

Integrating research data into the publication workflow: eBank UK experience

Rachel Heery, UKOLN, University of Bath

<http://www.ukoln.ac.uk/projects/ebank-uk/>

PV-2004, ESRIN Centre, Frascati, 5-7 October 2004



Overview

More effective curation by integrating research data and publications

- eScience agenda
 - Imperative to re-use data
 - Publication at source
- Innovations in scholarly communications
 - Open Access
 - Institutional repositories
- eBank UK
 - Integrating research data and journal articles
 - Information architecture and data flow
 - Data model and schemas
- Challenges for the future

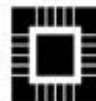


eBank project team

- **University of Southampton**
- Les Carr
- Simon Coles
- Jeremy Frey
- Chris Gutteridge
- Mike Hursthouse

- **University of Manchester**
John Blunden-Ellis

- **UKOLN, University of Bath**
- Michael Day
- Monica Duke
- Rachel Heery
- Liz Lyon



Electronics and
Computer Science



University
of Southampton

Imperative to re-use research data



“The next generation of research breakthroughs will rely upon new ways of handling the immense amounts of data that are being produced by modern research methods and equipment, such as telescopes, particle accelerators, genome sequencers and biological imagers....Similar developments are having an impact in the arts and humanities, and in the social sciences.”

*A Vision for Research,
Research Councils UK, December 2003*



UK Parliamentary Committee report



“It is envisaged that the sharing of primary data would prevent unnecessary repetition of experiments and enable scientists to build directly on each others’ work, creating greater efficiencies and productivity in the research process.”

Calls for new modes of curation for digital data

- Publication
- Discovery
- Re-use
- Preservation



eBank motivation

- Publication bottleneck in many scientific communities
- Small percentage of data referenced in literature
- Limited amount of results data
- Publication at source
- Open repositories
- Link data to research literature
- More timely access

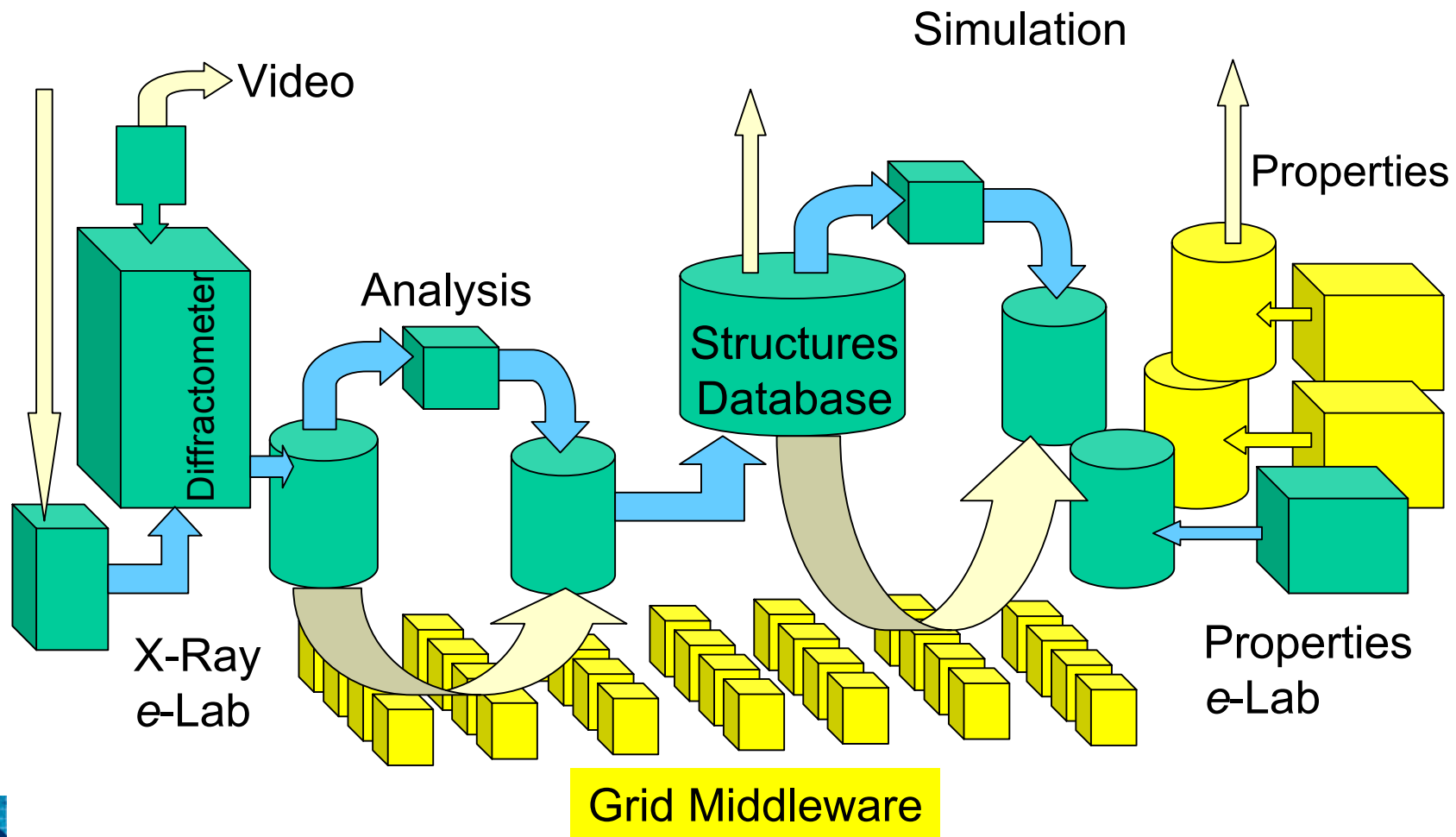


eBank focus on crystallography

- Computer controlled instruments
- Generates large quantities of digital data and metadata automatically
- Requirement for curaton of data
- Strict workflow
- Data formatted to international standard
 - Crystallographical Information File (CIF) maintained by the International Union of Crystallography
- CombeChem: funded by UK eScience programme



CombeChem: an eScience project



Emerging infrastructure to support curation of digital data



Improving access to research publications

- Repositories
 - Subject based (arXiv, CogPrints)
 - Institutional (CDL, MIT)
 - Supporting technology (DSpace, eprints.org)
- Open Access
 - Self archiving peer reviewed journal articles
 - ‘Toll free’ journals (free at point of use)
 - Supporting technology (OAI-PMH)



Potential for integrating access to data and publications

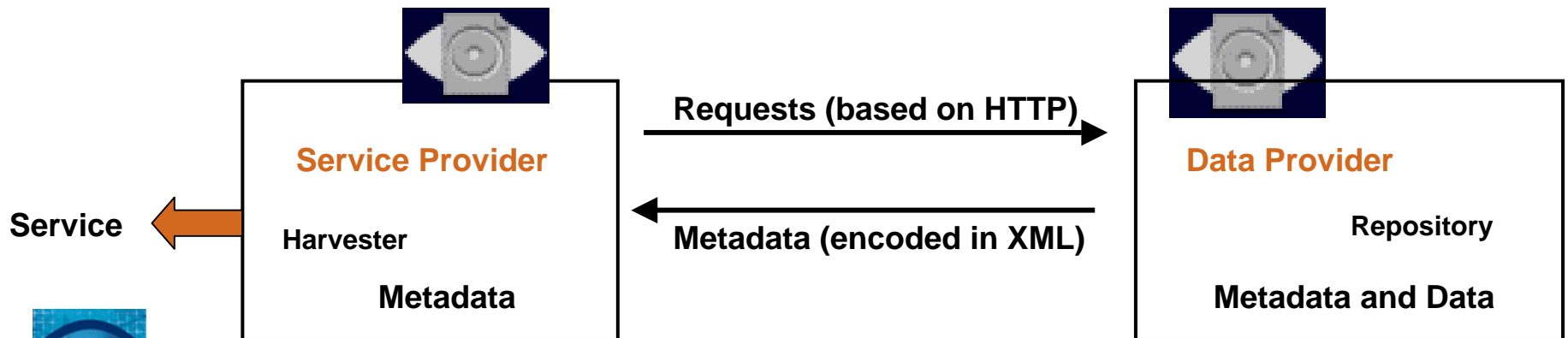
Supporting technology: Open Archives Initiative

- Protocol for Metadata Harvesting (OAI-PMH)
- Architecture of the OAI-PMH
 - Harvest available metadata from Data Providers
 - Place aggregated metadata in a repository
 - Expose aggregated metadata via a Web interface
- Potential for added value services...
- www.openarchives.org



Architecture of the OAI PMH

- Consistent interfaces for data provider and service provider
- Low barrier protocol / effortless implementation
- Based on existing standards (e.g. HTTP, XML, DC)



The ePrints UK Project

The ePrints UK project is developing a series of national, discipline-focused services through which the higher and further education community can access the collective output of e-print papers available from compliant Open Archive repositories, particularly those provided by UK



...assisting scholarly communication

[Home](#) | [About](#) | [Partners](#) | [Documents](#) | [Links](#) | [Contacts](#)

SHERPA

SHERPA aims to investigate issues to do with the future of scholarly communication and publishing. In particular, it is initiating the development of openly accessible institutional digital repositories of research output in a number of research universities. These so-called 'e-print archives' will contain papers by researchers from the participating institutions.

The project will investigate the IPR, quality control and other key management issues associated with making the research literature freely available to the research community. It will also investigate technical questions, including interoperability between repositories and digital preservation of e-prints.



News

[Wellcome Trust report confirms viability of "publication charge" model for Open Access Journals](#)



open archives forum

open archives

Home

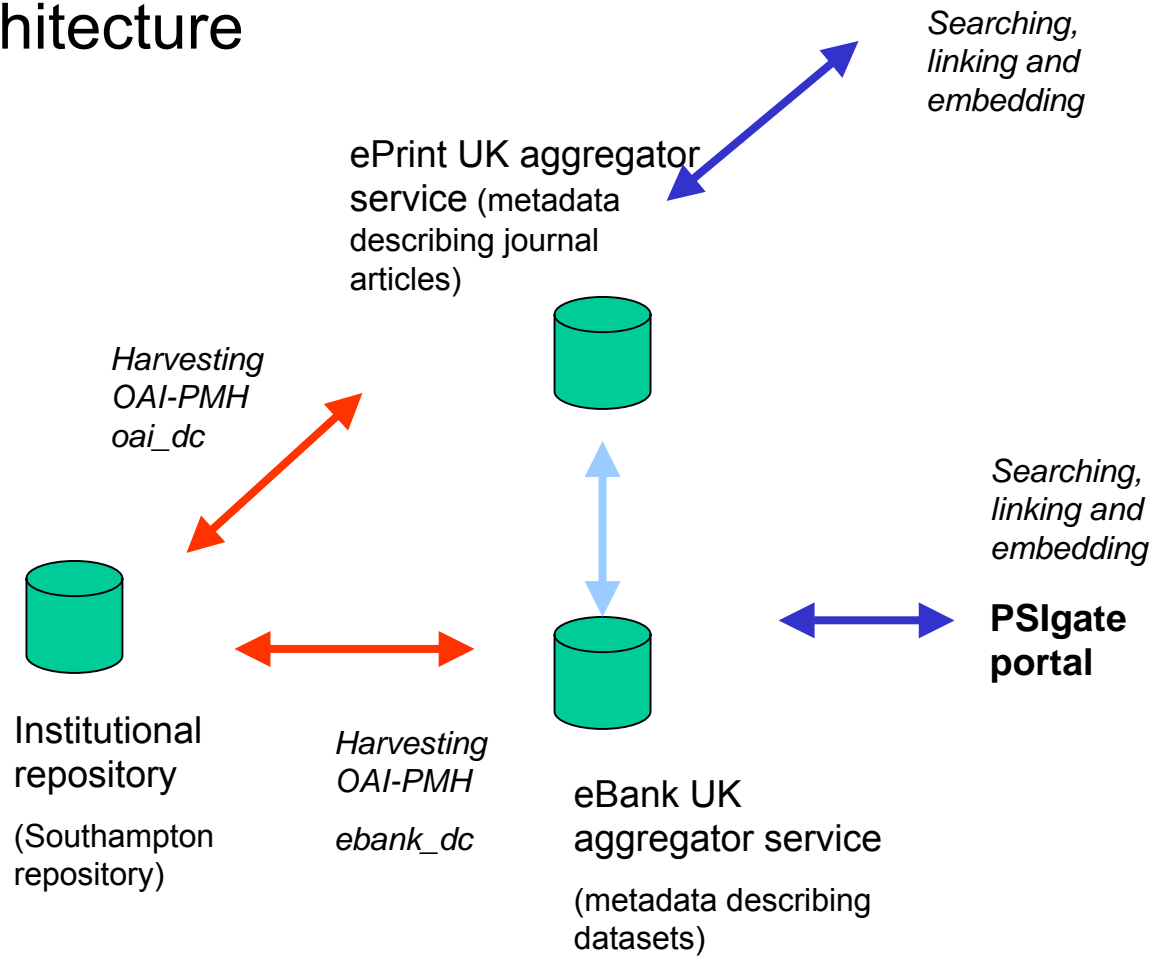
eBank in a nutshell

To develop pilot service linking journal articles and scientific datasets (September 2003 - October 2005)

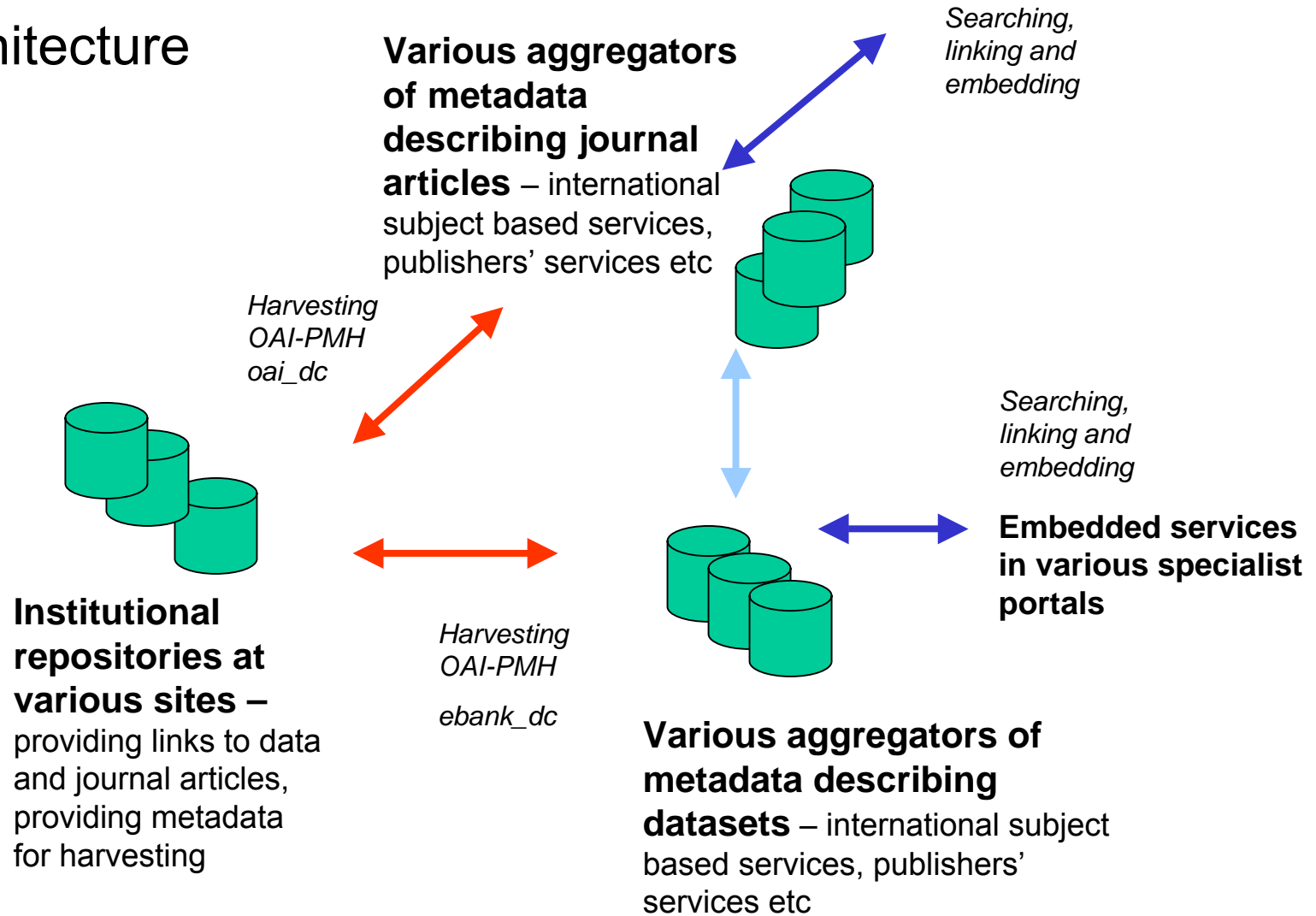
- Create institutional repository of Crystallography Data (at Southampton)
- Modify repository software to handle datasets (eprints.org at Southampton)
- Demonstrate eBank search service linked to ePrints UK, indexing harvested descriptions of datasets and journal articles (at UKOLN)
- Embed eBank service into PSIGate subject gateway (at Manchester)



eBank architecture



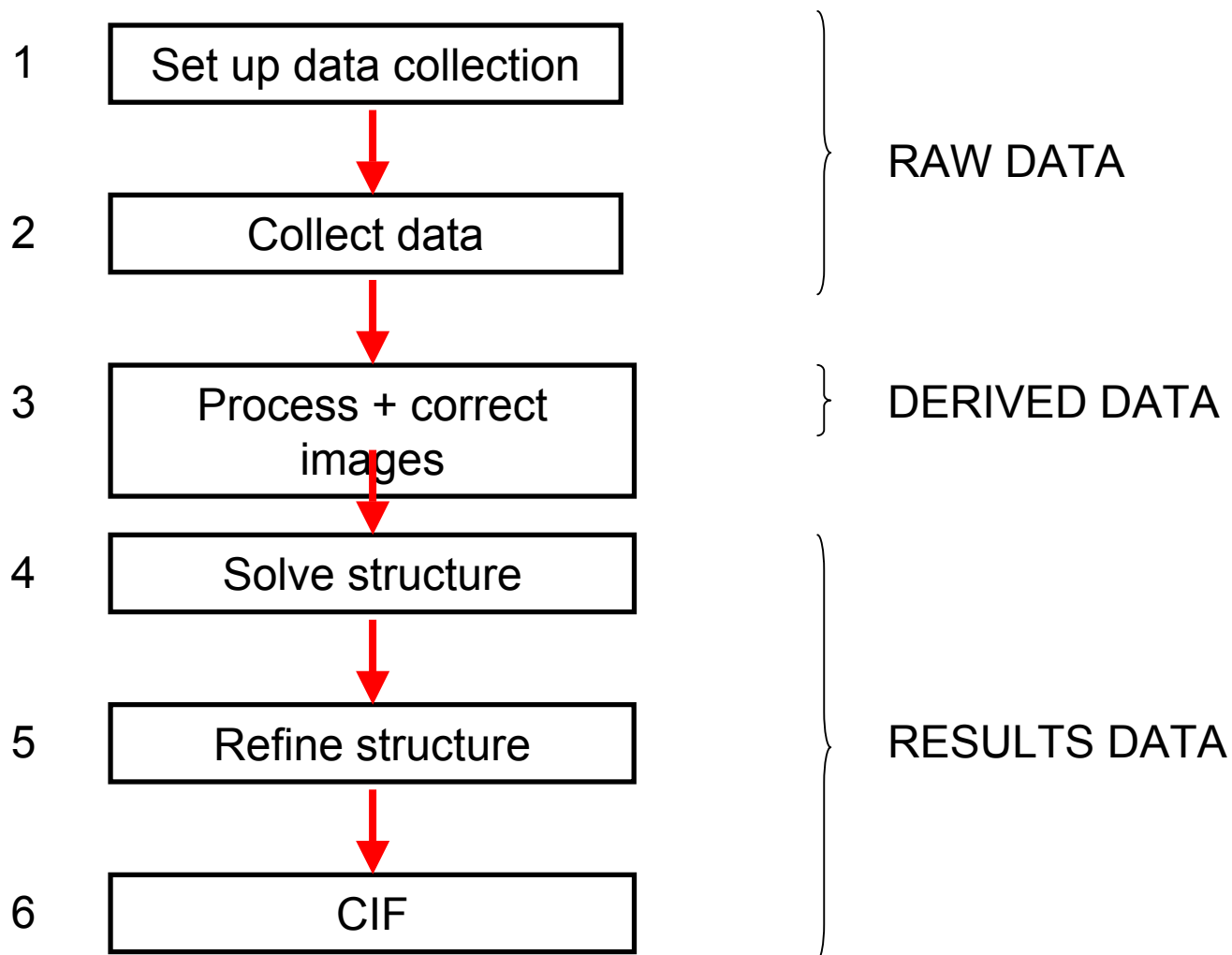
Potential extended architecture



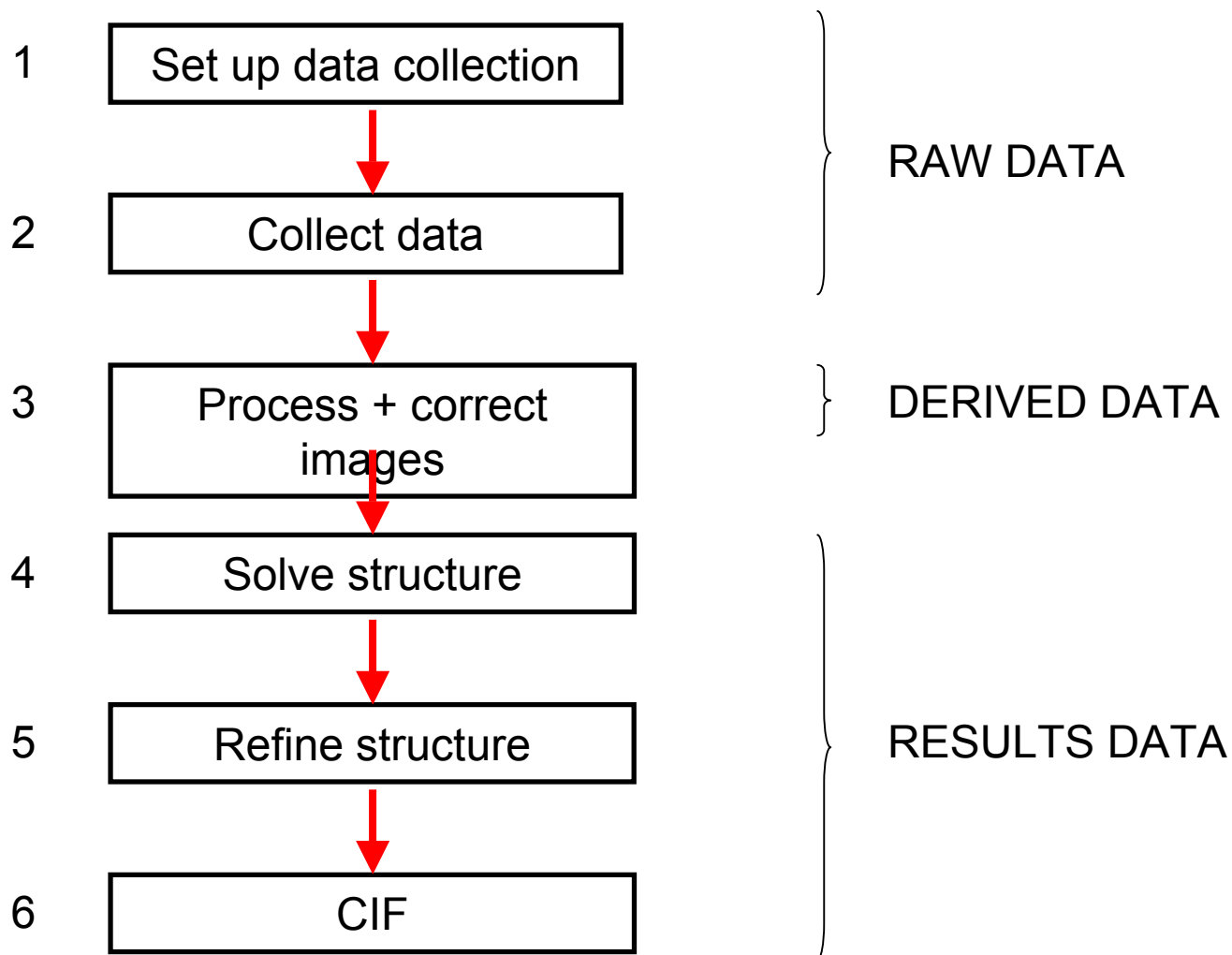
First steps: establishing common ground...

- Understand the data creation process
- Terminology and definitions
 - Data
 - Metadata
 - Datafile
 - Dataset
 - Data holding
- Different views
 - Digital library researchers, computer scientists, chemists
 - Generic vs specific
 - Modeller vs practitioner
- Data modelling
- Defining metadata schema

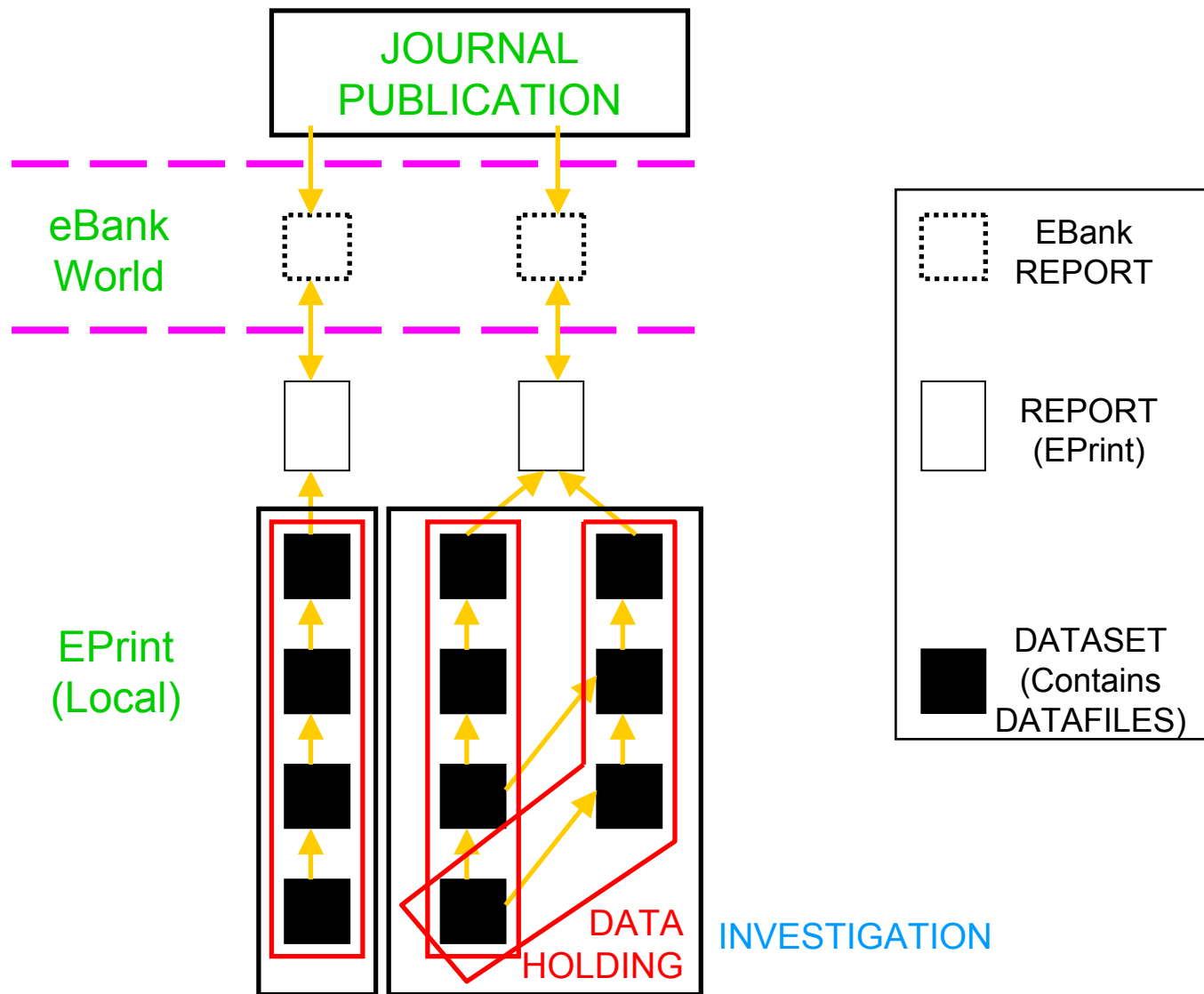
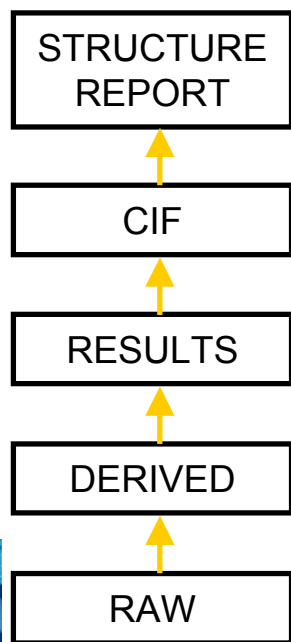
Crystallographic data workflow



Crystallographic data workflow



Linking Crystallography data and journal ePrints



Crystallography data model

Name	Description of the stage	Files associated with this stage			Metadata associated with this stage	
		File	Type	Description	Name	Data Type
Initialisation	Mount new sample on diffractometer Parameterisation to set up data collection	datoinfon i*.kcd .dx	ASCII BINARY ASCII	Parameters for producing i*.kcd files Unit cell determination images Unit cell	Morphology	STRING
Collection	Collect Data	collect.mat datoinfon .py s*.kcd *scan*.jpg	ASCII ASCII ASCII BINARY JPG	Orientation of the crystal Command file Proprietary configuration file Diffraction images Visual version of kcd file	Instrument_Type Temperature Software_Name (x n) Software_Version (x n) Software_URL (x n)	STRING INTEGER STRING (x n) INTEGER (x n) URL (x n)
Processing	Process and correct images	scale_all.in scale_all.out .hkl .htm	ASCII ASCII ASCII HTML	Result of processing Result of correction on processed data Derived data set Report file	Cell a Cell b Cell c Cell alpha Cell beta Cell gamma Crystal system Completeness Software_Name (x n) Software_Version (x n) Software_URL (x n)	INTEGER INTEGER INTEGER INTEGER INTEGER INTEGER STRING INTEGER STRING (x n) INTEGER (x n) URL (x n)
Solution	Solve Structures	.ppp xs.lst	ASCII ASCII	Symmetry file, log of process Solution log file	Space_group Figure_of_merit Software_Name (x n) Software_Version (x n) Software_URL (x n)	STRING INTEGER STRING (x n) INTEGER (x n) URL (x n)
Refinement	Refine Structure	xl.lst .res	ASCII ASCII	Final refinement listing Output coordinates	R1_obs wR2_obs R1_all wR2_all Software_Name (x n) Software_Version (x n) Software_URL (x n)	INTEGER INTEGER INTEGER INTEGER STRING (x n) INTEGER (x n) URL (x n)
CIF	Produce CIF	.cif	ASCII	Final results		
Report	Generate e-Print report	.html	HTML	Publication format (HTML/XHTML)	Authors Affiliations Formula Compound_name 2D_diagram	STRING STRING STRING STRING STRING

Metadata approach

- Extended Dublin Core for structure reports within institutional repository
- Both simple Dublin Core and extended Dublin Core are offered as alternative schemas for harvesting using OAI-PMH
- Exploring use of extended DC schema within DCMI
 - impact on aggregator service
- Engaging the broader scientific community to ensure different schemas are compliant and standards can emerge



Extended Dublin Core schema

- Additional chemical information in schema for harvesting e.g. empirical formula
- Schema contains International Chemical Identifier (InChI)
- Links to all datasets associated with an experiment
- Links to individual datasets *within* an experiment
- Links to eprints (and other published literature) derived from the data
- Using vocabularies specific to crystallography



Bis(μ^2 -4,6-bis((diphenylphosphino)oxy)-5-methyl-1,3-phenylene-C,C',P,P')-tetrakis(μ^2 -trifluoroacetato-O,O')-tetra-palladium chloroform solvate

cif

- [02SRC841.cif](#)
(20336)

rfne

- [02src841.res](#) (8460)
- [02src841_xl.lst](#)
(58006)

soln

- [02src841.PRP](#) (6019)
- [02src841_xs.lst](#)
(73362)

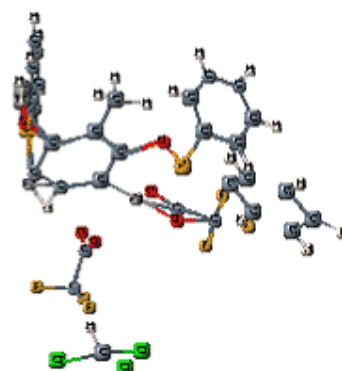
proc

- [02SRC841.HTM](#)
(6535)
- [02src841.HKL](#)
(1180322)

?

- [02src841_0KL.JPG](#)

Simon J Coles, Robin B Bedford, M E Blake, Michael B Hursthouse and P N Scully.



Creation Date: 18 March 2004

Deposited By: [Christopher Gutteridge](#)

Deposited On: 18 March 2004

_CHEMICAL_FORMULA_SUM:

C288 H200
Cl24 F48 O48
P16 Pd16

Structure reports link back to the underlying data...

Available Files

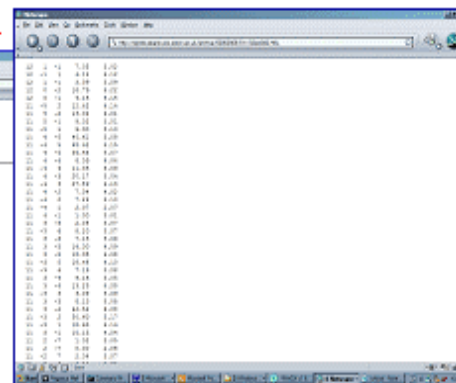
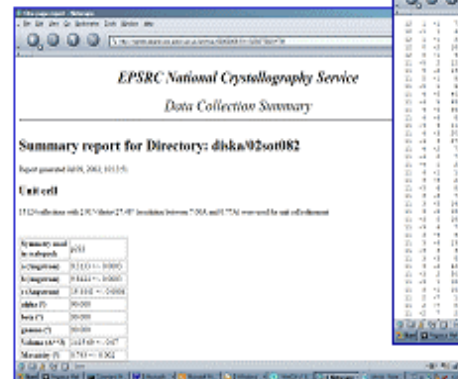
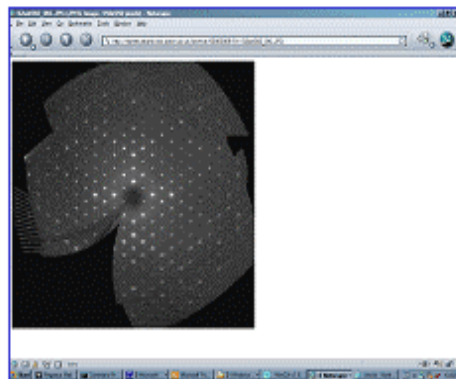
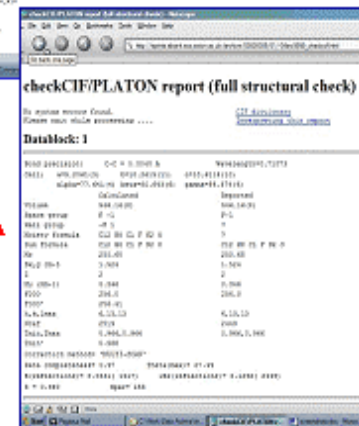
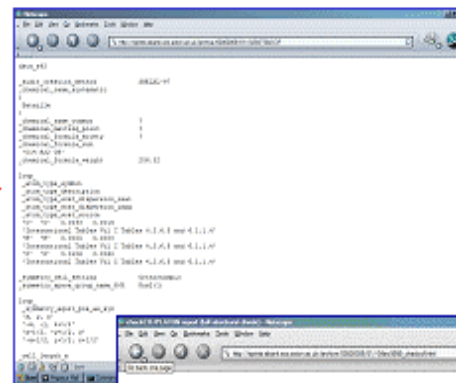
File Name	Size
02sot064.CIF	19k
02sot064.cml	8k
02sot064_checkcif.html	14k
02sot064.RES	9k
02sot064.PRP	5k
02SOT064.HTM	6k
02sot064.HKL	336k
02sot064.DOC	113k
02sot064.LST	49k

Data collection parameters

Chemical formula	C30 H26 Fe N2 O3
Crystallisation Solvent	
Crystal morphology	
Crystal system	Orthorhombic
Space group symbol	Pbca
Cell length a	6.0816(4)
Cell length b	24.8503(16)
Cell length c	31.120(3)
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	
R Factor (Obs)	0.0573
R Factor (All)	0.1185
Weighted R Factor (Obs)	0.1046
Weighted R Factor (All)	0.1243



eBank aggregator : search

eBank UK Demo

This is a prototype interface for the [eBank UK](#) JISC-funded project. It demonstrates an OAI-PMH aggregator service which cross searches a small sample of metadata records describing crystallography experiments (provided by the National Crystallography Service at the University of Southampton), and a small number of metadata records describing articles from the Crystallography literature (made available for use in this demo only by IUCr.) Links to the crystallography data sets and to the articles on line at the IUCr website are available in the search results.

Search for entries matching all the following:

Author =
CCDC Code =
IUPAC name =
Empirical Formula =
Compound Class =
General keywords =
Date released =
OR published in the last

Search within: Data Reports Publications e.g. journal articles

eBank UK Demo

This is a prototype interface for the [eBank UK](#) JISC-funded project. It demonstrates an OAI-PMH aggregator service which cross searches a small sample of metadata records describing crystallography experiments (provided by the National Crystallography Service at the University of Southampton), and a small number of metadata records describing articles from the Crystallography literature (made available for use in this demo only by IUCr.) Links to the crystallography data sets and to the articles on line at the IUCr website are available in the search results.

Search for entries matching all the following:

Author =
CCDC Code =
IUPAC name =
Empirical Formula =
Compound Class =
General keywords =
Date released =
OR published in the last day(s)

Search within: Data Reports Publications e.g. journal articles



Ebank aggregator: browse

eBank UK Demo

Crystal Structure Data Reports

[Crystal Structure Report of 5alpha-cholestane](#)
Creator(s): Coles, Simon J., Hursthouse, Michael B., Frampton, C. S.
Date released: 23/05/2004
Empirical Formula: C₂₇H₄₈
IUPAC name: 5alpha-cholestane
CCDC code: ZZZKGI01
Compound Class: Organic
Related article: <http://scripts.iucr.org/cgi-bin/getarticleid?issn=1500-5368&volume=58&page=0445&details=yes>

Publications

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 Online Vol 58 Issue Pt 4 pp. 0445 - 0446
DOI: 10.1107/S1500536802004786
Download from: <http://scripts.iucr.org/cgi-bin/getarticleid?issn=1500-5368&volume=58&page=0445&details=yes>
Related dataset: <http://icrystals.chem.soton.ac.uk/archive/0000051/>

eBank UK Demo

Crystal Structure Data Reports

[Crystal Structure Report of 2-\(N-Ferrocenylmethylcarbamoyl\)-5-\(N-phenylcarbamoyl\)-3,4-diphenylpyrrole](#)
Creator(s): Hursthouse, Michael B., Light, Mark E., Coles, Simon J., Horton, Peter N., Gale, Phil A., Desautel, G., Wanstler, C. N.
Date released: 23/05/2004
Empirical Formula: C₃₈H₃₅N₃O₂
IUPAC name: 2-(N-Ferrocenylmethylcarbamoyl)-5-(N-phenylcarbamoyl)-3,4-diphenylpyrrole
CCDC code: KUZSUU
Compound Class: Organic
General keywords: Supramolecular Chemistry
Related article: [2A URL status?](#)

Publications

A supramolecular assembly, aquatri(pentafluorophenyl)borane as its mixed dimethyl sulfone and water solvate, (H₂O)(C₆F₅)₃Me₂SO₂·H₂O
The title compound, C₁₉H₂₈F₁₅O₄S₂, 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 tri(pentafluorophenyl)borane, (H₂O)(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 Online Vol 59 Issue Pt 3 pp. 0134 - 0136
DOI: <http://scripts.iucr.org/cgi-bin/getarticleid?issn=1500-5368&volume=59&page=0134&details=yes>

Structural investigations of phosphorus-nitrogen compounds. 5. Relationships between molecular parameters of 2,2-diphenyl-4,6-cis-oxetra[etha]ferrocenyl-4,6-R₂-cyclophosphazenes (R = Cl, OCH₂CF₃, OPh, OMe, NHPh, NHBu) and substituent basicity constants
The syntheses and crystal structures of six new cis-ansa derivatives (NDP3Ph₂O)(CH₂CH₂CO₂R)₂ (R = Cl, OCH₂CF₃, OPh, OMe, NHPh, NHBu) are reported and the observed relationship between molecular parameters of the NCP ring and substituent basicity constants is discussed.

Creator(s): Beak, S., Coles, S. J., Hursthouse, M. B., Klic, A., Mayer, T. A., Shaw, R. A.
Acta Crystallogr B Vol 58 Pt 6 pp. 1067 - 1073
DOI: 10.1107/S0108768210201868
Download from: <http://scripts.iucr.org/cgi-bin/getarticleid?issn=0108-7682&volume=58&page=1067&details=yes>
Related dataset: <http://icrystals.chem.soton.ac.uk/archive/0000062/>

Your search returned 1 data reports and 1 publications. Viewing

Search for entries matching all the following

Author =

eBank UK Demo

Crystal Structure Data Reports

[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., Desautel, G., Wanstler, C. N.
Date released: 23/05/2004
Empirical Formula: C₂₉H₂₄F₁₀O₃
IUPAC name: 2-(N-Ferrocenylcarbamoyl)-5-(methoxycarbonyl)-3,4-diphenylpyrrole
CCDC code: KUZSUU
Compound Class: Organic
General keywords: Supramolecular Chemistry
Related article: [2A URL status?](#)

Publications

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 Online Vol 58 Issue Pt 4 pp. 0445 - 0446
DOI: 10.1107/S1500536802004786
Download from: <http://scripts.iucr.org/cgi-bin/getarticleid?issn=1500-5368&volume=58&page=0445&details=yes>
Related dataset: <http://icrystals.chem.soton.ac.uk/archive/0000051/>

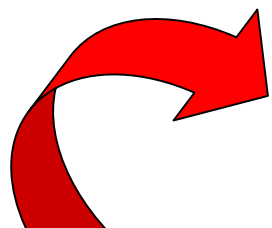
Ethyl (2S)-2-(2R)-2R',5S)-2',5-dimethyl-5-oxopenthydro-2,2'-bifuranyl-5-yl]-2-hydroxyethanoate

The framework of K₂Zn(H₂P₂O₇)₂·2H₂O contains acid dihydrophosphate-metalate layers linked by H₂O interactions and weak hydrogen bonds. Zn²⁺ cations are coordinated octahedrally by O atoms from two bidentate (H₂P₂O₇)₂ anions and two water molecules.



And finally...

eBank search embedded in a science portal



PSigate
Physical Sciences Information Gateway

PSigate Home > eBank

This is a prototype test interface to the eBank UK service providing access to data in the University of Southampton eCrystal data repository and elsewhere. eBank UK is a JISC-funded project which is a part of the Semantic Grid Programme. The project is being led by UKOLN in partnership with the Combechem project at the University of Southampton and PSigate.

Enter your search term(s)

Author

CCDC Code

IUPAC name

Empirical Formula

Compound Class

General keywords

Date released

OR published in the last

Search within Data Reports Publications e.g. journal articles

Search Clear

Copyright © PSigate. All rights reserved. PSigate is a service of the Resource Discovery Network (RDN). feedback@psigate.ac.uk

PSigate
Physical Sciences Information Gateway

About Us | Contacts | Site Map | Help

PSigate Home > eBank > Search Results

Your search returned 28 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

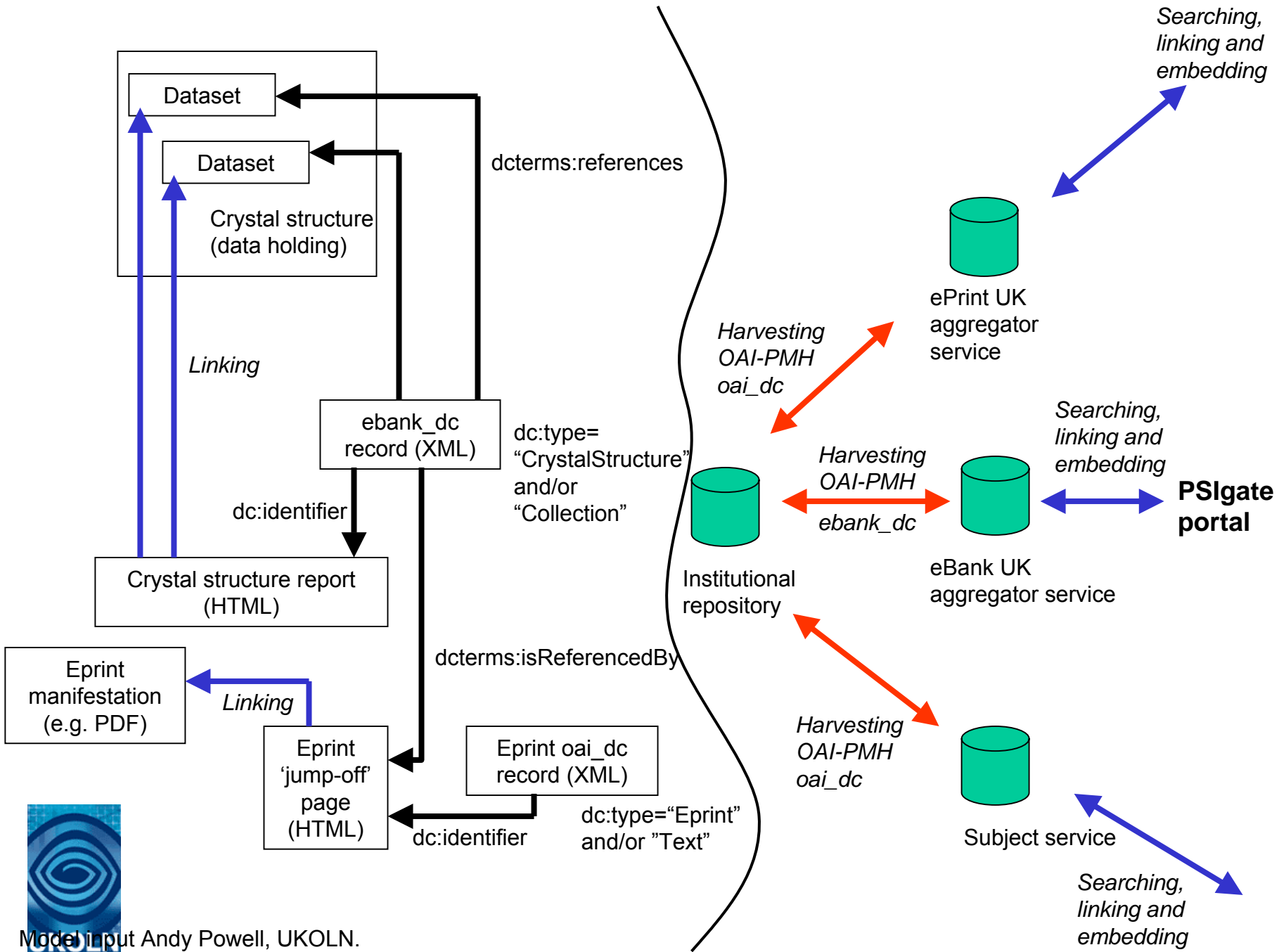
IUPAC name: 2-(N-Ferrocenylmethylcarbamoyl)-5-(N-phenylcarbamoyl)-3,4-diphenyl pyrrole

Compound Class: Organic

General keywords: Supramolecular Chemistry

Related article: [2A IIRL citation?](#)





Challenges for the future



Progress update

- Version 2.0 eBank metadata schema
- Enhanced ePrints.org software
- Pilot institutional e-data repository for harvesting (raw, derived, results data)
- Exports records as ebank_dc and oai_dc
- Pilot eBank UK aggregator service
- Developing search interface Version 1.0
- Testing with PSIgate physical sciences portal – embedding eBank UK



Plans for eBank Phase 2

- Progress towards generic data model for description of research datasets
 - Validate eBank schema against other schema
 - CLRC Scientific Metadata Model
- Modify eprints.org software to allow for more varied scientific data and schemas
- Investigate identifiers e.g. International Chemical Identifier (InChI code)



Plans for eBank Phase 2.....(contd.)

- Explore embedding in chemistry workflow

Potential to expand remit to

- wider range of crystallography data
- other chemistry sub-domains
- broader physical sciences



eBank (potential) links with eLearning

- Provide access to primary research data within learning materials
 - in the taught postgraduate curriculum in chemistry, undergraduate project work, chemical informatics courses
- Inclusion of e-research data in e-learning courses.
 - through links in reading lists, through essay assignments, through analytical problems, through practical work, through RDN PsiGATE links



In conclusion

- eBank demonstrates benefits to research community
- Potential for integration into digital library services
 - Moving from demonstrator to service, need to involve publishers and specialist services

The end...

Questions?

<http://www.ukoln.ac.uk/projects/ebank-uk/>

