The eCrystals Federation

Repository Curation Service Environments (RECURSE) Workshop
National e-Science Centre, Edinburgh

4th International Digital Curation Conference
"Radical Sharing: Transforming Science?"
1-3rd December 2008
Edinburgh, Scotland

Manjula Patel
UKOLN, University of Bath, UK

Context

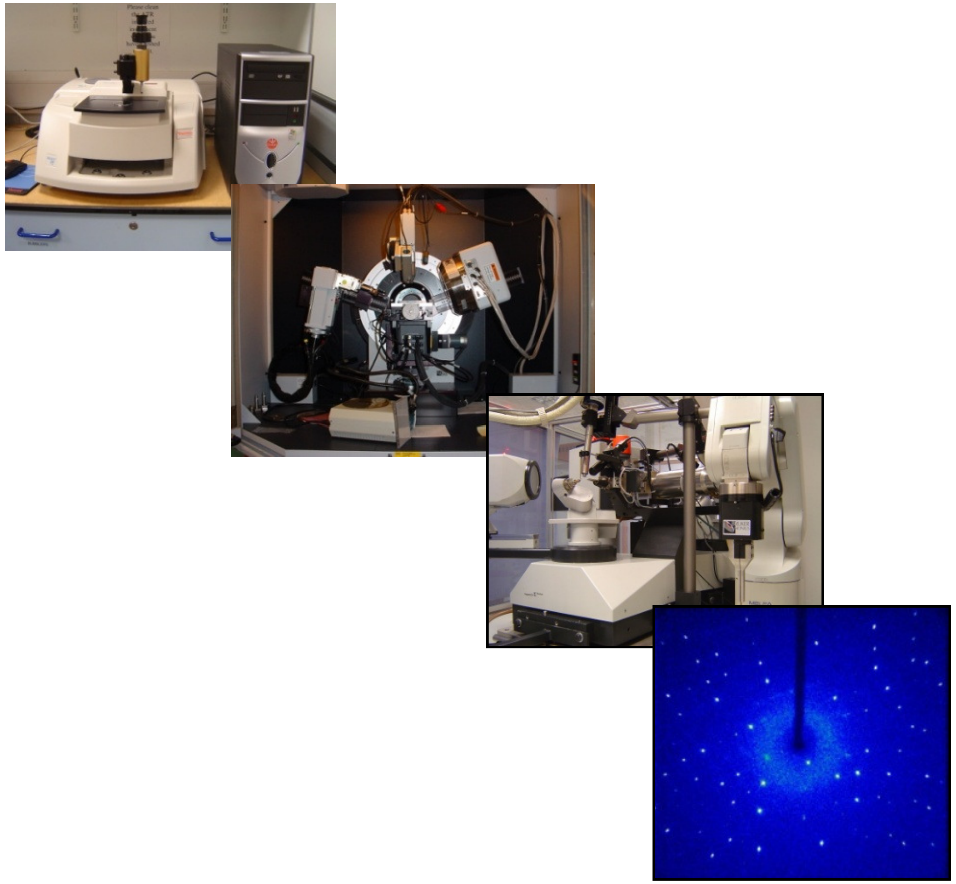
- The data deluge
  - Advances in instrumentation, data storage technologies, computational power and improvements in algorithms
  - Development of grid and cyber infrastructures
- Actual nature of science is changing
  - Mining and analysis of large datasets (e.g. Protein Data Bank, GenBank)
  - Open Science (e.g.Open Notebook Science; myExperiment)
- High quality data are the raw materials of contemporary e-science
  - Verification; Validation; Replication
  - Predictive science
  - Innovative scientific endeavour
- S. Carlson, *Lost in a Sea of Science Data*, The Chronicle of Higher Education, June 2006

> “To vet experiments, correct errors, or find new breakthroughs, scientists desperately need better ways to store and retrieve research data”

> “Data from Big Science is … easier to handle, understand and archive. Small Science is horribly heterogeneous and far more vast. In time Small Science will generate 2-3 times more data than Big Science.”

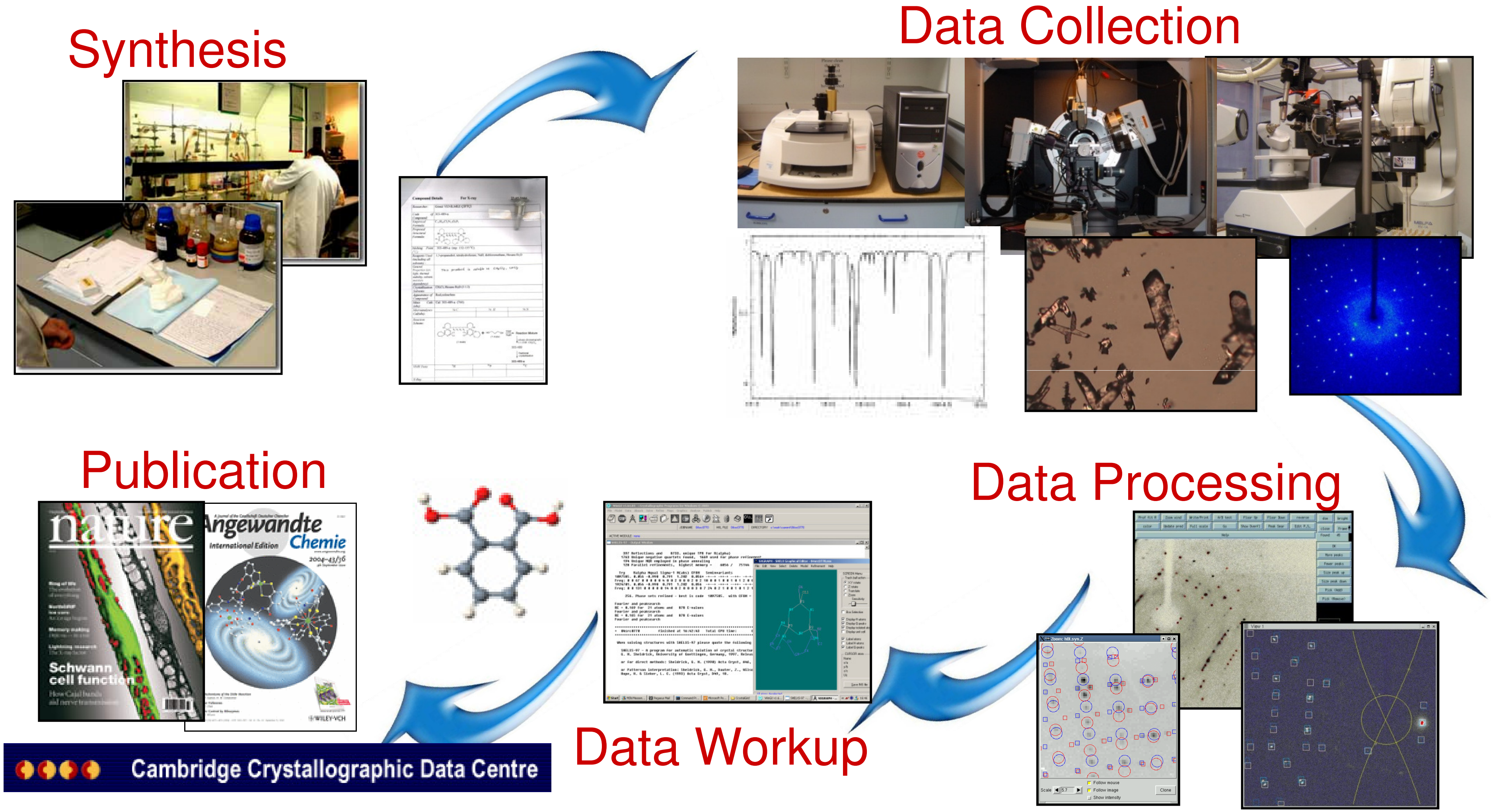
Crystallography –The Science

- Sub-discipline of chemistry
- Concerned with determining the structure of a molecule and its 3D orientation with respect to other molecules in a crystal
- Analysis of diffraction patterns obtained from X-ray scattering experiments
- Focus on laboratory based experimental technique of chemical crystallography undertaken at the EPSRC National Crystallography Service (NCS), UK



Images from Simon Coles (NCS), 2006

Data Generation



Synthesis

Data Collection

Data Processing

Data Workup

Publication

Cambridge Crystallographic Data Centre

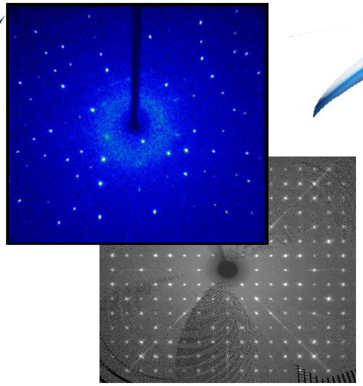
Adapted from Simon Coles (NCS), 2007

Data Volumes

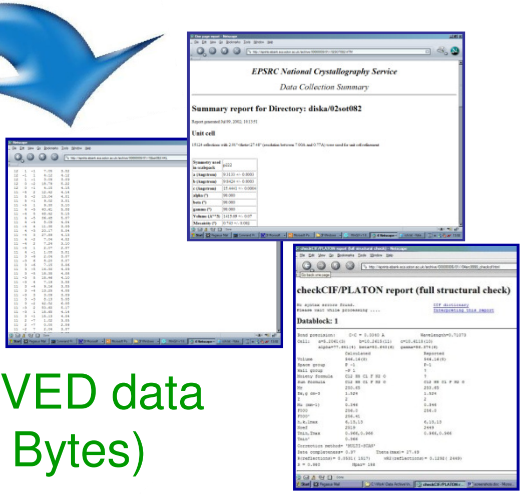
Science & Technology Facilities Council

Laboratory; Institution

RAW data (G Bytes)



DERIVED data (M Bytes)



EPSRC National Crystallography Service

Data Collection Summary

Summary report for Directory: diska/02sot082

Unit cell

Symmetry used in scalepack	p222
a (Angstrom)	9.3133 +/- 0.0003
b (Angstrom)	9.8424 +/- 0.0003
c (Angstrom)	15.4441 +/- 0.0004
alpha (°)	90.000
beta (°)	90.000
gamma (°)	90.000
Volume (A**3)	1415.69 +/- 0.07
Mosaicity (°)	0.743 +/- 0.002

checkCIF/PLATON report (full structural check)

```
No syntax errors found.
Please wait while processing ....
```

Datablock: 1

Subject Repository; Data Centre; Public Domain

Cambridge Crystallographic Data Centre

RESULTS data (K Bytes)

Adapted from Simon Coles (NCS), 2007

Community & Current Practice (1)

- Relatively organised approach to data (crystallography data are highly structured)
- Convention is to share derived or reduced data, access to raw data is rare
- Crystallography Information File (CIF) is a de facto exchange standard
  - Maintained by International Union of Crystallography (IUCr)
- Heterogeneity in instrumentation and associated software
- Established system for publishing crystallographic data in UK (Cambridge Crystallographic Data Centre-CCDC)
- Other major databanks
  - Germany (inorganic molecule database)
  - Canada (metals database)
  - US (Protein Data Bank -PDB)

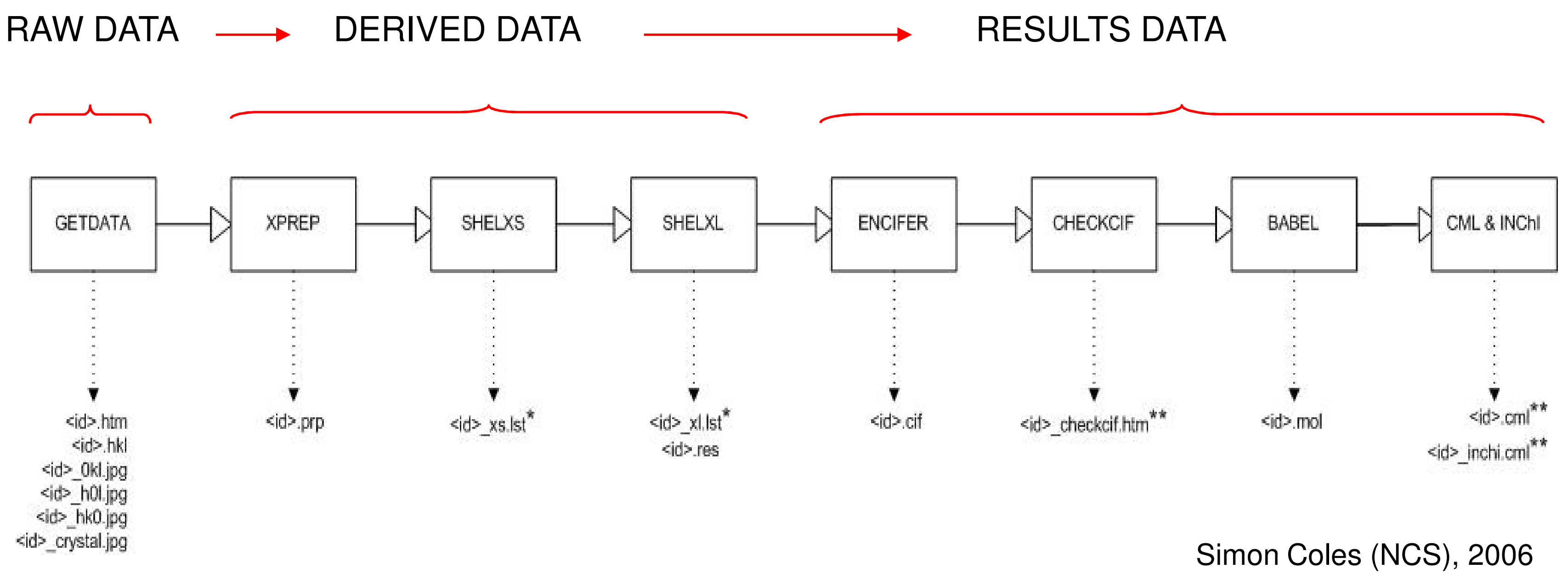
Community & Current Practice (2)

- Publishing datasets
  - Alongside journal articles through publisher mandates
  - Researchers often wish to retain exclusive use of their data
  - Lack of career rewards with respect to data creation and publishing
- Smaller projects at greatest risk
  - Sometimes CIF retained but raw data discarded
  - Data often stored on DVDs or laptops
  - Distributed, local storage -shortage of local curation expertise
  - Quality of metadata for datasets is variable
- Open access
  - eCrystals Federation Project
  - CrystalEye
  - ReciprocalNet (US, Australia, UK)
  - Crystallography Open Database (COD)
  - Chemistry Central (open access publisher)

Building the eCrystals Repository

- Phenomenal growth in amount of data generated from experiments
  - 40 years ago a PhD student would determine 2-3 structures for a thesis; this can now be easily achieved in a single day
- Only a small proportion is widely and easily accessible
  - Estimated that < 50% of crystal structures are published [Allen 2004]
  - Current data publication process is a bottleneck
- eBank-UK Project
  - JISC funded; three phases Sept. 2003-June 2007
  - UKOLN (lead), University of Southampton, University of Manchester
- eCrystals data repository
  - Open access and rapid dissemination of derived and results data from crystallography experiments
  - Repository platform: ePrints.org software V3
  - Supported by learned society (IUCr) and subject repository (CCDC)
- Linking research data to publications and scholarly communication
- Metadata harvesting and aggregation (OAI-PMH)

EPSRC NCS Crystal Structure Determination Workflow

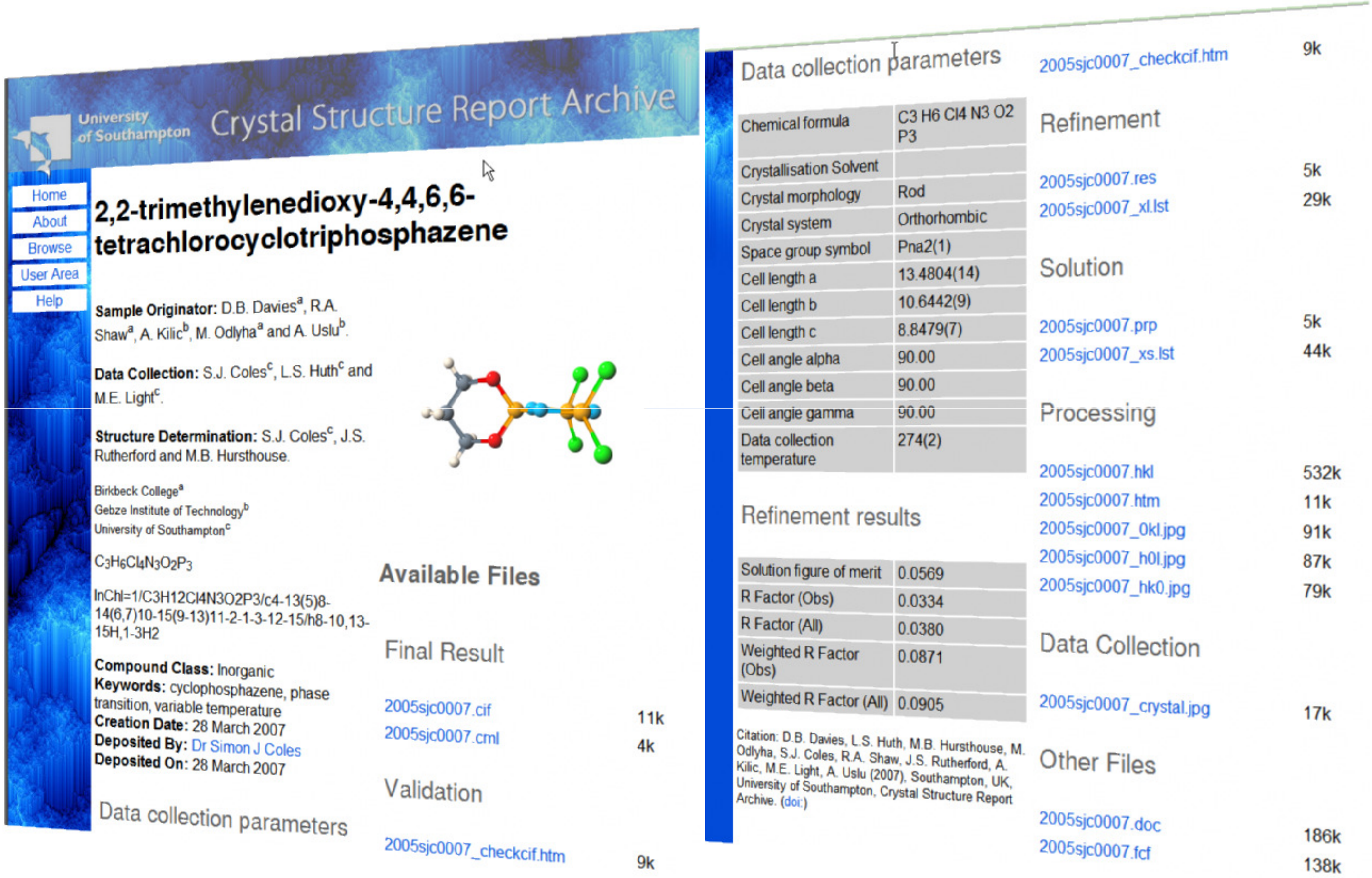


Simon Coles (NCS), 2006

- Initialisation: mount new sample
- Collection: collect data
- Processing: process and correct images
- Solution: solve structures

- Refinement: refine structure
- CIF: produce Crystallographic Information File
- Validation: chemical & crystallographic checks
- Report: generate Crystal Structure Report
- CML, INChI

eCrystals Data Repository: Example Crystal Structure Report



2,2-trimethylenedioxy-4,4,6,6-tetrachlorocyclotriphosphazene

Sample Originator: D.B. Davies[a], R.A. Shaw[a], A. Kilic[b], M. Odlyha[a] and A. Uslu[b].

Data Collection: S.J. Coles[c], L.S. Huth[c] and M.E. Light[c].

Structure Determination: S.J. Coles[c], J.S. Rutherford and M.B. Hursthouse.

Birkbeck College[a]
Gebze Institute of Technology[b]
University of Southampton[c]

$C_3H_6Cl_4N_3O_2P_3$

InChI=1/C3H12Cl4N3O2P3/c4-13(5)8-14(6,7)10-15(9-13)11-2-1-3-12-15/h8-10,13-15H,1-3H2

Compound Class: Inorganic
Keywords: cyclophosphazene, phase transition, variable temperature
Creation Date: 28 March 2007
Deposited By: Dr Simon J Coles
Deposited On: 28 March 2007

Available Files

Final Result

2005sjc0007.cif	11k
2005sjc0007.cml	4k

Validation

2005sjc0007_checkcif.htm	9k

Data collection parameters

Chemical formula	C3 H6 Cl4 N3 O2 P3
Crystallisation Solvent	
Crystal morphology	Rod
Crystal system	Orthorhombic
Space group symbol	Pna2(1)
Cell length a	13.4804(14)
Cell length b	10.6442(9)
Cell length c	8.8479(7)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	274(2)

Refinement results

Solution figure of merit	0.0569
R Factor (Obs)	0.0334
R Factor (All)	0.0380
Weighted R Factor (Obs)	0.0871
Weighted R Factor (All)	0.0905

Citation: D.B. Davies, L.S. Huth, M.B. Hursthouse, M. Odlyha, S.J. Coles, R.A. Shaw, J.S. Rutherford, A. Kilic, M.E. Light, A. Uslu (2007). Southampton, UK, University of Southampton, Crystal Structure Report Archive. (doi:)

2005sjc0007_checkcif.htm 9k

Refinement

2005sjc0007.res	5k
2005sjc0007_xl.lst	29k

Solution

2005sjc0007.prp	5k
2005sjc0007_xs.lst	44k

Processing

2005sjc0007.hkl	532k
2005sjc0007.htm	11k
2005sjc0007_0kl.jpg	91k
2005sjc0007_h0l.jpg	87k
2005sjc0007_hk0.jpg	79k

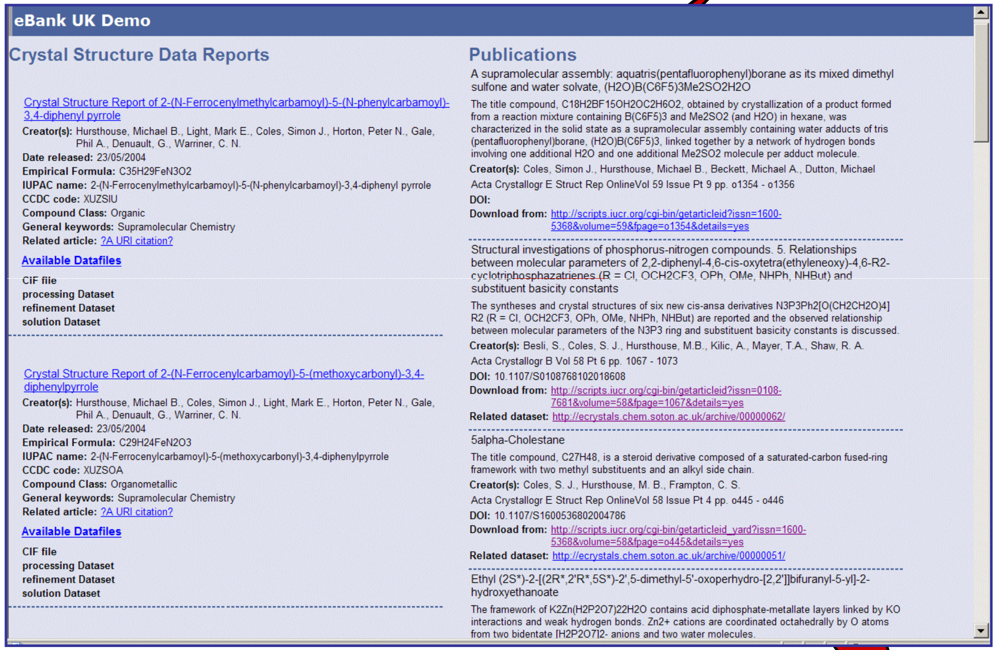
Data Collection

2005sjc0007_crystal.jpg	17k

Other Files

2005sjc0007.doc	186k
2005sjc0007.fcf	138k

Linking Data to Publications



Crystal Structure Data Reports

Crystal Structure Report of 2-(N-Ferrocenylmethylcarbamoyl)-5-(N-phenylcarbamoyl)-3,4-diphenyl pyrrole

Creator(s): Hursthouse, Michael B., Light, Mark E., Coles, Simon J., Horton, Peter N., Gale, Phil A., Denuault, G., Warriner, C. N.

Date released: 23/05/2004

Empirical Formula: C35H29FeN3O2

IUPAC name: 2-(N-Ferrocenylmethylcarbamoyl)-5-(N-phenylcarbamoyl)-3,4-diphenyl pyrrole

CCDC code: XUZSIU

Compound Class: Organic

General keywords: Supramolecular Chemistry

Related article: ?A URI citation?

Available Datafiles

CIF file
processing Dataset
refinement Dataset
solution Dataset

Crystal Structure Report of 2-(N-Ferrocenylcarbamoyl)-5-(methoxycarbonyl)-3,4-diphenylpyrrole

Creator(s): Hursthouse, Michael B., Coles, Simon J., Light, Mark E., Horton, Peter N., Gale, Phil A., Denuault, G., Warriner, C. N.

Date released: 23/05/2004

Empirical Formula: C29H24FeN2O3

IUPAC name: 2-(N-Ferrocenylcarbamoyl)-5-(methoxycarbonyl)-3,4-diphenylpyrrole

CCDC code: XUZSOA

Compound Class: Organometallic

General keywords: Supramolecular Chemistry

Related article: ?A URI citation?

Available Datafiles

CIF file
processing Dataset
refinement Dataset
solution Dataset

Publications

A supramolecular assembly: aquatris(pentafluorophenyl)borane as its mixed dimethyl sulfone and water solvate, (H2O)B(C6F5)3Me2SO2H2O

The title compound, C18H2BF15OH2OC2H6O2, obtained by crystallization of a product formed from a reaction mixture containing B(C6F5)3 and Me2SO2 (and H2O) in hexane, was characterized in the solid state as a supramolecular assembly containing water adducts of tris (pentafluorophenyl)borane, (H2O)B(C6F5)3, linked together by a network of hydrogen bonds involving one additional H2O and one additional Me2SO2 molecule per adduct molecule.

Creator(s): Coles, Simon J., Hursthouse, Michael B., Beckett, Michael A., Dutton, Michael

Acta Crystallogr E Struct Rep OnlineVol 59 Issue Pt 9 pp. o1354 - o1356

DOI:

Download from: http://scripts.iucr.org/cgi-bin/getarticleid?issn=1600-5368&volume=59&fpage=o1354&details=yes

Structural investigations of phosphorus-nitrogen compounds. 5. Relationships between molecular parameters of 2,2-diphenyl-4,6-cis-oxytetra(ethyleneoxy)-4,6-R2-cyclotriphosphazatrienes (R = Cl, OCH2CF3, OPh, OMe, NHPh, NHBut) and substituent basicity constants

The syntheses and crystal structures of six new cis-ansa derivatives N3P3Ph2[O(CH2CH2O)4] R2 (R = Cl, OCH2CF3, OPh, OMe, NHPh, NHBut) are reported and the observed relationship between molecular parameters of the N3P3 ring and substituent basicity constants is discussed.

Creator(s): Besli, S., Coles, S. J., Hursthouse, M.B., Kilic, A., Mayer, T.A., Shaw, R. A.

Acta Crystallogr B Vol 58 Pt 6 pp. 1067 - 1073

DOI: 10.1107/S0108768102018608

Download from: http://scripts.iucr.org/cgi-bin/getarticleid?issn=0108-7681&volume=58&fpage=1067&details=yes

Related dataset: http://ecrystals.chem.soton.ac.uk/archive/00000062/

5alpha-Cholestane

The title compound, C27H48, is a steroid derivative composed of a saturated-carbon fused-ring framework with two methyl substituents and an alkyl side chain.

Creator(s): Coles, S. J., Hursthouse, M. B., Frampton, C. S.

Acta Crystallogr E Struct Rep OnlineVol 58 Issue Pt 4 pp. o445 - o446

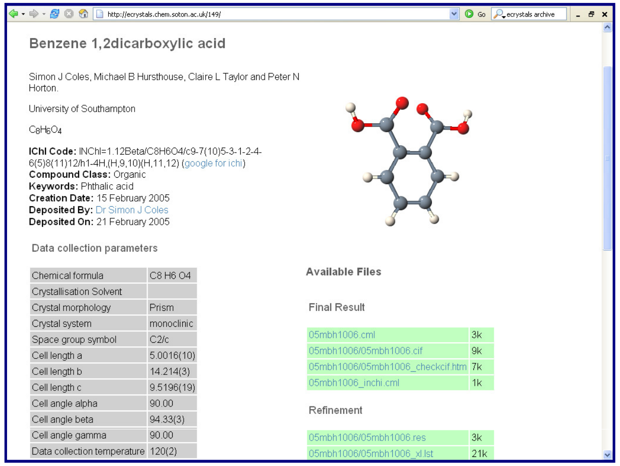
DOI: 10.1107/S1600536802004786

Download from: http://scripts.iucr.org/cgi-bin/getarticleid_yard?issn=1600-5368&volume=58&fpage=o445&details=yes

Related dataset: http://ecrystals.chem.soton.ac.uk/archive/00000051/

Ethyl (2S*)-2-[(2R*,2'R*,5S*)-2',5-dimethyl-5'-oxoperhydro-[2,2']]bifuranyl-5-yl]-2-hydroxyethanoate

The framework of K2Zn(H2P2O7)22H2O contains acid diphosphate-metallate layers linked by KO interactions and weak hydrogen bonds. Zn2+ cations are coordinated octahedrally by O atoms from two bidentate [H2P2O7]2- anions and two water molecules.



Benzene 1,2dicarboxylic acid

Simon J Coles, Michael B Hursthouse, Claire L Taylor and Peter N Horton.

University of Southampton

$C_8H_6O_4$

IChI Code: INChI=1.12Beta/C8H6O4/c9-7(10)5-3-1-2-4-6(5)8(11)12/h1-4H,(H,9,10)(H,11,12) (google for ichi)
Compound Class: Organic
Keywords: Phthalic acid
Creation Date: 15 February 2005
Deposited By: Dr Simon J Coles
Deposited On: 21 February 2005

Data collection parameters

Chemical formula	C8 H6 O4
Crystallisation Solvent	
Crystal morphology	Prism
Crystal system	monoclinic
Space group symbol	C2/c
Cell length a	5.0016(10)
Cell length b	14.214(3)
Cell length c	9.5196(19)
Cell angle alpha	90.00
Cell angle beta	94.33(3)
Cell angle gamma	90.00
Data collection temperature	120(2)

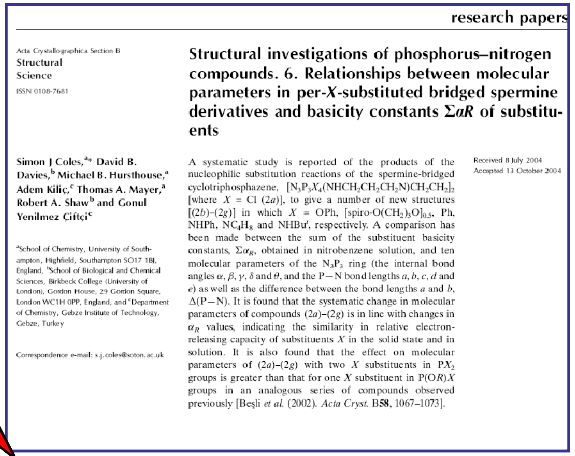
Available Files

Final Result

05mbh1006.cml	3k
05mbh1006/05mbh1006.cif	9k
05mbh1006/05mbh1006_checkcif.htm	7k
05mbh1006_inchi.cml	1k

Refinement

05mbh1006/05mbh1006.res	3k
05mbh1006/05mbh1006_xl.lst	21k



research papers

Acta Crystallographica Section B
Structural
Science
ISSN 0108-7681

Structural investigations of phosphorus–nitrogen compounds. 6. Relationships between molecular parameters in per-*X*-substituted bridged spermine derivatives and basicity constants Σ*αR* of substituents

Simon J Coles,[a]* David B. Davies,[b] Michael B. Hursthouse,[a] Adem Kiliç,[c] Thomas A. Mayer,[a] Robert A. Shaw[b] and Gonul Yenilmez Çiftçi[c]

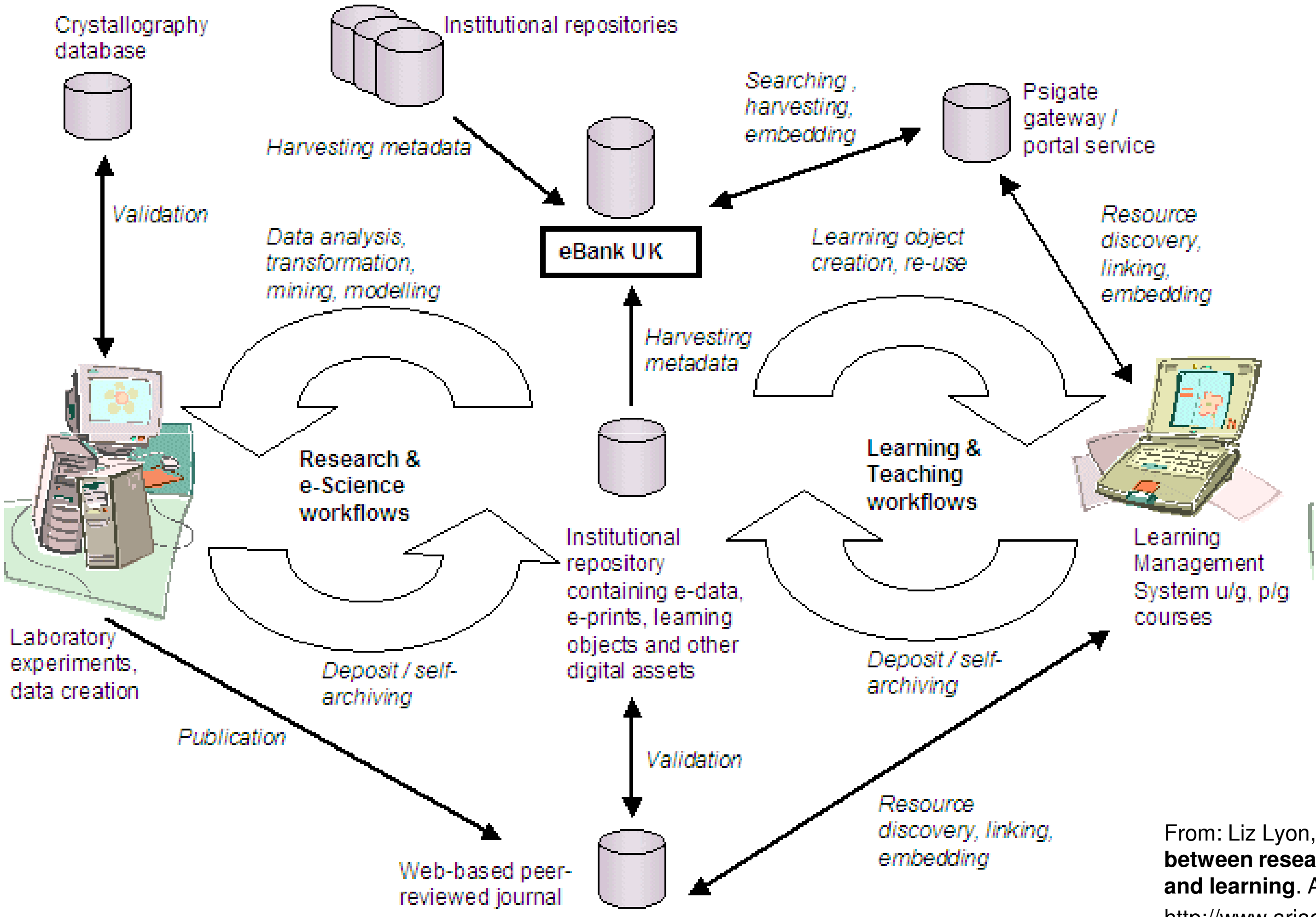
[a]School of Chemistry, University of Southampton, Highfield, Southampton SO17 1BJ, England, [b]School of Biological and Chemical Sciences, Birkbeck College (University of London), Gordon House, 29 Gordon Square, London WC1H 0PP, England, and [c]Department of Chemistry, Gebze Institute of Technology, Gebze, Turkey

Correspondence e-mail: s.j.coles@soton.ac.uk

Received 8 July 2004
Accepted 13 October 2004

A systematic study is reported of the products of the nucleophilic substitution reactions of the spermine-bridged cyclotriphosphazene, $[N_3P_3X_4(NHCH_2CH_2CH_2N)CH_2CH_2]_2$ [where X = Cl (2*a*)], to give a number of new structures [(2*b*)–(2*g*)] in which X = OPh, $[\text{spiro-O(CH}_2)_3\text{O}]_{0.5}$, Ph, NHPh, NC_4H_8 and $NHBu^t$, respectively. A comparison has been made between the sum of the substituent basicity constants, $\Sigma\alpha_R$, obtained in nitrobenzene solution, and ten molecular parameters of the N_3P_3 ring (the internal bond angles $\alpha, \beta, \gamma, \delta$ and θ, and the P—N bond lengths a, b, c, d and e) as well as the difference between the bond lengths a and b, Δ(P—N). It is found that the systematic change in molecular parameters of compounds (2*a*)–(2*g*) is in line with changes in α_R values, indicating the similarity in relative electron-releasing capacity of substituents X in the solid state and in solution. It is also found that the effect on molecular parameters of (2*a*)–(2*g*) with two X substituents in PX_2 groups is greater than that for one X substituent in P(O*R*)*X* groups in an analogous series of compounds observed previously [Beşli *et al.* (2002). *Acta Cryst.* B**58**, 1067–1073].

The Scholarly Knowledge Lifecycle


Crystallography database
Institutional repositories
Harvesting metadata
Searching , harvesting, embedding
Psigate gateway / portal service
Validation
Data analysis, transformation, mining, modelling
eBank UK
Learning object creation, re-use
Resource discovery, linking, embedding
Harvesting metadata
Research & e-Science workflows
Learning & Teaching workflows
Institutional repository containing e-data, e-prints, learning objects and other digital assets
Learning Management System u/g, p/g courses
Laboratory experiments, data creation
Deposit / self-archiving
Deposit / self-archiving
Publication
Validation
Resource discovery, linking, embedding
Web-based peer-reviewed journal

Both research and learning are cyclical processes

- Research outputs feed into and contribute to knowledge
- Research outputs are based on continuous use and reuse of data i.e. derivative in nature

From: Liz Lyon, **eBank UK: Building the links between research data, scholarly communication and learning**. ARIADNE, July 2003
http://www.ariadne.ac.uk/issue36/lyon/intro.html

Resource Discovery & Reuse

- Simple Dublin Core
  - Crystal structure
  - Title (Systematic IUPAC Name)
  - Authors
  - Affiliation
  - Creation Date
- Qualified Dublin Core (for additional chemical metadata)
  - Empirical formula
  - International Chemical Identifier (InChI)
  - Compound Class and Keywords
- Application Profile: http://www.ukoln.ac.uk/projects/ebank-uk/schemas/
- DOI links: http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145
- Rights & Citation: http://ecrystals.chem.soton.ac.uk/rights.html

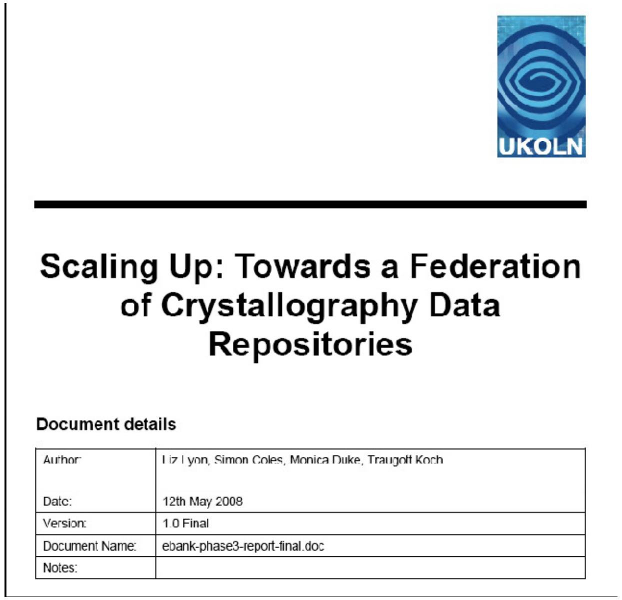
Scaling Up: Towards a Federation

Interviews, analysis & synthesis:

IR Policy & Practice, Laboratory Practice & Workflows, Technical Interoperability & Standards, Metadata Schema & Application Profiles, Semantic Interoperability, Data Citation, Identifiers & Linking, Federation Architectures & Third Party Services, Rights & Licensing, Data Quality & Validation, Preservation, Curation & Sustainability

Selected Issues (& Recommendations):

- Diverse laboratory practice
- Instrument manufacturers have proprietary formats
- Data policy needs to reflect laboratory practice
- Data quality criteria and validation (access to raw data)
- Repository must provide control over timing of public visibility-"prior publication" problem
- No disciplinary preservation model



UKOLN

Scaling Up: Towards a Federation of Crystallography Data Repositories

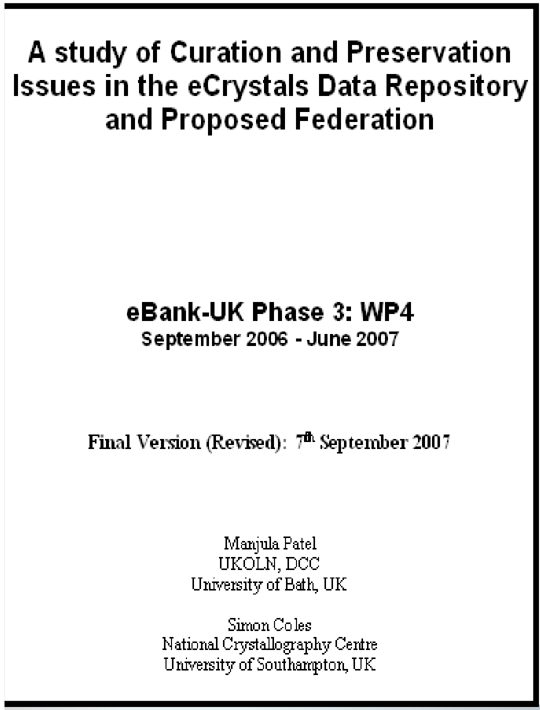
Document details

Author: Datc:	Liz Lyon, Simon Coles, Monica Duke, Traugott Koch 12th May 2008
Version:	1.0 Final
Document Name:	ebank-phase3-report-final.doc
Notes:	

Data Curation & Preservation

eBank-UK Phase 3: "A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed Federation", Sept. 2007

- Development of preservation strategies and policies
- Audit and certification issues (TRAC, DRAMBORA, NESTOR, ISO International repository audit and certification BOF Group)
- OAIS and Representation Information for crystallography data
- eBank-UK Application Profile and preservation metadata
- e-Prints.org repository platform



A study of Curation and Preservation Issues in the eCrystals Data Repository and Proposed Federation

eBank-UK Phase 3: WP4
September 2006 - June 2007

Final Version (Revised): 7th September 2007

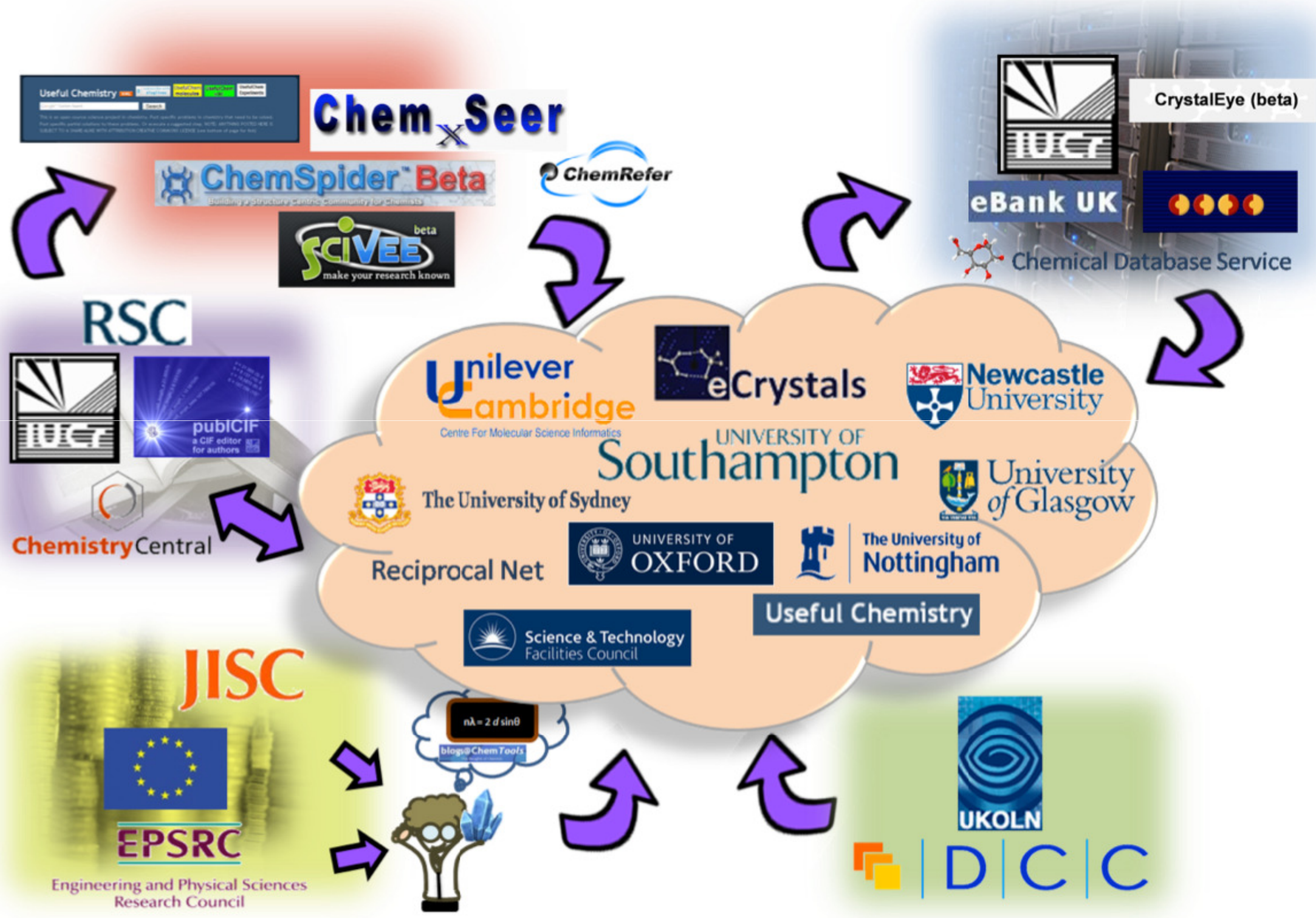
Manjula Patel
UKOLN, DCC
University of Bath, UK

Simon Coles
National Crystallography Centre
University of Southampton, UK

Data Curation & Preservation: Recommendations

- Develop a preservation and curation strategy and formal policies to indicate levels of service (e.g. deposit, ingest, validation, dissemination)
- Promote community-supported sustainability plan
- Self-assessment using DRAMBORA toolkit
  - Implement regular audits e.g. annually
  - Produce documentary evidence of compliance
- Maintenance and open access of critical file formats and software
  - Crystallography Information File (CIF)
  - Work-up software e.g. XPREP; SHELX{S,L}; ENCIFER; checkCIF, BABEL
  - Advocate export of raw data from instrumentation as IMG CIF
- Capture relevant Representation Information
- Capture preservation metadata (e.g. versioning; provenance)
  - OAIS Preservation Description Information
  - PREMIS Data Dictionary
  - Extend or augment eBank Metadata Application Profile
- Obtain consensus on Metadata Application Profile
- Seek to automate metadata generation, extraction and maintenance

Building a Federation of Repositories


Useful Chemistry
ChemxSeer
ChemSpider Beta
ChemRefer
SciVee make your research known
CrystalEye (beta)
IUCr
eBank UK
Chemical Database Service
RSC
IUCr
pubICIF a CIF editor for authors
ChemistryCentral
Unilever Cambridge Centre For Molecular Science Informatics
eCrystals
Newcastle University
University of Southampton
University of Glasgow
The University of Sydney
Reciprocal Net
University of Oxford
The University of Nottingham
Useful Chemistry
Science & Technology Facilities Council
JISC
EPSRC Engineering and Physical Sciences Research Council
nλ = 2 d sinθ
blogs@ChemTools
UKOLN
DCC

eCrystals Federation Project

- eCrystals Federation Project, Nov 2007 – Mar 2009
- Builds on eBank-UK Phase 3 results
- Led by the UK National Crystallography Service (University of Southampton) with core partners at UKOLN (University of Bath), the Digital Curation Centre and the Unilever Centre (University of Cambridge) – currently 14 supporting partners.
- Integrate and embed open data repository approach into current research practice by engaging data centres, librarians, researchers, publishers and third party information providers
- Harmonise Federation metadata application profile
- Investigate aggregation issues arising from harvesting metadata from Federation repositories
- Enable the Federation of institutional repositories to interoperate with international subject archives (IUCr and CCDC) and other third party harvesters
- Develop approaches to preservation and curation of scientific data in open repositories

Federation Interoperability

- Roll-out in 2 phases led by University of Southampton
  - Universities Sydney, Drexel, Birmingham, Newcastle with eprints.org platform
  - University Cambridge, STFC, ReciprocalNet, ARCHER with other platforms
  - Establish Federation policies, metadata application profile etc.
- Bi-directional links with derived articles in “publisher repositories”, IUCr, RSC, Chemistry Central
- StORe middleware -linking “source” and “output” repositories
- CLADDIER –linking data to publications
- OAI-ORE (Open Archives Initiative – Object Reuse and Exchange)
  - Enable distributed repositories to fully describe and exchange content
  - MicroSoft eChemistry Project

Some challenges

- Data management plans
- Dealing with diverse laboratory practice and workflows
- Appraisal and selection
- Data provenance, audit, tracking
- Citations and versions –persistent identifiers
- Granularity of citations: dataset or values within a dataset
- Instrumentation –proprietary formats
- Access to raw data files for mining and quality control purposes
- Preservation beyond “data” e.g. workflows, blogs, discourse
- Linking across disciplines and sectors
- Collaborative social networks; also “citizen science”
- Semantic integration –controlled vocabularies, ontology etc.

Selected References

- Liz Lyon, **eBank UK: Building the links between research data, scholarly communication and learning**, ARIADNE, July 2003
- F. H. Allen, **High-throughput crystallography: the challenge of publishing, storing and using the results.** Crystallography Reviews, 10, pp3-15, 2004
- Monica Duke, Michael Day, Rachel Heery, Leslie A. Carr, Simon J. Coles
  Enhancing access to research data: the challenge of crystallography
  JCDL 2005 Digital Libraries: Cyberinfrastructure for Research and Education, Denver, Colorado, USA June 7-11, 2005
- Scott Carlson, **Lost in a Sea of Science Data,** The Chronicle of Higher Education, June 2006
- Simon Coles, Jeremy Frey and Andrew Milstead
  Curation of Chemistry from Laboratory to Publication
  UK e-Science All Hands Meeting 2006, Nottingham, UK, 18-21 September 2006
- Liz Lyon, **Dealing with Data: Roles, Rights, Responsibilities and Relationships,** JISC Consultancy Report, June 2007
- Manjula Patel and Simon Coles
  A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed Federation, September 2007
- Liz Lyon, Simon Coles, Monica Duke, Traugott Koch
  Scaling Up: Towards a Federation of Crystallography Data Repositories, May 2008
- **To Share or not to Share, Publication and Quality Assurance of Research Data Outputs,** A Report commissioned by the Research Information Network (RIN), Annex: detailed findings for the eight research areas, June 2008

Thanks …

…for your attention

…to

Simon Coles: EPSRC National Crystallography Service,
School of Chemistry, University of Southampton
Liz Lyon: UKOLN, University of Bath

Questions?

Manjula Patel
UKOLN
University of Bath, UK
http://www.ukoln.ac.uk/