Electronics and Computer Science

eprints.org

Institutional Data Repositories for Chemistry

Simon Coles

School of Chemistry,

University of Southampton, U.K.

s.j.coles@soton.ac.uk

eBank UK

R4L Repository for the Laboratory

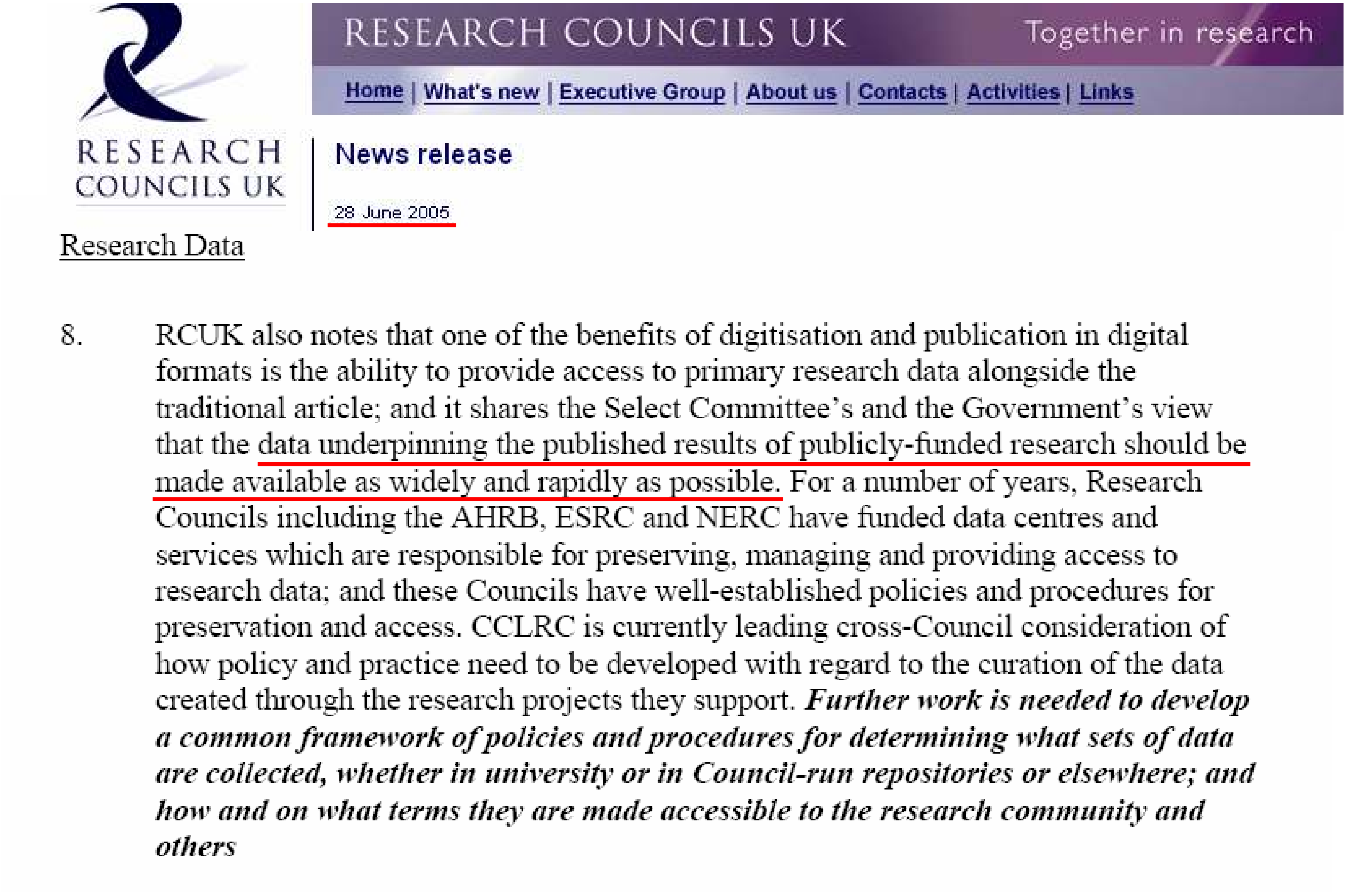
UKOLN

intute

ALPSP

JISC

Why?
Funding Body Viewpoint


RESEARCH COUNCILS UK
Together in research
Home | What's new | Executive Group | About us | Contacts | Activities | Links
RESEARCH COUNCILS UK

News release

28 June 2005

Research Data

8. RCUK also notes that one of the benefits of digitisation and publication in digital formats is the ability to provide access to primary research data alongside the traditional article; and it shares the Select Committee's and the Government's view that the data underpinning the published results of publicly-funded research should be made available as widely and rapidly as possible. For a number of years, Research Councils including the AHRB, ESRC and NERC have funded data centres and services which are responsible for preserving, managing and providing access to research data; and these Councils have well-established policies and procedures for preservation and access. CCLRC is currently leading cross-Council consideration of how policy and practice need to be developed with regard to the curation of the data created through the research projects they support. ***Further work is needed to develop a common framework of policies and procedures for determining what sets of data are collected, whether in university or in Council-run repositories or elsewhere; and how and on what terms they are made accessible to the research community and others***

Why?
Curation in the Laboratory

“Data from experiments conducted as recently as six months ago might be suddenly deemed important, but those researchers may never find those numbers – or if they did might not know what those numbers meant”

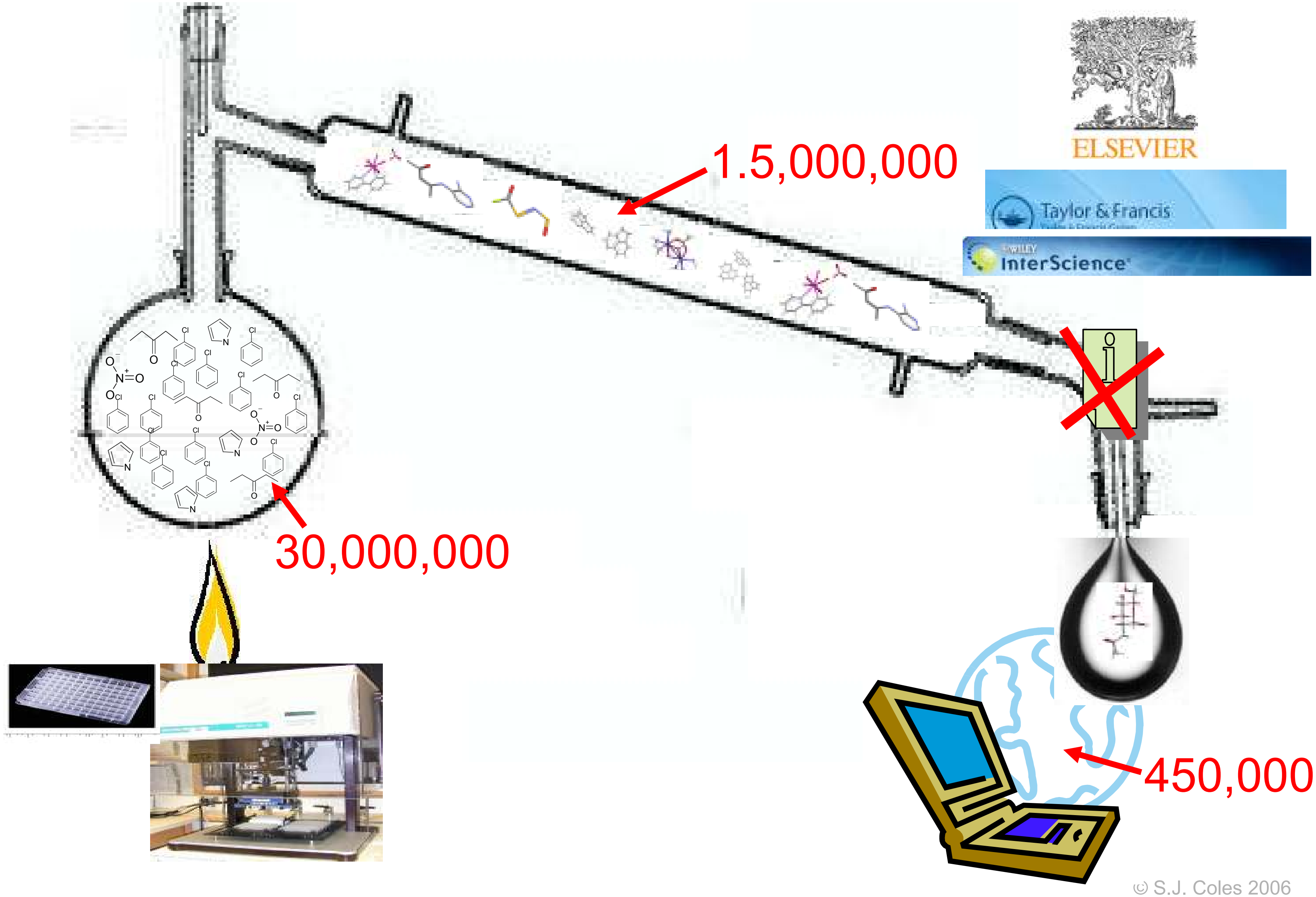
“Lost in some research assistant’s computer, the data are often irretrievable or an undecipherable string of digits”

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

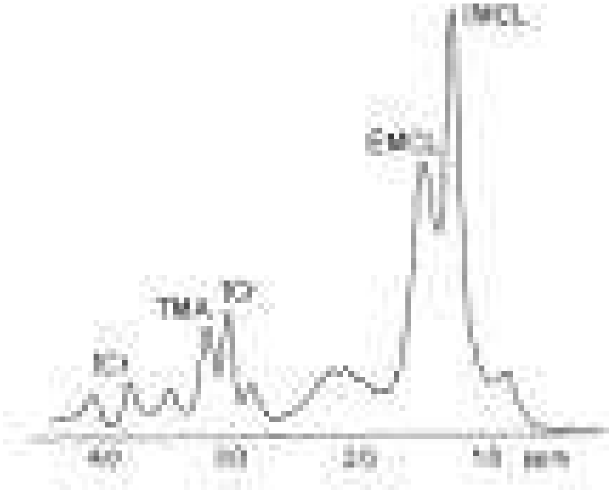
“Data from Big Science is … easier to handle, understand and archive. Small Science is horribly heterogeneous and far more vast. In time Small Science will generate 2-3 times more data than Big Science.”

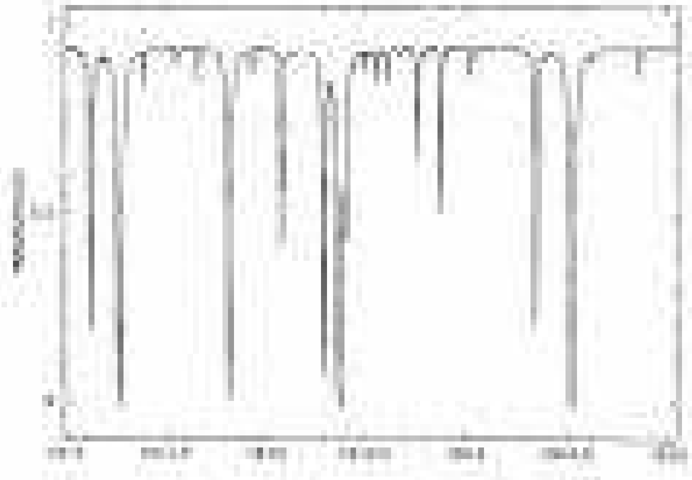
‘Lost in a Sea of Science Data’ S.Carlson, The Chronicle of Higher Education (23/06/2006)

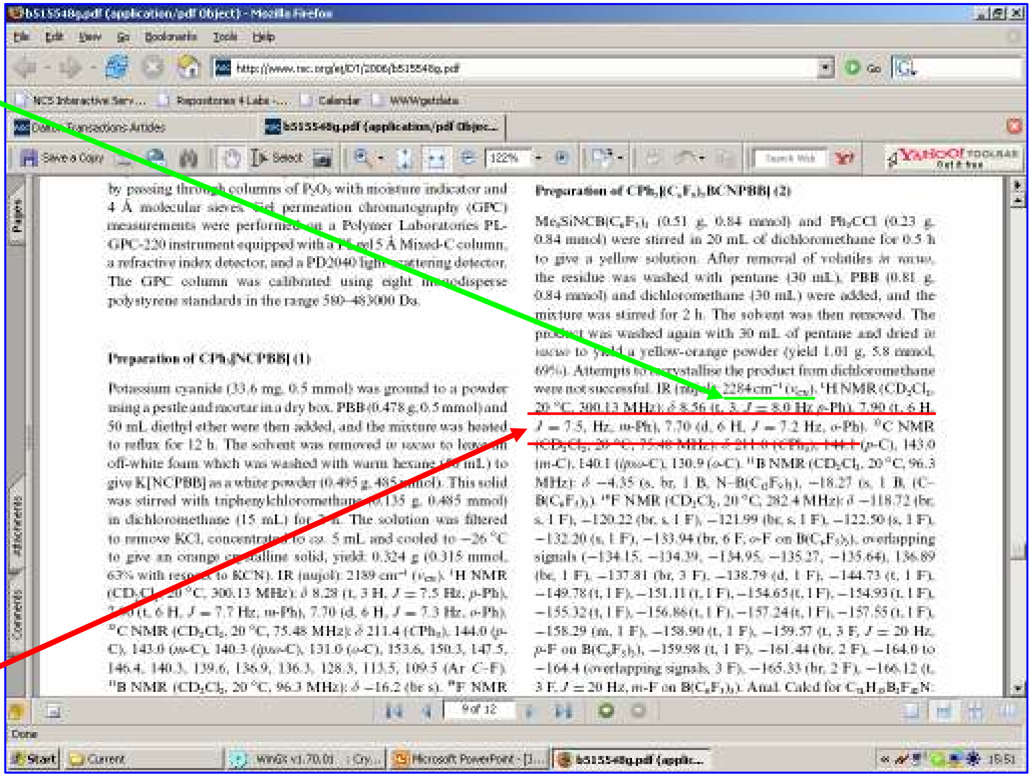
Why?
Publishing and the Data Deluge



Why?
Publishing Data and Information Loss






b515548g.pdf (application/pdf Object) - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://www.rsc.org/ej/DT/2006/b515548g.pdf
Go
NCS Interactive Serv...
Repositories 4 Labs -...
Calendar
WWWgetdata
Dalton Transactions Articles
b515548g.pdf (application/pdf Objec...
Save a Copy
Select
122%
Search Web
YAHOO! TOOLBAR
Pages
Attachments
Comments
9 of 12
Done
Start
Current
WinGX v1.70.01 : Cry...
Microsoft PowerPoint - [I...
b515548g.pdf (applic...
15:51

by passing through columns of P_2O_5 with moisture indicator and 4 Å molecular sieves. Gel permeation chromatography (GPC) measurements were performed on a Polymer Laboratories PL-GPC-220 instrument equipped with a PLgel 5 Å Mixed-C column, a refractive index detector, and a PD2040 light scattering detector. The GPC column was calibrated using eight monodisperse polystyrene standards in the range 580–483000 Da.

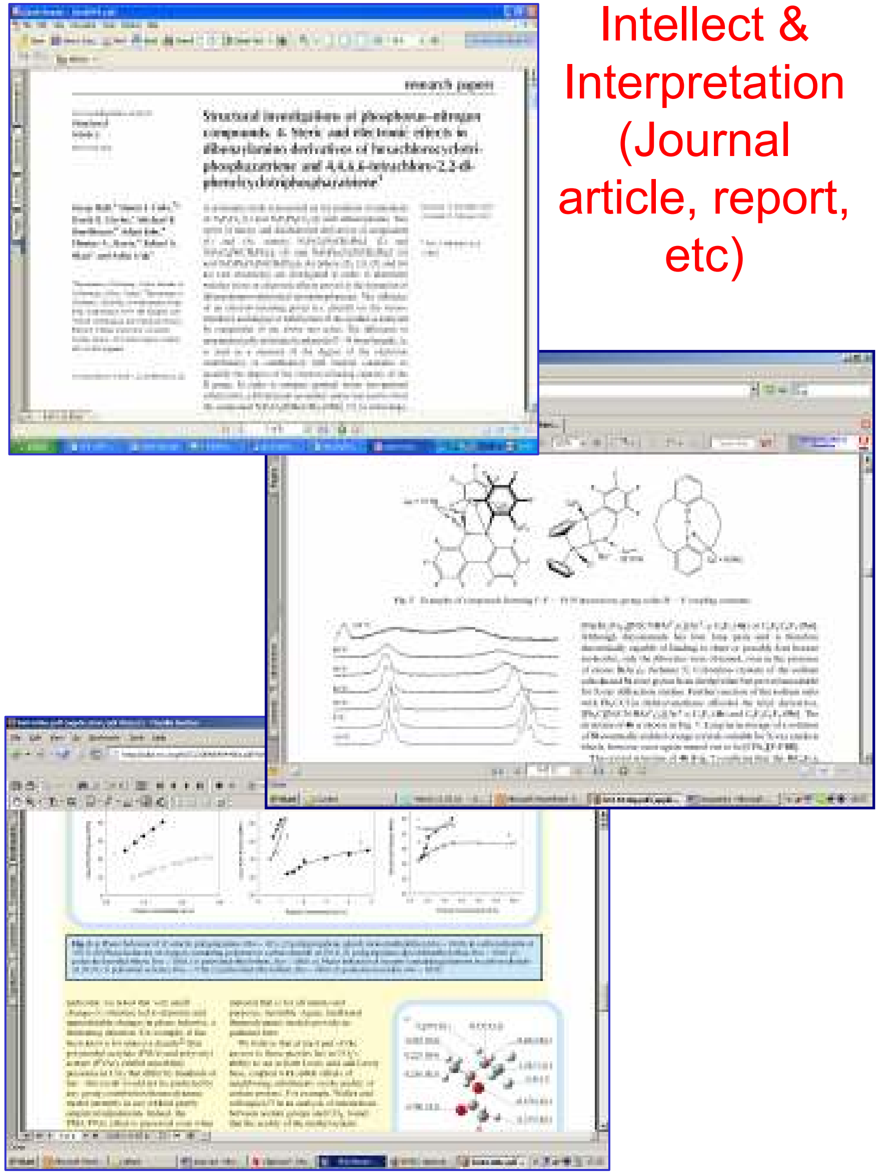
Preparation of $CPh_3[NCPBB]$ (1)

Potassium cyanide (33.6 mg, 0.5 mmol) was ground to a powder using a pestle and mortar in a dry box. PBB (0.478 g, 0.5 mmol) and 50 mL diethyl ether were then added, and the mixture was heated to reflux for 12 h. The solvent was removed *in vacuo* to leave an off-white foam which was washed with warm hexane (30 mL) to give K[NCPBB] as a white powder (0.495 g, 485 mmol). This solid was stirred with triphenylchloromethane (0.135 g, 0.485 mmol) in dichloromethane (15 mL) for 2 h. The solution was filtered to remove KCl, concentrated to *ca.* 5 mL and cooled to −26 °C to give an orange crystalline solid, yield: 0.324 g (0.315 mmol, 63% with respect to KCN). IR (nujol): 2189 cm^{-1} (ν_{CN}). ^{1}H NMR (CD_2Cl_2, 20 °C, 300.13 MHz): δ 8.28 (t, 3 H, *J* = 7.5 Hz, *p*-Ph), 7.90 (t, 6 H, *J* = 7.7 Hz, *m*-Ph), 7.70 (d, 6 H, *J* = 7.3 Hz, *o*-Ph). ^{13}C NMR (CD_2Cl_2, 20 °C, 75.48 MHz): δ 211.4 (CPh_3), 144.0 (*p*-C), 143.0 (*m*-C), 140.3 (*ipso*-C), 131.0 (*o*-C), 153.6, 150.3, 147.5, 146.4, 140.3, 139.6, 136.9, 136.3, 128.3, 113.5, 109.5 (Ar *C*–F). ^{11}B NMR (CD_2Cl_2, 20 °C, 96.3 MHz): δ −16.2 (br s). ^{19}F NMR

Preparation of $CPh_3[(C_6F_5)_3BCNPBB]$ (2)

$Me_3SiNCB(C_6F_5)_3$ (0.51 g, 0.84 mmol) and Ph_3CCl (0.23 g, 0.84 mmol) were stirred in 20 mL of dichloromethane for 0.5 h to give a yellow solution. After removal of volatiles *in vacuo*, the residue was washed with pentane (30 mL). PBB (0.81 g, 0.84 mmol) and dichloromethane (30 mL) were added, and the mixture was stirred for 2 h. The solvent was then removed. The product was washed again with 30 mL of pentane and dried *in vacuo* to yield a yellow-orange powder (yield 1.01 g, 5.8 mmol, 69%). Attempts to recrystallise the product from dichloromethane were not successful. IR (nujol): 2284 cm^{-1} (ν_{CN}). ^{1}H NMR (CD_2Cl_2, 20 °C, 300.13 MHz): δ 8.56 (t, 3, *J* = 8.0 Hz *p*-Ph), 7.90 (t, 6 H, *J* = 7.5, Hz, *m*-Ph), 7.70 (d, 6 H, *J* = 7.2 Hz, *o*-Ph). ^{13}C NMR (CD_2Cl_2, 20 °C, 75.48 MHz): δ 211.0 (CPh_3), 144.1 (*p*-C), 143.0 (*m*-C), 140.1 (*ipso*-C), 130.9 (*o*-C). ^{11}B NMR (CD_2Cl_2, 20 °C, 96.3 MHz): δ −4.35 (s, br, 1 B, N–$B(C_6F_5)_3$), −18.27 (s, 1 B, (C–$B(C_6F_5)_3$)). ^{19}F NMR (CD_2Cl_2, 20 °C, 282.4 MHz): δ −118.72 (br, s, 1 F), −120.22 (br, s, 1 F), −121.99 (br, s, 1 F), −122.50 (s, 1 F), −132.20 (s, 1 F), −133.94 (br, 6 F, *o*-F on $B(C_6F_5)_3$), overlapping signals (−134.15, −134.39, −134.95, −135.27, −135.64), 136.89 (br, 1 F), −137.81 (br, 3 F), −138.79 (d, 1 F), −144.73 (t, 1 F), −149.78 (t, 1 F), −151.11 (t, 1 F), −154.65 (t, 1 F), −154.93 (t, 1 F), −155.32 (t, 1 F), −156.86 (t, 1 F), −157.24 (t, 1 F), −157.55 (t, 1 F), −158.29 (m, 1 F), −158.90 (t, 1 F), −159.57 (t, 3 F, *J* = 20 Hz, *p*-F on $B(C_6F_5)_3$), −159.98 (t, 1 F), −161.44 (br, 2 F), −164.0 to −164.4 (overlapping signals, 3 F), −165.33 (br, 2 F), −166.12 (t, 3 F, *J* = 20 Hz, *m*-F on $B(C_6F_5)_3$). Anal. Calcd for $C_{74}H_{15}B_2F_{42}N$:

Separating Data from Interpretations

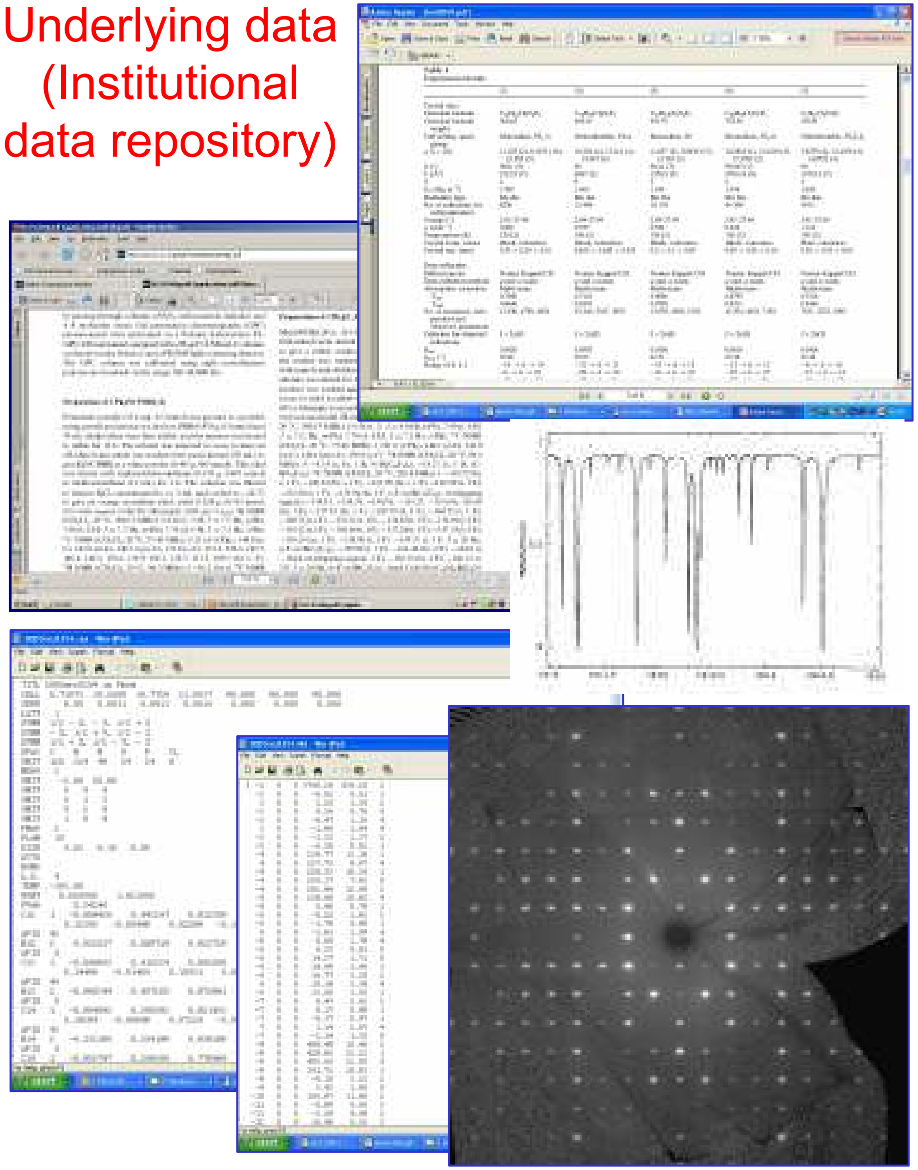


research papers

Structural investigations of phosphorus–nitrogen compounds. 4. Steric and electronic effects in dibenzylamino derivatives of hexachlorocyclotriphosphazatriene and 4,4,6,6-tetrachloro-2,2-diphenylcyclotriphosphazatriene

Intellect & Interpretation (Journal article, report, etc)

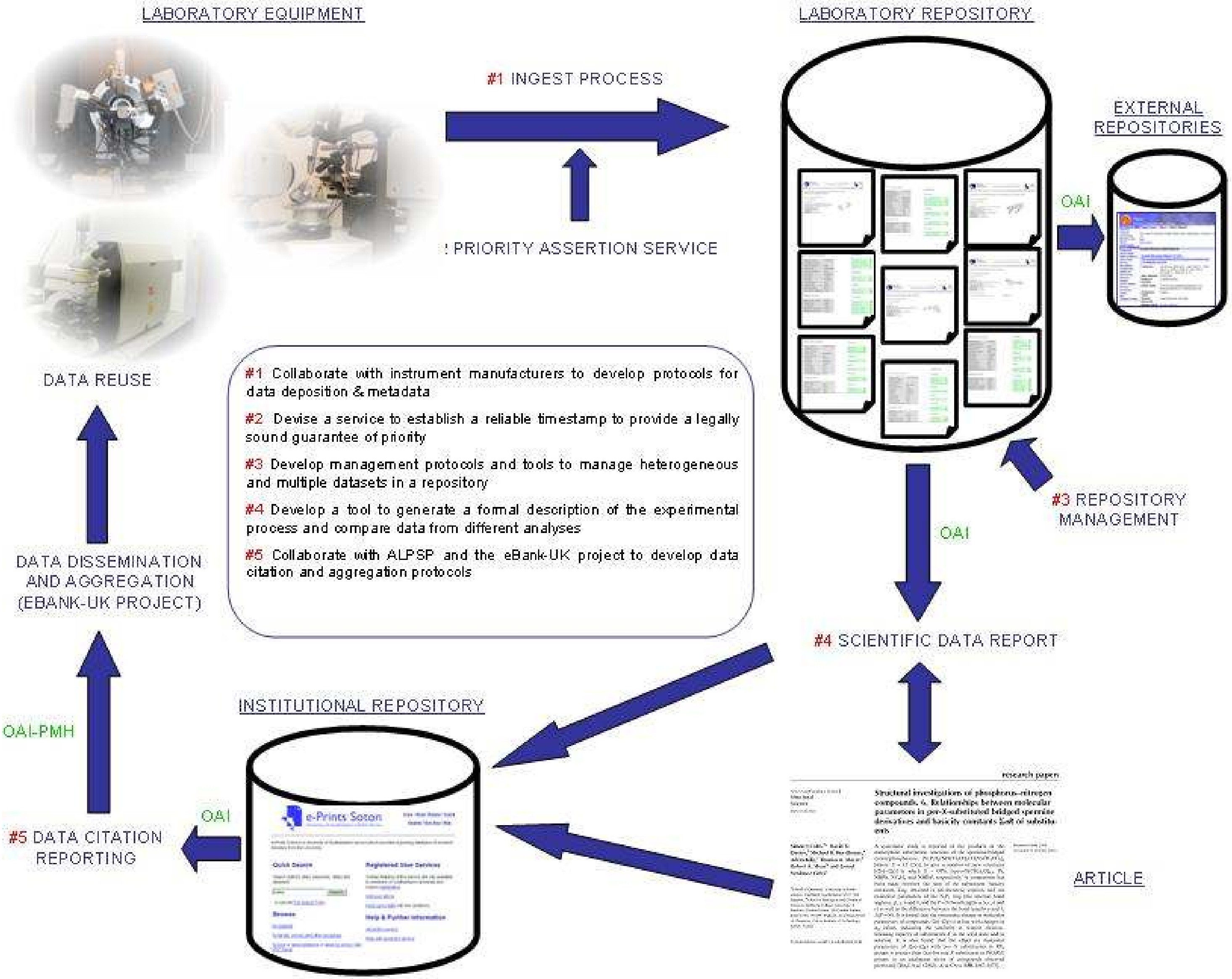
Underlying data (Institutional data repository)



Data capture and curation at the point of generation in the laboratory

The Repository for the Laboratory – R4L

Laboratory IRs and Information Management



#1 Collaborate with instrument manufacturers to develop protocols for data deposition & metadata

#2 Devise a service to establish a reliable timestamp to provide a legally sound guarantee of priority

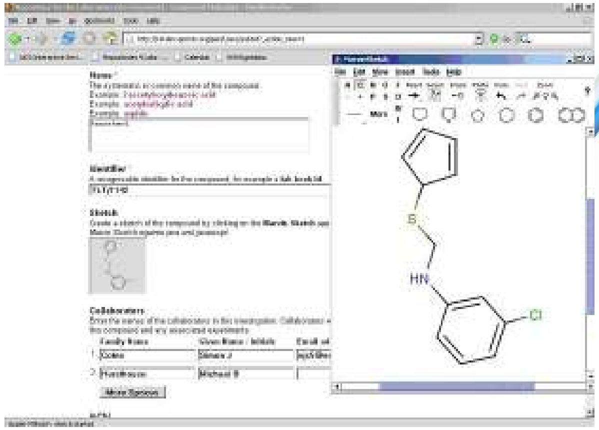
#3 Develop management protocols and tools to manage heterogeneous and multiple datasets in a repository

#4 Develop a tool to generate a formal description of the experimental process and compare data from different analyses

#5 Collaborate with ALPSP and the eBank-UK project to develop data citation and aggregation protocols

The R4L Repository

Create new compound



Name*
The systematic or common name of the compound.
Example: 2-acetyloxybenzoic acid
Example: acetylsalicylic acid
Example: aspirin

Identifier
A recognisable identifier for the compound, for example a lab. book id
TLT/1/40

Sketch
Create a sketch of the compound by clicking on the Marvin Sketch app
Marvin Sketch requires java and javascript

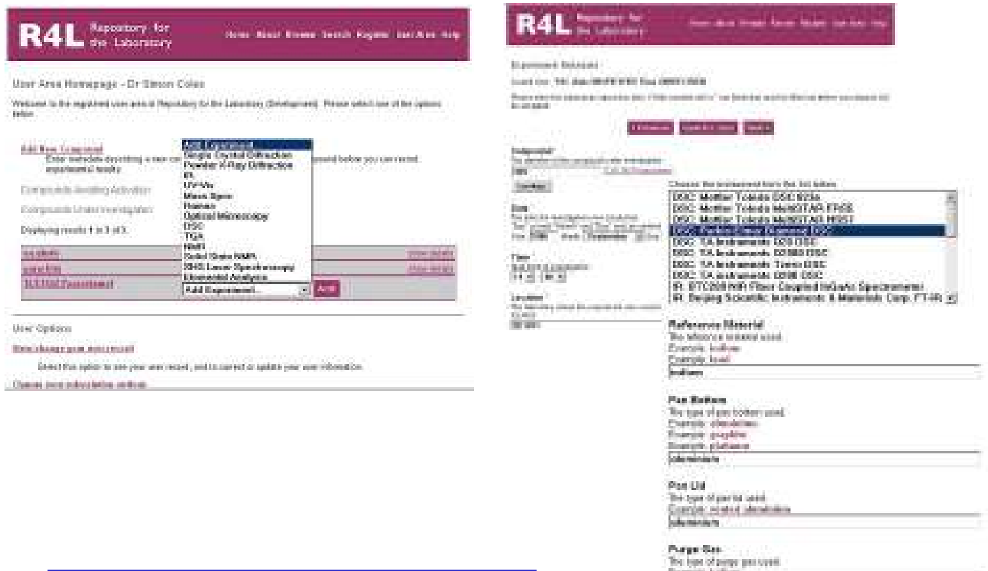
Collaborators
Enter the names of the collaborators in this investigation. Collaborators
the compound and any associated experiments

	Family Name	Given Name / Initials	Email ad
1	Coles	Simon J	[illegible]
2	Hursthouse	Michael B	

More Options

Deposit

Add experiment data and metadata



R4L Repository for the Laboratory
Home About Browse Search Register User Area Help

User Area Homepage - Dr Simon Coles

Welcome to the registered user area of Repository for the Laboratory (Development). Please select one of the options below.

Add New Compound
Enter metadata describing a new co[illegible] [illegible]ound below you can record experimental results

Compounds Awaiting Activation

Compounds Under Investigation

Displaying results 1 to 3 of 3.

TLT/1/40 Furosemide

Add Experiment...
Single Crystal Diffraction
Powder X-Ray Diffraction
IR
UV-Vis
Mass Spec
Raman
Optical Microscopy
DSC
TGA
NMR
Solid State NMR
SHG Laser Spectroscopy
Elemental Analysis

Add Experiment... Add

User Options
View, change your user record
Select this option to see your user record, and to correct or update your user information.
Change your subscription settings

R4L Repository for the Laboratory

Experiment Metadata

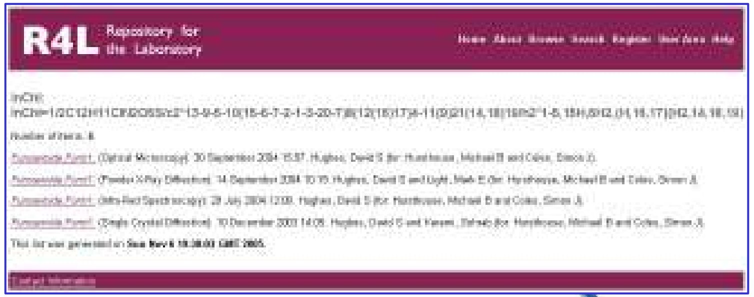
Choose the instrument from the list below
DSC: Mettler Toledo DSC 823e
DSC: Mettler Toledo MultiSTAR FRS5
DSC: Mettler Toledo MultiSTAR HRS7
DSC: Perkin Elmer Diamond DSC
DSC: TA Instruments Q20 DSC
DSC: TA Instruments Q2000 DSC
DSC: TA Instruments Tzero DSC
DSC: TA Instruments Q200 DSC
IR: BTC208 MIR Fiber Coupled InGaAs Spectrometer
IR: Beijing Scientific Instruments & Materials Corp. FT-IR

Reference Material
The reference material used.
Example: indium
Example: lead
Indium

Pan Bottom
The type of pan bottom used.
Example: aluminium
Example: graphite
Example: platinum
aluminium

Pan Lid
The type of pan lid used.
Example: vented aluminium
aluminium

Purge Gas
The type of purge gas used.
Example: helium
helium



R4L Repository for the Laboratory
Home About Browse Search Register User Area Help

InChI:
InChI=1/C12H11ClN2O5S/c13-9-5-10(15-6-7-2-1-3-20-7)8(12(16)17)4-11(9)21(14,18)19/h1-5,15H,6H2,(H,16,17)(H2,14,18,19)

Number of items: 4

Furosemide Form1 (Optical Microscopy) 30 September 2004 15:37. Hughes, David S (for: Hursthouse, Michael B and Coles, Simon J).

Furosemide Form1 (Powder X-Ray Diffraction) 14 September 2004 10:15. Hughes, David S and Light, Mark E (for: Hursthouse, Michael B and Coles, Simon J).

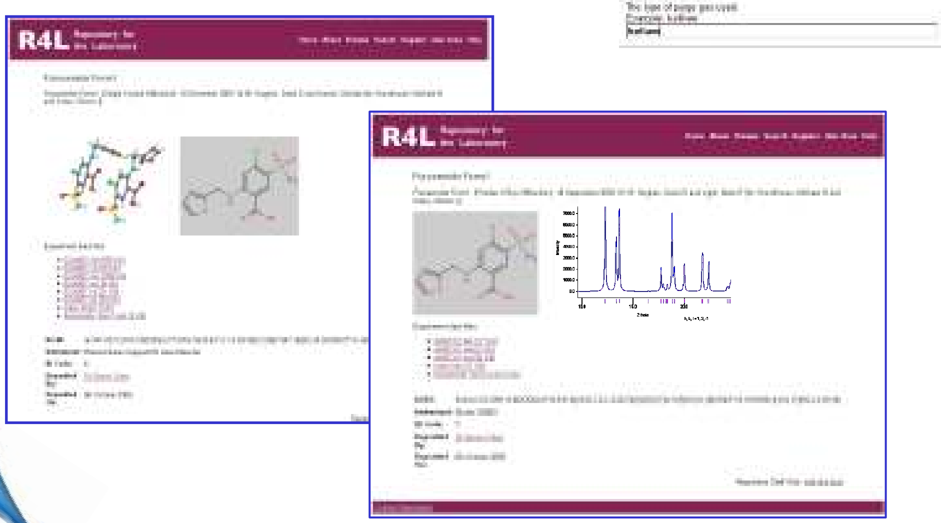
Furosemide Form1 (Infra-Red Spectroscopy) 29 July 2004 12:00. Hughes, David S (for: Hursthouse, Michael B and Coles, Simon J).

Furosemide Form1 (Single Crystal Diffraction) 10 December 2003 14:08. Hughes, David S and Karami, Sahar (for: Hursthouse, Michael B and Coles, Simon J).

This list was generated on Sun Nov 6 19:38:03 GMT 2005.

Contact Information

Search / Browse



R4L Repository for the Laboratory

Furosemide Form1

Experiment List

R4L Repository for the Laboratory

Furosemide Form1

Experiment List

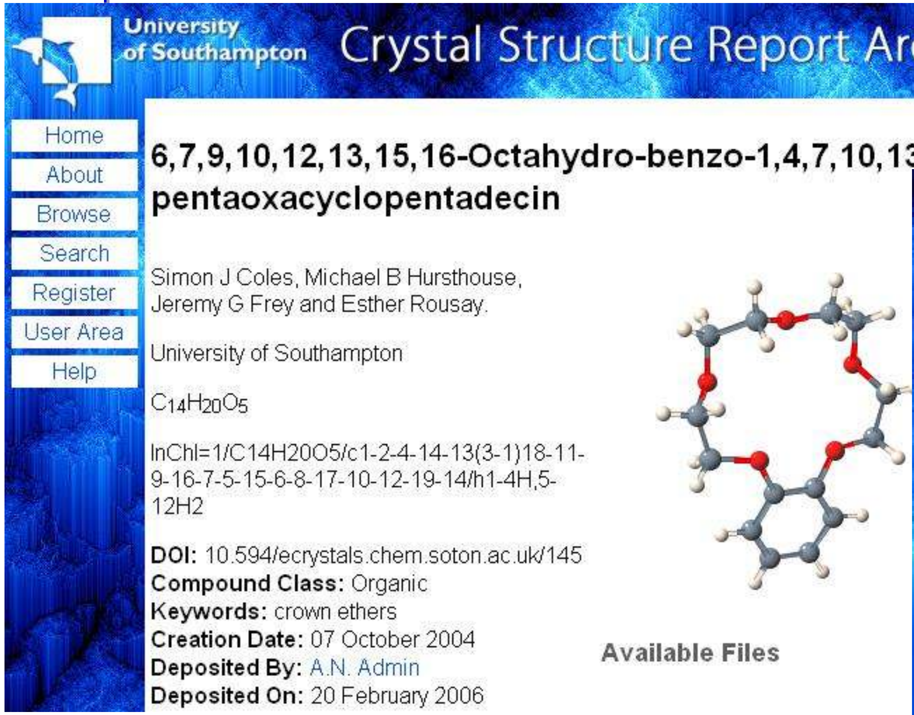
Data dissemination and curation by the scientist and host institution

eBank-UK and the eCrystals Repository

University of Southampton Crystal Structure Report Archive

eBank UK

The eCrystals Data Archive



University of Southampton Crystal Structure Report Archive

6,7,9,10,12,13,15,16-Octahydro-benzo-1,4,7,10,13-pentaoxacyclopentadecin

Simon J Coles, Michael B Hursthouse, Jeremy G Frey and Esther Rousay.

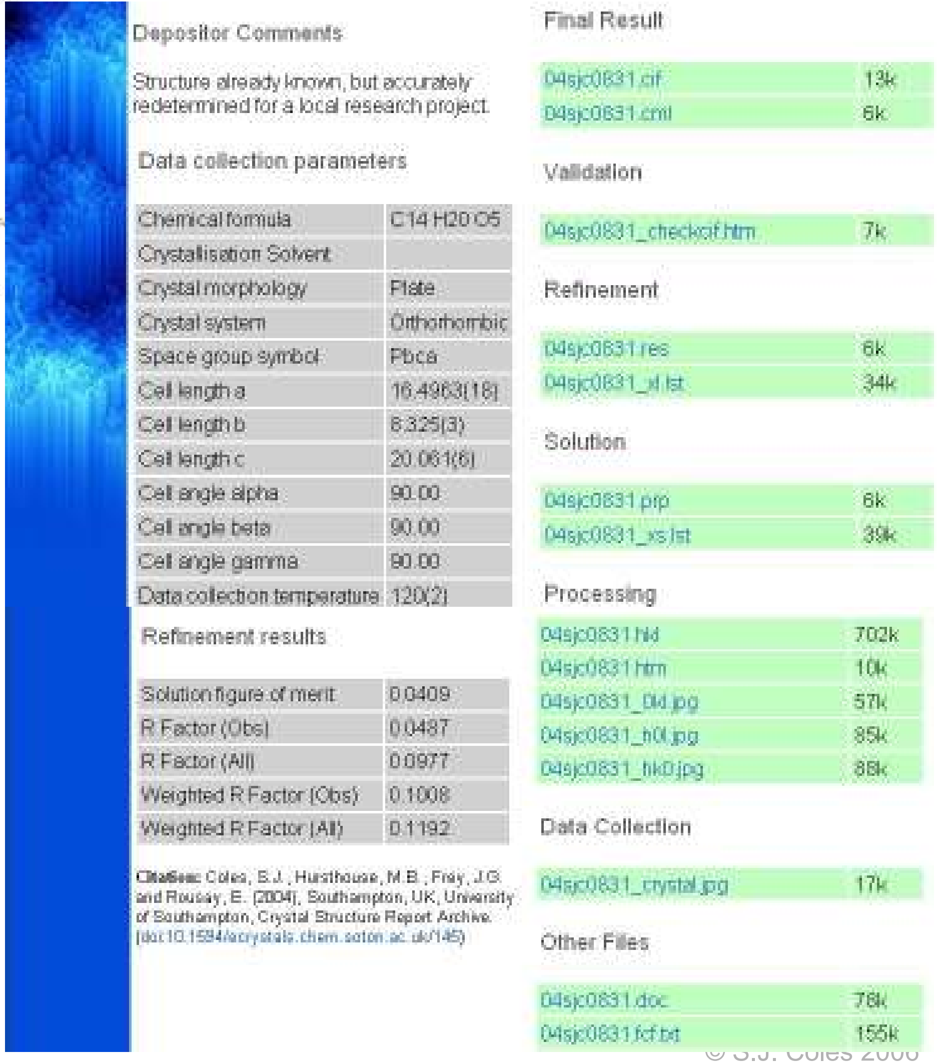
University of Southampton

$C_{14}H_{20}O_5$

InChI=1/C14H20O5/c1-2-4-14-13(3-1)18-11-9-16-7-5-15-6-8-17-10-12-19-14/h1-4H,5-12H2

DOI: 10.594/ecrystals.chem.soton.ac.uk/145
Compound Class: Organic
Keywords: crown ethers
Creation Date: 07 October 2004
Deposited By: A.N. Admin
Deposited On: 20 February 2006

Available Files



Depositor Comments

Structure already known, but accurately redetermined for a local research project

Data collection parameters

Chemical formula	C 14 H20 O5
Crystallisation Solvent	
Crystal morphology	Plate
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	16.4963(18)
Cell length b	8.325(3)
Cell length c	20.061(6)
Cell angle alpha	90.00
Cell angle beta	90.00
Cell angle gamma	90.00
Data collection temperature	120(2)

Refinement results

Solution figure of merit	0.0409
R Factor (Obs)	0.0487
R Factor (All)	0.0977
Weighted R Factor (Obs)	0.1008
Weighted R Factor (All)	0.1192

Citation: Coles, S.J., Hursthouse, M.B., Frey, J.G. and Rousay, E. (2004), Southampton, UK, University of Southampton, Crystal Structure Report Archive. (doi:10.1594/ecrystals.chem.soton.ac.uk/145)

Final Result

04sjc0831.cif	13k
04sjc0831.cml	6k

Validation

04sjc0831_checkcif.htm	7k

Refinement

04sjc0831.res	6k
04sjc0831_xl.lst	34k

Solution

04sjc0831.prp	6k
04sjc0831_xs.lst	39k

Processing

04sjc0831.hkl	702k
04sjc0831.htm	10k
04sjc0831_0kl.jpg	57k
04sjc0831_h0l.jpg	85k
04sjc0831_hk0.jpg	88k

Data Collection

04sjc0831_crystal.jpg	17k

Other Files

04sjc0831.doc	78k
04sjc0831.fcf.txt	155k

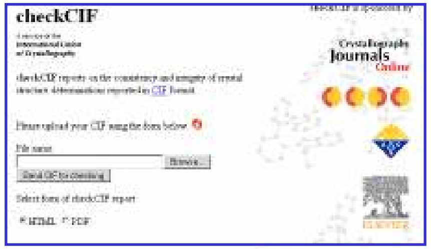
http://ecrystals.chem.soton.ac.uk

Metadata Publication

- Using simple Dublin Core
  - Crystal structure
  - Title (Systematic IUPAC Name)
  - Authors
  - Affiliation
  - Creation Date
- Additional chemical information through Qualified Dublin Core
  - Empirical formula
  - International Chemical Identifier (InChI)
  - Compound Class & Keywords
- Specifies which ‘datasets’ are present in an entry
- DOI *http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145*
- Rights & Citation *http://ecrystals.chem.soton.ac.uk/rights.html*
- Application Profile *http://www.ukoln.ac.uk/projects/ebank-uk/schemas/*

Metadata and Data Quality Control

Data manipulation toolbox



checkCIF

A service of the International Union of Crystallography

checkCIF reports on the consistency and integrity of crystal structure determinations reported in CIF format.

Please upload your CIF using the form below

File name

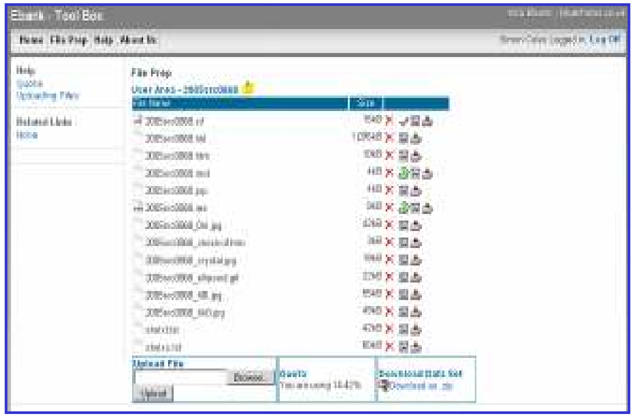
Browse...

Send CIF for checking

Select form of checkCIF report

● HTML ○ PDF

Value added



Ebank - Tool Box

Home | File Prep | Help | About Us

Help
Quota
Uploading Files

Related Links
Home

File Prep

User Area - 2005src0888

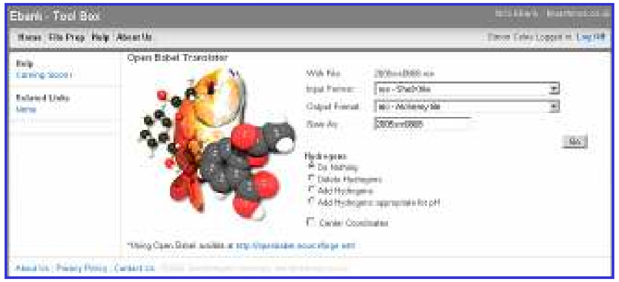
File Name	Size
2005src0888.cif	5540
2005src0888.hkl	129540
2005src0888.htm	1060
2005src0888.mol	4400
2005src0888.pgs	4400
2005src0888.res	3400
2005src0888_Ort.jpg	4260
2005src0888_checkcif.htm	3600
2005src0888_crystal.jpg	9860
2005src0888_ellipsoid.gif	2260
2005src0888_fit.jpg	5540
2005src0888_hkl.jpg	4560
shelx.lst	4260
shelx.lst	8040

Upload File
Browse...
Upload

Quota
You are using 10.42%

Download Data Set
Download as .zip

Format conversion



Ebank - Tool Box

Home | File Prep | Help | About Us

Simon Coles Logged in. Log Off

Help
Converting Files

Related Links
Home

Open Babel Translator

With File: 2005src0888.res

Input Format: res - ShelX file

Output Format: alc - Alchemy file

Save As: 2005src0888

Go

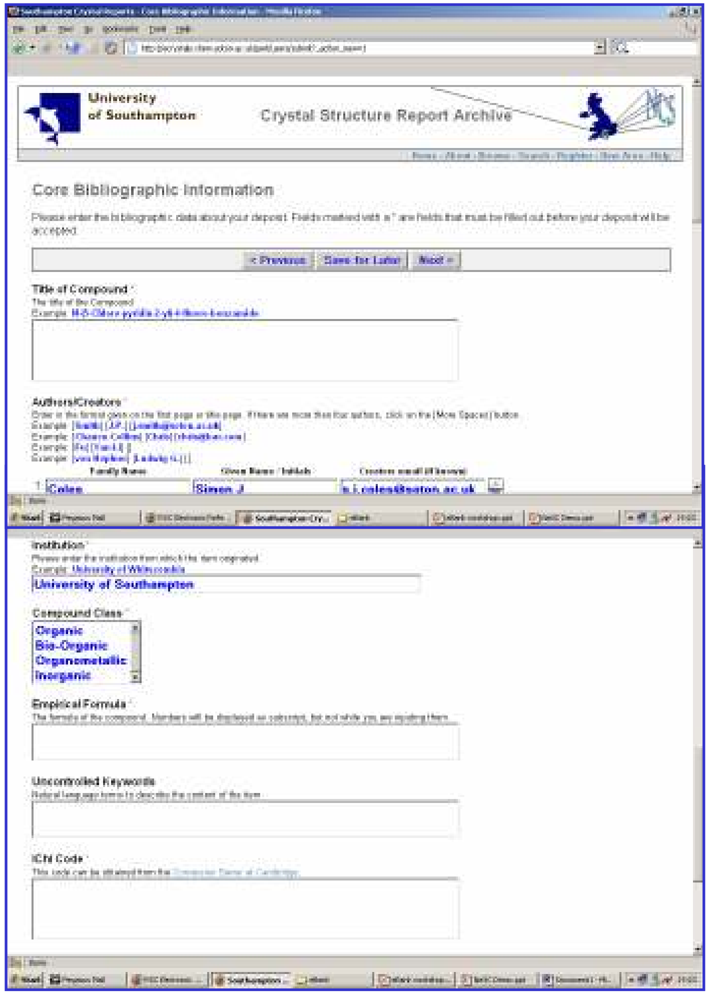
Hydrogens
● Do Nothing
○ Delete Hydrogens
○ Add Hydrogens
○ Add Hydrogens appropriate for pH

□ Center Coordinates

*Using Open Babel available at http://openbabel.sourceforge.net/

About Us | Privacy Policy | Contact Us

Associated Metadata



Southampton Crystal Reports - Core Bibliographic Information - Mozilla Firefox

University of Southampton

Crystal Structure Report Archive

Home | About | Browse | Search | Register | User Area | Help

Core Bibliographic Information

Please enter the bibliographic data about your deposit. Fields marked with a * are fields that must be filled out before your deposit will be accepted.

< Previous | Save for Later | Next >

Title of Compound *
The title of the Compound
Example: N-(5-Chloro-pyridin-2-yl)-4-fluoro-benzamide

Authors/Creators *
Enter in the format given on the first page of the page. If there are more than four authors, click on the [More Spaces] button
Example: [Smith] [J.P.] [j.smith@soton.ac.uk]
Example: [Chaucer-Collins] [Chris] [chris@bar.com]
Example: [Fu] [Yan-Li] []
Example: [von Hopfen] [Ludwig H.] []

	Family Name	Given Name / Initials	Creators email (if known)
1.	Coles	Simon J	s.j.coles@soton.ac.uk

Institution *
Please enter the institution from which the item originated
Example: University of Whitecastle
University of Southampton

Compound Class *
Organic
Bio-Organic
Organometallic
Inorganic

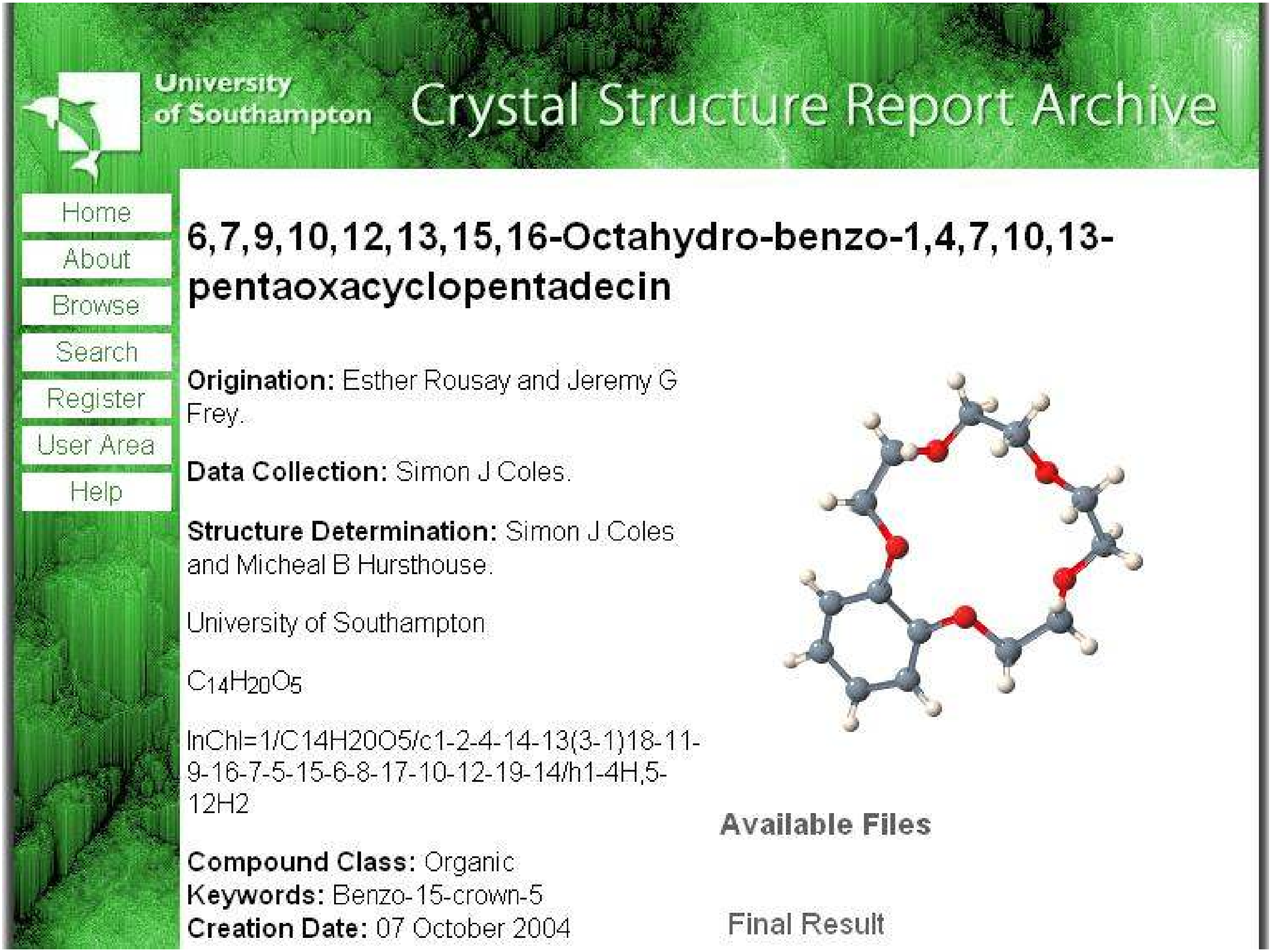
Empirical Formula *
The formula of the compound. Numbers will be displayed as subscripts, but not while you are editing them.

Uncontrolled Keywords
Natural language terms to describe the content of the item

ICN Code *
This code can be obtained from the Chemical Data of Cambridge

© S.J. Coles 2006

Laboratory Data Management and Archive



6,7,9,10,12,13,15,16-Octahydro-benzo-1,4,7,10,13-pentaoxacyclopentadecin

Origination: Esther Rousay and Jeremy G Frey.

Data Collection: Simon J Coles.

Structure Determination: Simon J Coles and Micheal B Hursthouse.

University of Southampton

$C_{14}H_{20}O_5$

InChI=1/C14H20O5/c1-2-4-14-13(3-1)18-11-9-16-7-5-15-6-8-17-10-12-19-14/h1-4H,5-12H2

Compound Class: Organic
Keywords: Benzo-15-crown-5
Creation Date: 07 October 2004

Available Files

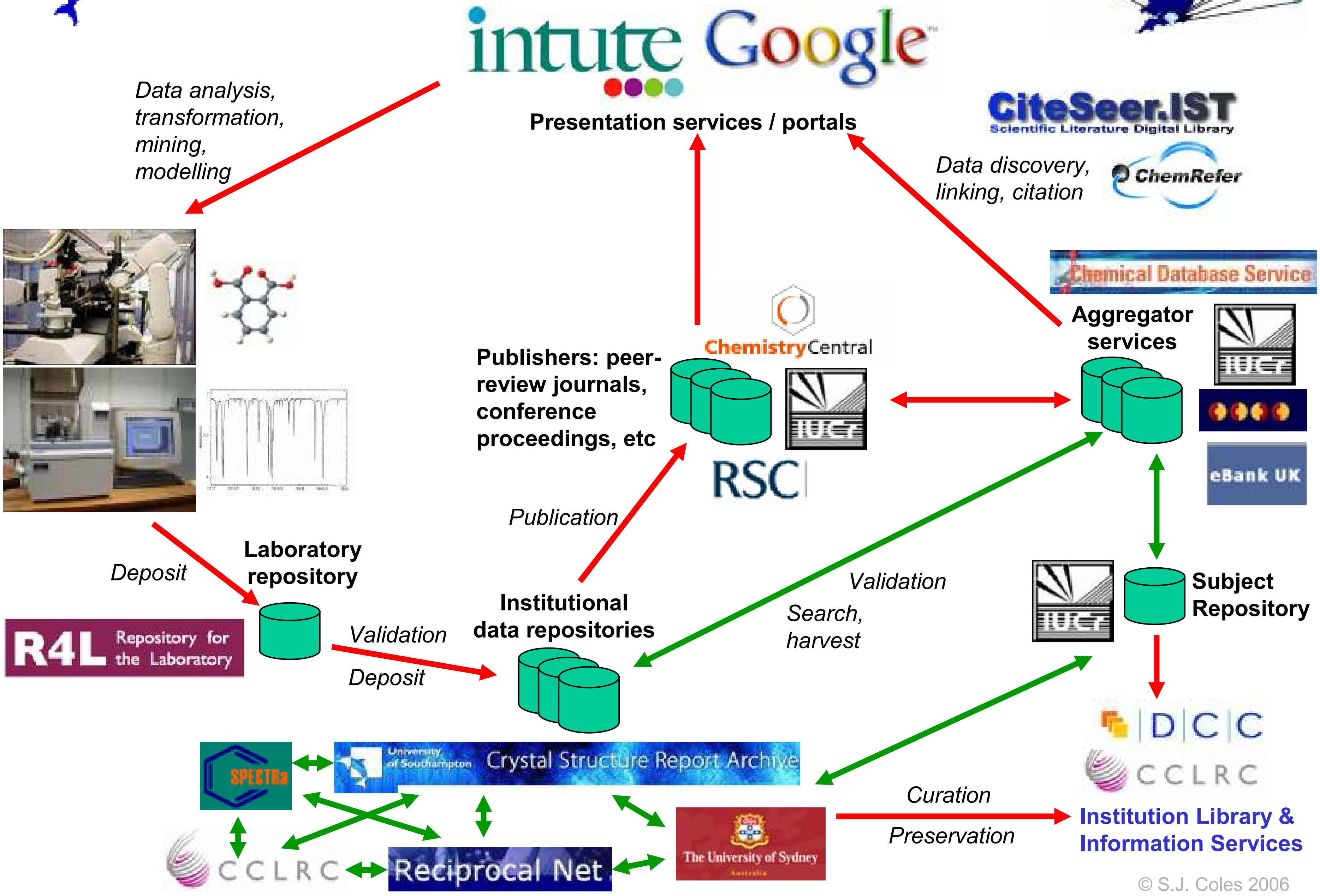
Final Result

Institutional data repositories and harvesting, aggregation and curation by data centres and third party services

eBank-UK Phase 3 – The eCrystals Federation

eBank UK

The eCrystals ‘Global Federation’ Model



Presentation services / portals

Data analysis, transformation, mining, modelling

Data discovery, linking, citation

Publishers: peer-review journals, conference proceedings, etc

Aggregator services

Publication

Deposit

Laboratory repository

Validation

Deposit

Institutional data repositories

Validation

Search, harvest

Subject Repository

Curation

Preservation

Institution Library & Information Services

intute
Google
CiteSeer.IST
Scientific Literature Digital Library
ChemRefer
Chemical Database Service
ChemistryCentral
IUCr
RSC
eBank UK
R4L Repository for the Laboratory
SPECTRa
University of Southampton
Crystal Structure Report Archive
CCLRC
Reciprocal Net
The University of Sydney
Australia
DCC

Exploring the heterogeneous landscape of data repositories

University of Southampton Crystal Structure Report Archive

- Different software platforms
- Different administrative domains
- Different Institutional structure
- Institutional vs Subject repositories
- Data Repository Interoperability

SPECTRa

Reciprocal Net

The University of Sydney
Australia

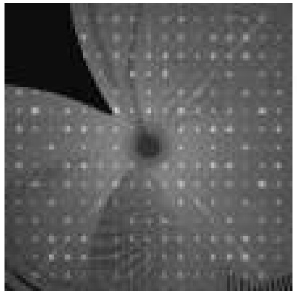
CCLRC

IUCr

ORE

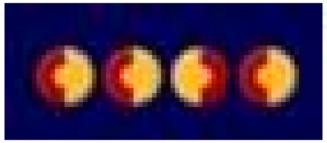
Preservation and curation by data centres & Institutions

G bytes



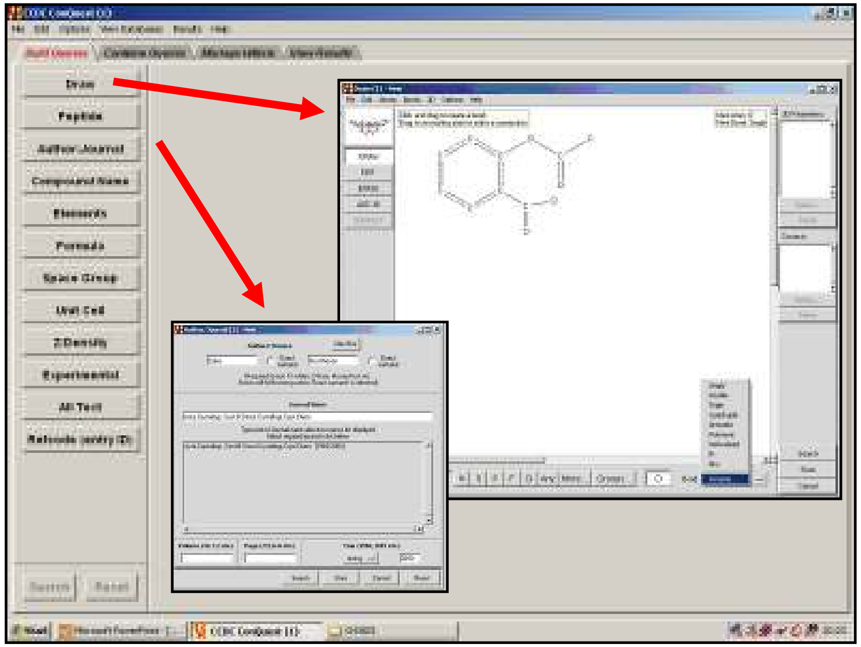
M bytes

k bytes

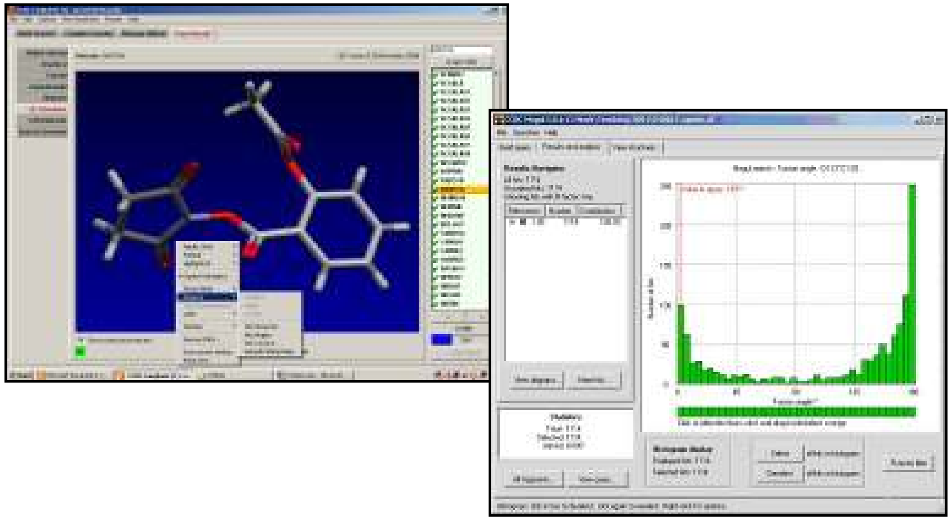


Institution Library &
Information Services

Harvesting, aggregation, value addition and curation by data centres



Cambridge Crystallographic Data Centre



IUCr

Chemical Database Service

eBank UK

The relationship with (conventional?) publication protocols and procedures

- Discipline-based publication

IUCr

- Domain-based publication

RSC

- Open Access publication

Chemistry Central

Aggregation, linking and information provision by third party services

- Indexing and aggregating with other datasets

Google

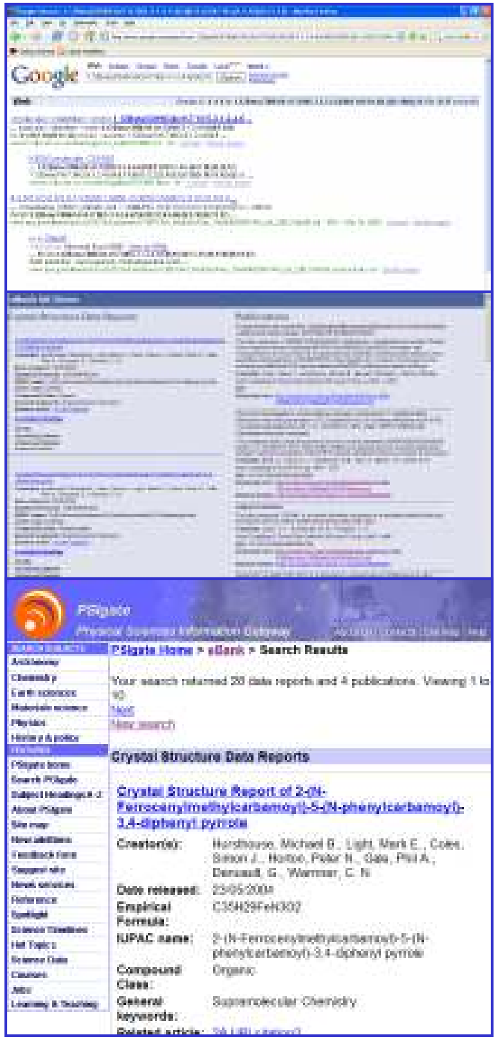
Chemical Database Service

- Aggregating and linking between datasets and articles

eBank UK

- Integration into information portals

intute



PSigate Home > eBank > Search Results

Your search returned 20 data reports and 4 publications. Viewing 1 to 10

Next

New search

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
Compound Class:	Organic
General keywords:	Supramolecular Chemistry
Related article:	[illegible]