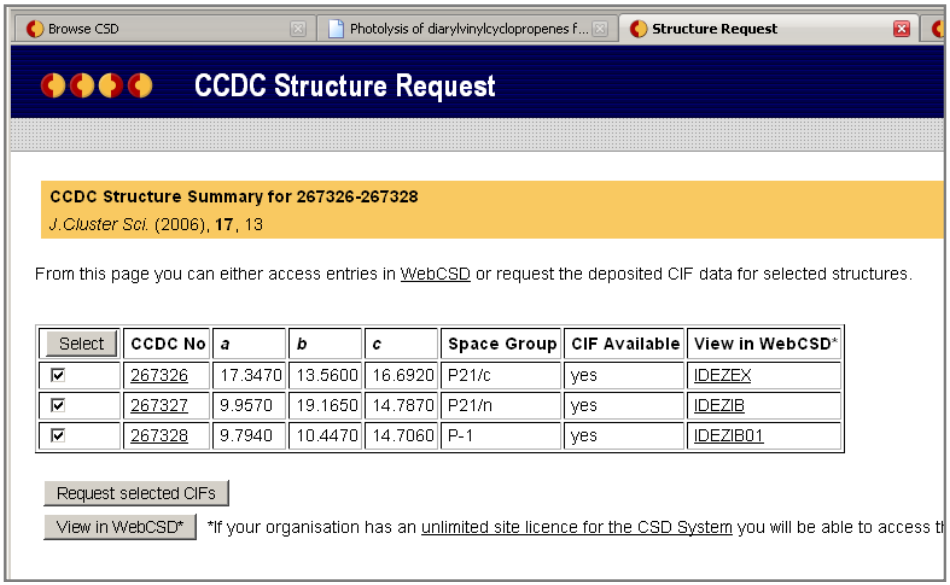
G. M. Battle, F. H. Allen, G. M. Ferrence,
J. Chem. Ed., (2010) 87, 809-812. 10.1021/ed100256k

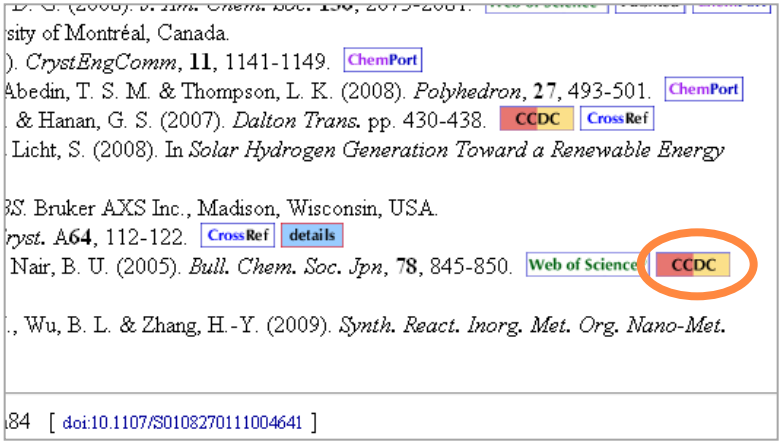
G. M. Battle, F. H. Allen, G. M. Ferrence,
J. Chem. Ed., (2010) 87, 813-818. 10.1021/ed100257t

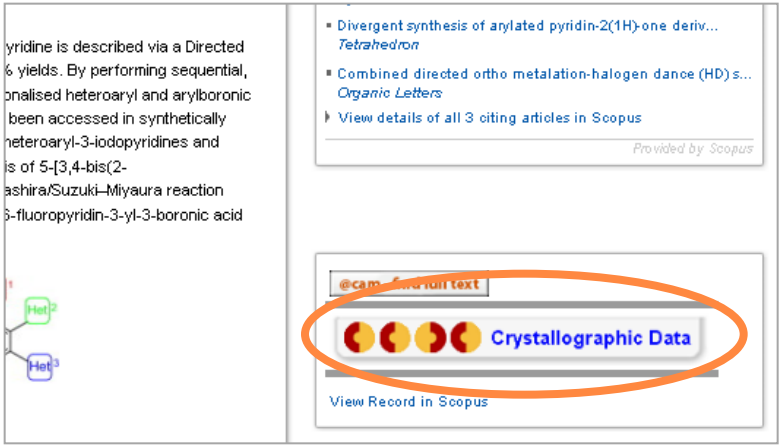
Supported by the United States National Science Foundation under Grant No. 0725294

Access to Original Deposited Data

- CCDC CIF Repository
  - individual data sets available to anyone
  - access to all electronically deposited data
  - free, no financial cost







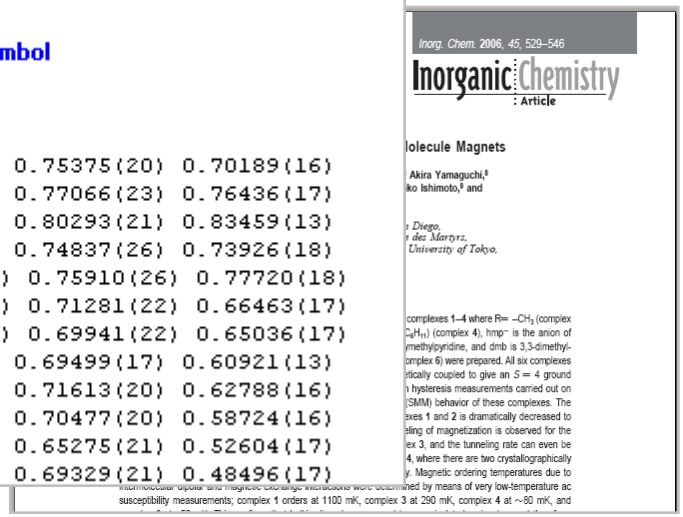
Added Value

- Aim is to add value to the originally deposited data
- **Save scientists time and effort – avoid pain**
- **Provide solutions to scientific problems**
- Editorial activities and software tools help achieve this

Overview of Editorial Process

CIF/Paper

```
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)
C4  C -0.00500(38)  0.71281(22)  0.66463(17)
H2  H -0.11609(38)  0.69941(22)  0.65036(17)
N1  N  0.11523(28)  0.69499(17)  0.60921(13)
C5  C  0.27057(33)  0.71613(20)  0.62788(16)
H3  H  0.35291(33)  0.70477(20)  0.58724(16)
C6  C  0.07764(37)  0.65275(21)  0.52604(17)
H4  H  0.12644(37)  0.69329(21)  0.48496(17)
```


Inorg. Chem. 2006, 45, 529–546

Inorganic Chemistry

: Article

olecule Magnets

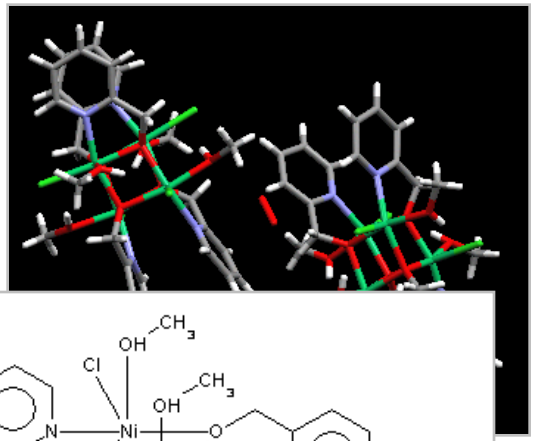
Akira Yamaguchi,[§]
ko Ishimoto,[‡] and

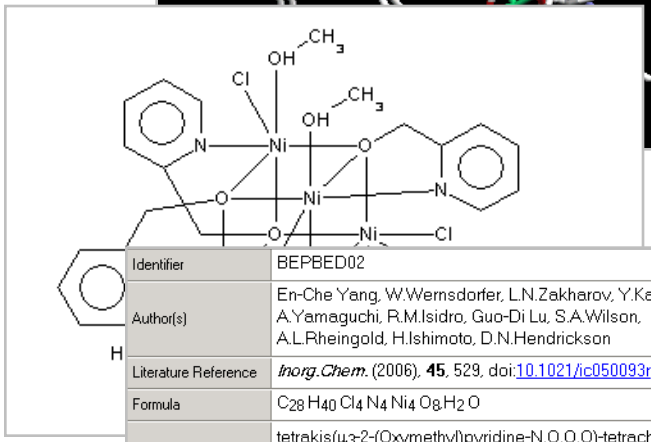
Diego,
des Martyrz,
University of Tokyo.

complexes 1–4 where R= –CH_3 (complex
C_6H_{11}) (complex 4), hmp⁻ is the anion of
methylpyridine, and dmb is 3,3-dimethyl-
omplex 6) were prepared. All six complexes
tically coupled to give an $S = 4$ ground
hysteresis measurements carried out on
SMM) behavior of these complexes. The
xes 1 and 2 is dramatically decreased to
ling of magnetization is observed for the
ex 3, and the tunneling rate can even be
4, where there are two crystallographically
y. Magnetic ordering temperatures due to
ned by means of very low-temperature ac

susceptibility measurements; complex 1 orders at 1100 mK, complex 3 at 290 mK, complex 4 at ~80 mK, and

CSD Entry



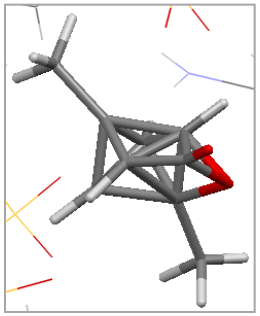


Identifier	BEPBED02
Author(s)	En-Che Yang, W.Wernsdorfer, L.N.Zakharov, Y.Karaki, A.Yamaguchi, R.M.Isidro, Guo-Di Lu, S.A.Wilson, A.L.Rheingold, H.Ishimoto, D.N.Hendrickson
Literature Reference	*Inorg.Chem.* (2006), **45**, 529, doi:10.1021/ic050093r
Formula	$C_{28}H_{40}Cl_4N_4Ni_4O_8.H_2O$
Compound Name	tetrakis(μ3-2-(Oxymethyl)pyridine-N,O,O,O)-tetrachloro-tetrakis(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.

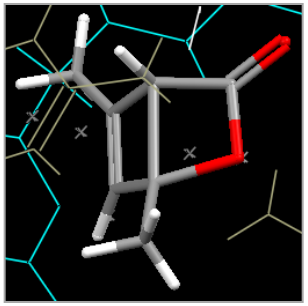
- Level of information varies between articles
- Degree of peer review of crystallography varies
- Differing chemical interpretations

Avoiding duplication of effort

- Curation of CSD aims to save others from having to duplicate effort



Unresolved disorder



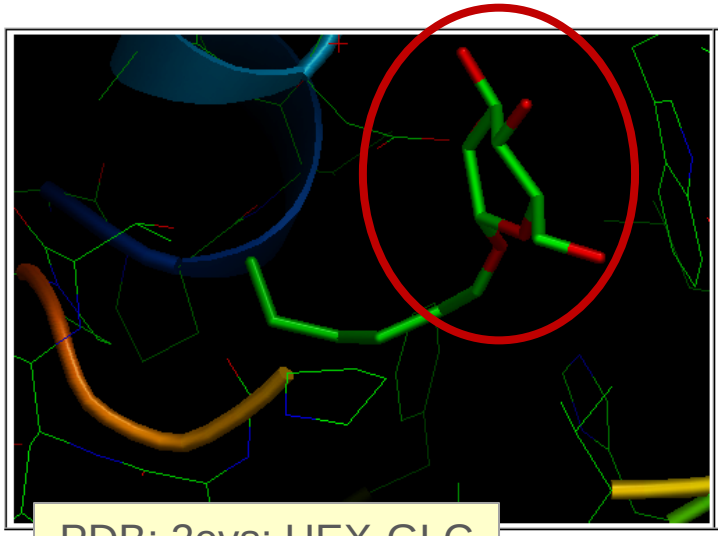
Under UV radiation the clathrated pyrone molecule converts to a disordered mixture of square-planar 1, 3-dimethylcyclobutadiene and rectangular-bent 1, 3-dimethylcyclobutadiene in van der Waals contact with a carbon dioxide molecule. The ratio of the square-planar to rectangular-bent 1, 3-dimethylcyclobutadiene clathrate is modelled with occupancies 0.6292:0.3708.

Resolved disorder with editorial comment based on discussion in paper

Key Editorial Tasks

- ✓ correct CIF syntax and match structures to publications
- ✓ analyse disorder and identify bonds
- ✓ for polymers, find an acceptable monomer unit
- ✓ assign bond types, atom charges and hydrogen positions
- ✓ generate chemical diagram and compound name

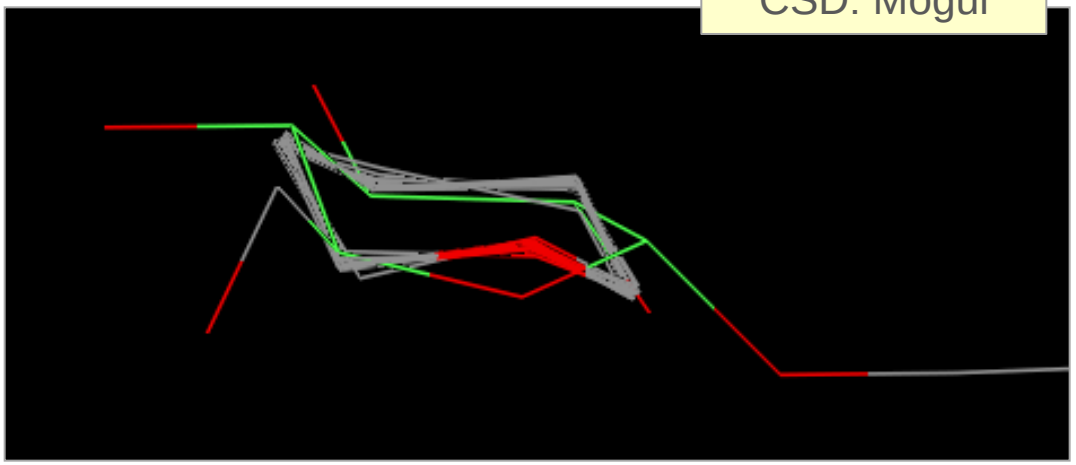
Assessment of molecular geometry



PDB: 2evs: HEX-GLC

angle			
	LIM__HEX...		
		C4 C3 C2	Not unusual (enough hits) 926
		C3 C4 C5	Not unusual (enough hits) 828
		O5 C1 C2	Unusual (enough hits) 3421
		C1 C2 C3	Unusual (enough hits) 3296
		O5 C5 C4	Unusual (enough hits) 891
		C1 O5 C5	Unusual (enough hits) 5171
torsion			
	LIM__HEX...		
ring			
	LIM__HEX...		
		O5 C1 C2 C3 C4 C5	Unusual (enough hits) 390

CSD: Mogul

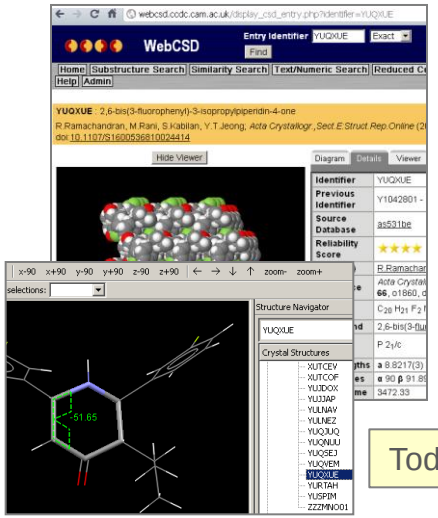


CCDC are collaborating with PDB to incorporate tools based on CSD data into PDB deposition and validation procedures.

Changing Times

- The economic environment has changed
- Science and technology has evolved
- Attitudes and expectations regarding data are changing
- *Does our current business model allow us to maximise accessibility **and** ensure sustainability?*

1971-1984

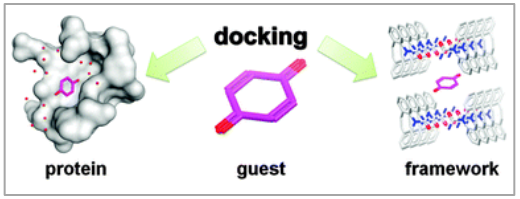


Today

Pressures on current business model

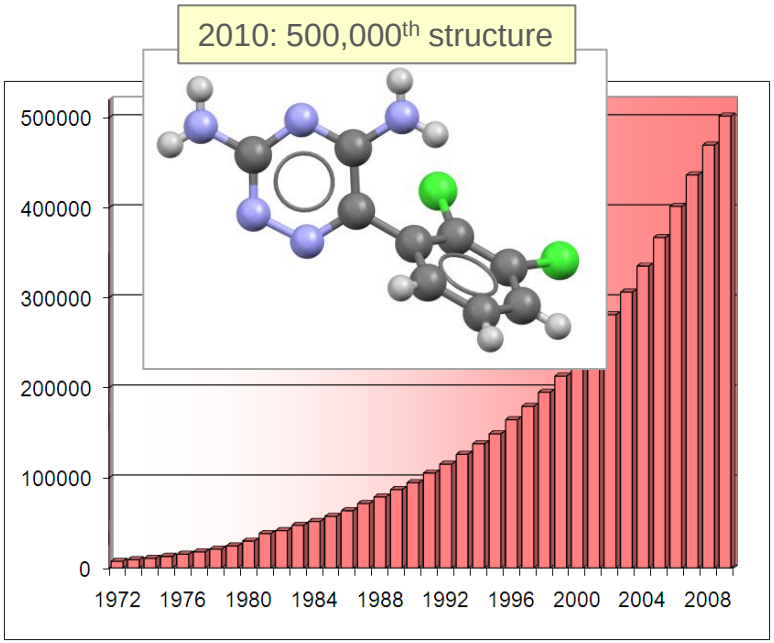
- Pharmaceutical sector
  - future of R&D in big pharma uncertain
  - upsurge in out-sourcing/CROs

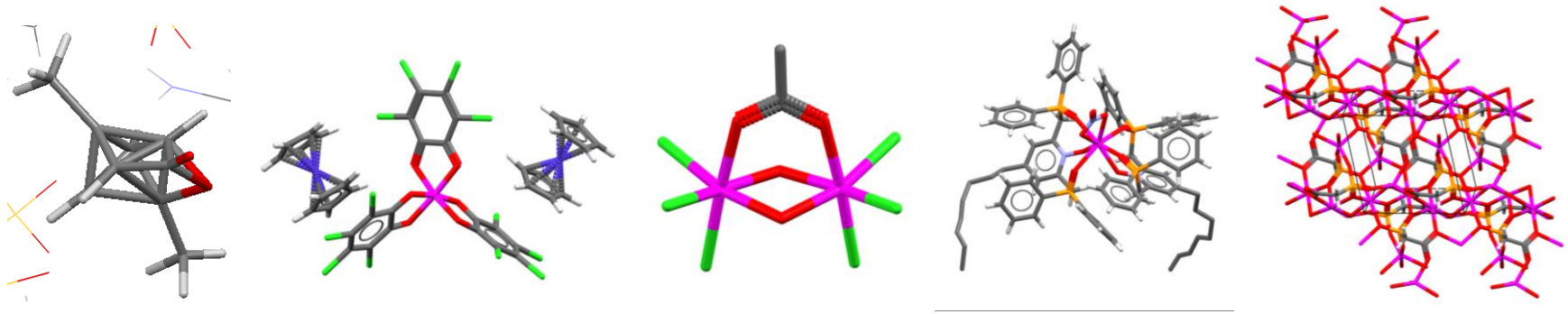
- Competing software
  - direct competition with commercial software
  - computational alternatives to experimental data



Scientific Pressures

- Throughput is increasing
- Complexity and diversity is increasing
- Issues faced with deposited data
  - disorder
  - poor geometry
  - polymeric structures
  - incomplete chemical representation





Attitudes and Expectations

Data-Driven High-Throughput Prediction of the 3-D Structure of Small Molecules: Review and Progress

J. Chem. Inf. Model., 2011, 51 (4), pp 760–776, doi:10.1021/ci100223t

We also hope that future versions of COSMOS, or other similar systems, will be able to **achieve the greater degree of data and software openess** that is indispensable for real scientific progress in the field.

One obstacle in this area may be **the closed nature of the CSD**, which unlike the PDB cannot be used without severe restrictions, even for academic research purposes*.

This is yet another example of the unfortunate state of affairs in chemoinformatics, where an overly zealous culture of closeness and secrecy, sometimes related to short-term profits, have greatly hampered scientific progress.

* *COSMOS is available at http://cosmos.igb.uci.edu/ with express permission from the CCDC. The number of molecules that can be uploaded at any one time is limited to 100 and the service ought not to be used for commercial benefit or gain. See J. Chem. Inf. Model., Article ASAP, August 30 2011 doi:10.1021/ci2002523 for a response from CCDC.*

Attitudes and Expectations

CCDC: Reasons why sourceCIF data must be Open

petermr's blog, A Scientist and the Web, August 2011

http://blogs.ch.cam.ac.uk/pmr/2011/08/31/ccdc-reasons-why-sourcecif-data-must-be-open/

"sourceCIFs" are raw data created as part of a crystallographic experiment by scientists (not in the CCDC) and required by community norms as part of the scholarly publication process. Some are published Openly, but others are sent by the author or publisher to CCDC in an exclusive process. **CCDC then control the further distribution of this data which are either made available in trivial amounts** (less than 0.1% of the CCDC's holding of sourceCIFs*) **or significant financial subscription (which many institutions cannot afford[+]).**

* *CCDC makes 100% of CIFs freely available but does not currently allow bulk download of the complete collection*

[+] *To our knowledge, no academic scientist has been denied access to the CSD due to genuine lack of funds*

Attitudes and Expectations

Letter from the COD* to the IUCr and the CCDC

April 2005 http://www.crystallography.net/petition/letter4.html

> We do not want to accept the idea that PDB or the AMCSD are able to obtain funding in order to make this free Web access possible, and that the ICSD, **CSD**, CRYSTMET and ICDD would not (**though, probably, these databases are already obtaining large public funding**[+]).

* *COD is the Crystallography Open Database, currently developed at the Institute of Biotechnology, Vilnius, Lithuania. The Vilnius development group is financed by the Research Council of Lithuania, contract No. MIP-124/2010*

[+] *CCDC receives no public funding to support creation of the CSD. In some regions public funds subsidise the cost of access by the academic community.*

What people say

- Some are campaigning for data to be “Open”
- Others complain that access isn’t “free”
- Funders talk in terms of “public domain”, “data sharing”
- Some think we’re worth every penny
- Many don’t say anything

Open Data Principles

Open Knowledge Foundation

- The Open Knowledge Definition (OKD):

> A piece of content or data is open if anyone is free to use, reuse, and redistribute it — subject only, at most, to the requirement to attribute and share-alike.

- Open does not necessarily mean cost-free
  - available as a whole and at no more than a reasonable reproduction cost
- Discrimination against persons, groups, fields of endeavour prohibited
  - does not allow different treatment of e.g. industrial and academic users
- Provide in a form where there are no technological barriers to access

Arguments why Data should be Open

- Data is donated by the community and should be freely available to the community
- Public funding supports generation of the data and the public should have access
- Limiting access to data can be argued to be holding back science
- Opening up data potentially encourages fresh thinking, drives innovation and future growth

The Hargreaves Review

Commissioned in November 2010 by UK Prime Minister

Identify barriers to growth within the IP framework ... particularly focus on how the IP system can be improved to help the new business models arising from the digital age.

Heavy focus on the Creative Industries but much to say about text and data mining

Research scientists ... are today being hampered from using computerised search and analysis techniques on data and text because copyright law can forbid or restrict such usage

http://www.ipo.gov.uk/ipreview-finalreport.pdf

UK Government response

Bring forward proposals for a substantial opening up of the UK's copyright exceptions including a wide non-commercial research exception covering text and data mining

http://www.ipo.gov.uk/ipresponse-full.pdf

Mapping Data

- Ordnance Survey – national mapping agency of Great Britain
  - April 2010, announced OS Open Data
  - selected data sets made available under CC-BY-like licence

John Denham
Communities Secretary
(*quondam*)

The move to free up public data encourages fresh thinking - people re-using information in different and more imaginative ways than may have originally been intended ...

*Increasing access to Ordnance Survey data will attract a **new wave of entrepreneurs** and result in new solutions to old problems that will benefit us all. It will also **drive a new industry, creating new jobs and driving future growth**.*

http://www.ordnancesurvey.co.uk/oswebsite/media/news/2010/April/OpenData.html

What we perceive people want

- One-click access to crystallographic data from within other resources
- Data available as soon as it is published
- Unified search across all crystallographic data
- Access to unpublished crystal structure data
- Ability to make services based on the CSD available to the community
- Integration of functionality and interoperability with other systems

Use, Reuse, Redistribution

CIF Repository	CSD
Deposited Data – Available for free	Added value – Subscription required
◦ Technical barriers – forms need to be filled, data e-mailed ◦ Complete collection of CIFs not currently available for bulk download ◦ Copyright statements are unhelpful or unclear	◦ No restrictions on what can be done within site boundaries ◦ Derivative works can be made publically available with written permission

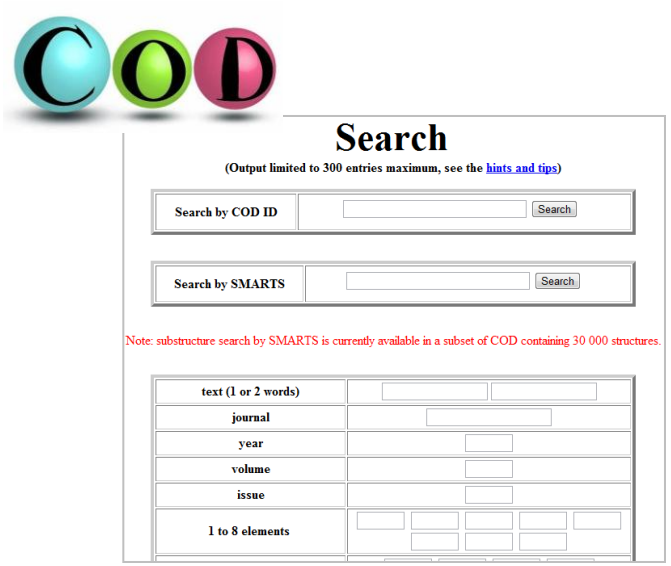
- Statements applied to deposited data reflect uncertainty over rights:
  - *may contain copyright material of the CCDC or of third parties*

Relationship with Publication Processes

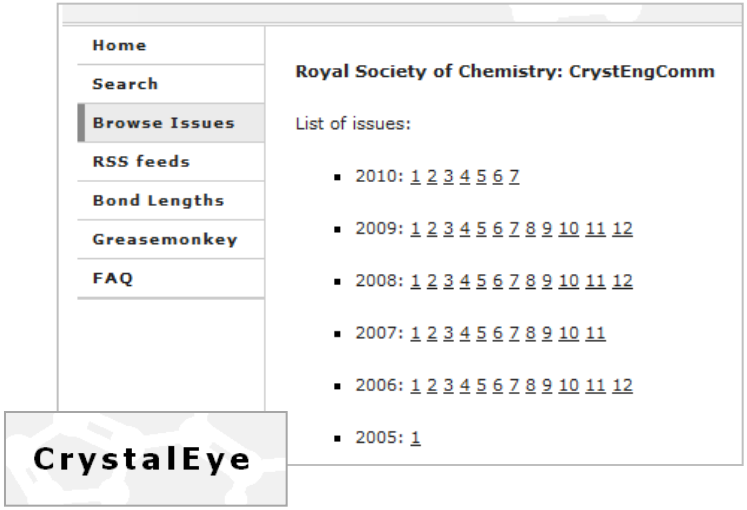
- Deposition of data with CCDC typically linked to publication
  - some data comes from publishers, some deposited directly with us
- Author agreements grant publishers rights over supplementary data
  - publisher claims rights then makes data available to the community
  - non-exclusive agreements between publisher and author
  - publisher claims blanket rights over “article”
- Currently, CCDC undertake to include data in CSD and CIF Repository
  - should we presume that depositors expect anything more?

Open Alternatives

- Crystallography Open Database (COD)
  - collection of CIFs, syntax correction
  - basic searching
  - recently started assigning chemistry



- CrystalEye
  - automatically assigns chemistry to CIFs
  - basic searching, bond length distributions
  - emphasis on current awareness



A lower cost “Open” Model?

- Structure acquisition
  - Automated harvesting of data published on web sites, in repositories
- Assignment of chemistry
  - Algorithmic processing enhanced by community curation
- Dissemination
  - Basic access through light-weight web services
- Advanced analysis
  - Open Source tools developed by the community

- Minimal overheads
- Financially supported by e.g. funding agencies or consultancy
- No subscriptions, no restrictions

A lower cost “Open” Model?

- Automatic harvesting, algorithmic assignment
  - currently difficult to ensure comprehensive coverage
  - data that is available isn’t sufficient to give complete picture

Needs development *and adoption* of tools to better capture data at source

- Community input – some foundation
  - tradition of crystallographic co-editing
  - developing body of relevant Open Source libraries and tools

Risks shifting burden back onto the wider scientific community

- Economic factors
  - market for consultancy as yet unproven
  - relying on funding agencies for long-term support not without risk

There is still value we can offer to the scientific community through expert curation of crystallographic data and development of associated software

Alternative Funding Models

Google

advertising

freemium

publically funded

publisher supported

public appeal

venture capital

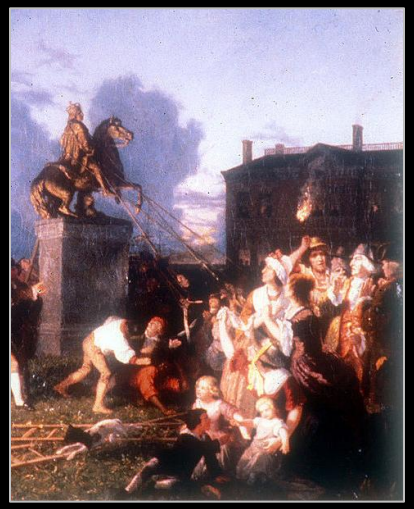
pay-to-publish

pay-per-download

open source
paid support

Purchase This Content
Choose from the following options:
» USD35.00 for 48 hours of access

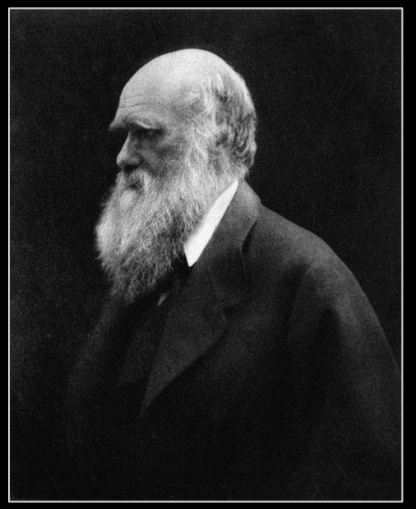
pay-per-view



Revolution

or

Evolution?



Improve access to deposited data sets

Anyone able to access individual deposited data sets for any purpose

Widely linked or embedded in other resources

PubChem

www.chemspider.com

Google

WIKIPEDIA
The Free Encyclopedia

Priorities

- Remove technical barriers and enhance linking mechanisms
- Establish conditions of use consistent with modern age and desires of rights holders

Builds on existing services that make data freely available to the community

Widen access to added-value services

Promote WebCSD as a portal to added value in the non-profit sector

Engage with the community to leverage public funds to support this

Value Proposition

- Income generated supports CCDC's role in data curation and preservation
- Wider community gains access to search and analysis tools
- Potential to exploit crystallographic data in science education

Builds on existing Web-based services that deliver added-value

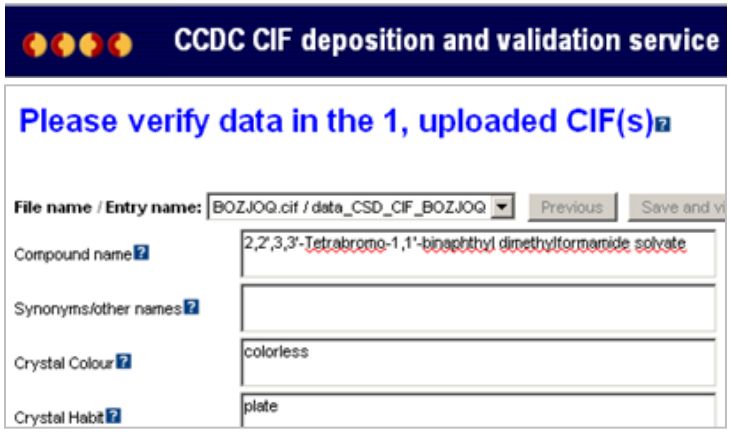
Streamlining and cost reduction

- Initiatives aimed at improving internal processes are ongoing
  - reworking internal systems to improve efficiency and data storage
  - establishing more effective deposition and validation routes
  - automated tools for curation and assessment of reliability
  - services that enable experts to provide insight at point of deposition

```
Low probability bond lengths:
C5-C6   1.405, av(CSD) = 1.505, prob = 0.001
C2-C3   1.345, av(CSD) = 1.514, prob = 0.001
C3-C4   1.338, av(CSD) = 1.514, prob = 0.001
C3-C6   1.798, av(CSD) = 1.546, prob = 0.001
Reliability level: 2
```

Decifer: Chemical Assignment + Reliability Report

Acta Cryst. (2011). B67, 333-349



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

Finding a Balance

- We still need to rely on commercial income streams
  - continue to seek out opportunities for development of novel scientific applications
  - maintain partnerships and collaborations with industrial customers
- We recognise we can’t do everything
  - provide access to toolkits, APIs, web services etc. that allow others to innovate
  - develop framework that enables others to share applications based on the CSD
- We can do more to improve accessibility to data
  - deposited data – remove technical and other barriers
  - value-added data and services – through WebCSD

We aspire to make data as openly available as possible

We feel we have a good understanding of the challenges – and the opportunities

We can see how we can build on our current business model

Is this enough to provide the accessibility the community needs and ensure sustainability?

Acknowledgements

- Colin Groom (Executive Director) and CCDC Staff

- CCDC Board of Governors

Professor Guy Orpen	University of Bristol, UK
Professor Christer Aakeroy	Kansas State University, USA
Professor Neil Isaacs	University of Glasgow, UK
Dr Andrew Leach	GlaxoSmithKline
Professor Val Gillet	University of Sheffield, UK
Professor Reiko Kuroda	University of Tokyo, Japan
Dr James Milne	RSC Publishing
Professor Robert Glen	University of Cambridge, UK

- The scientific community who contribute to the success of the CSD