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

UKOLN

UNIVERSITY OF BATH

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

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

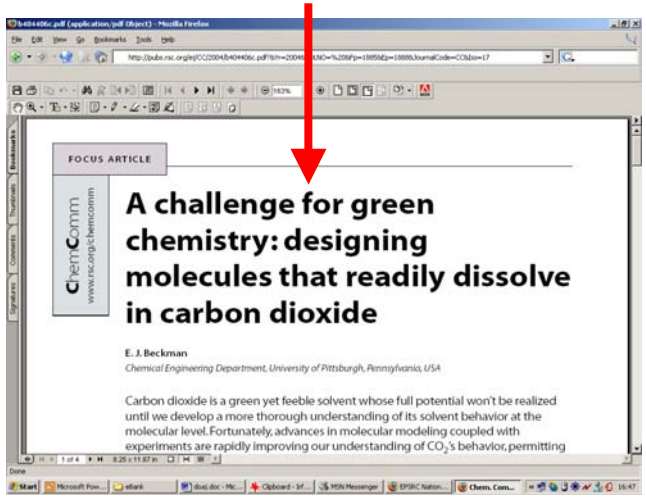
UK Parliamentary Committee report

"*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.*"

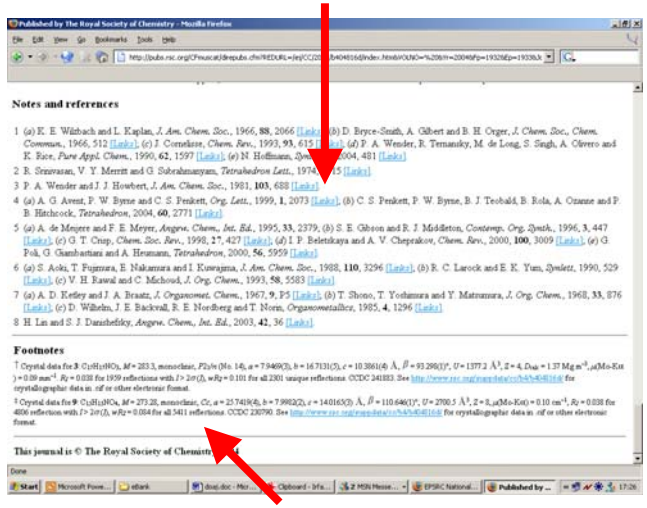
UKOLN

Current chemistry publishing protocols

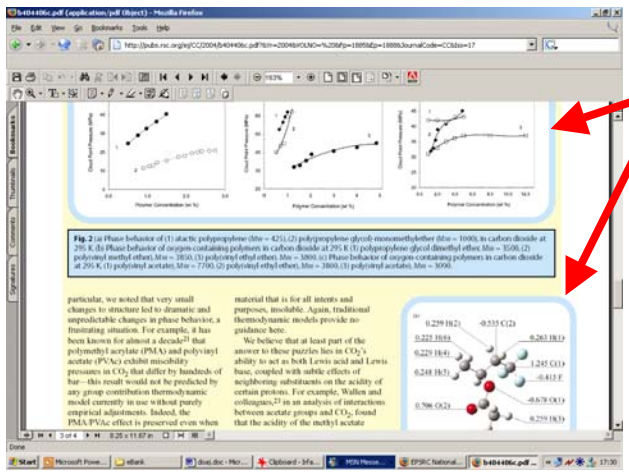
Ideas and interpretations



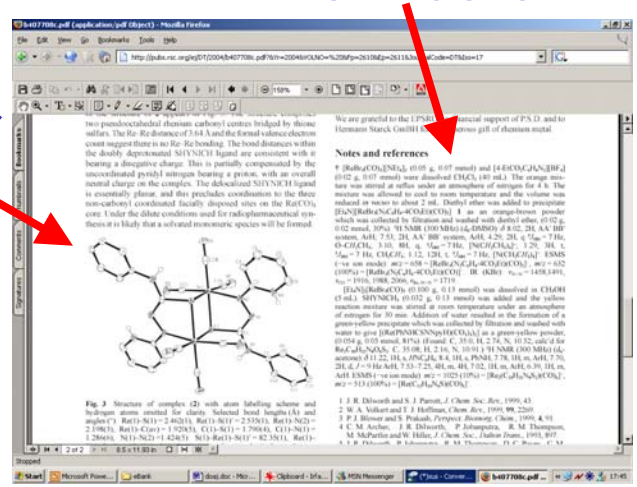
Hooks into the literature



Raw data!



Results & derived data



UKOLN

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

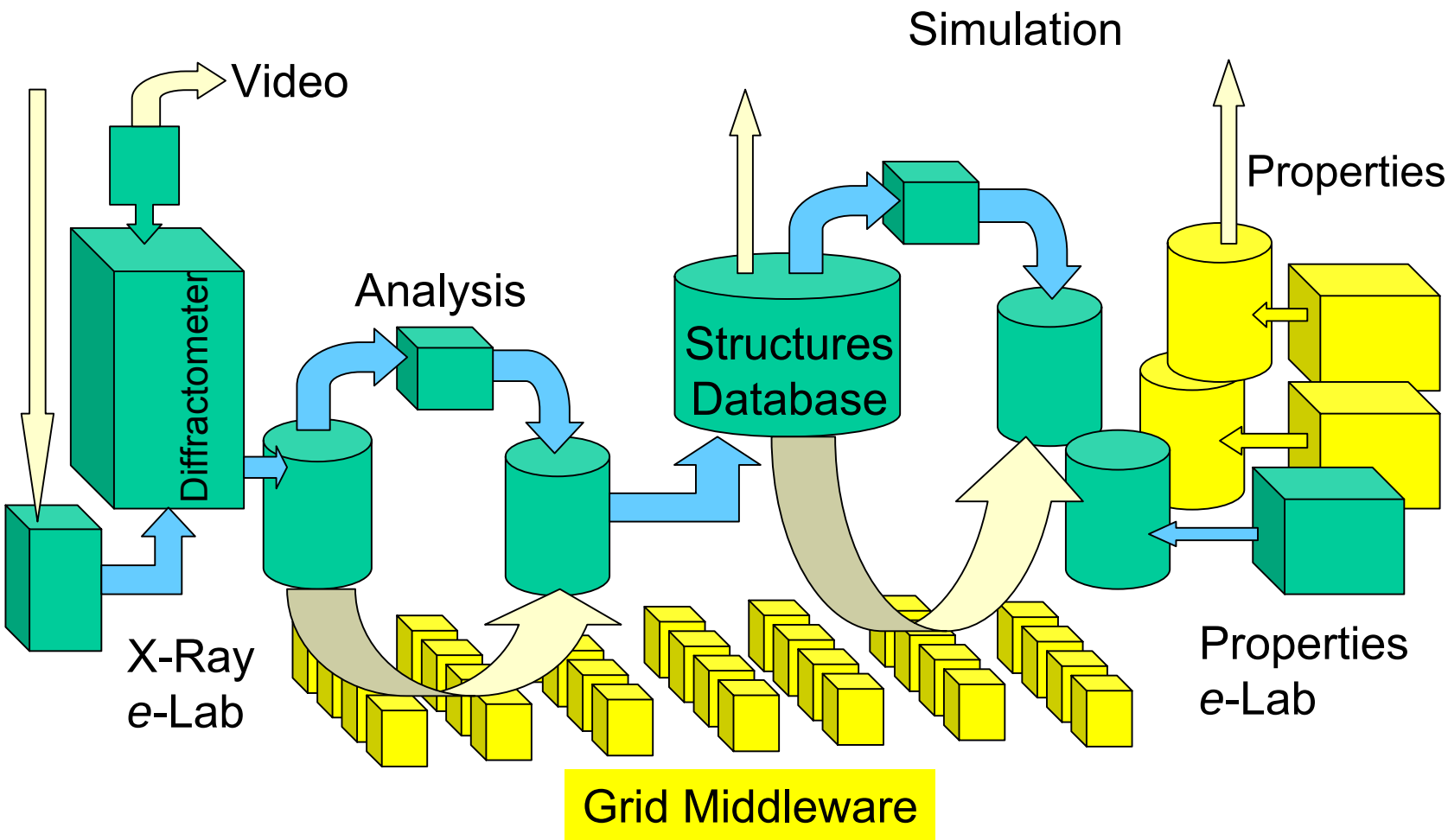
- Publication at source
- Open repositories
- Link data to research literature
- More timely access

UKOLN

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



UKOLN

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

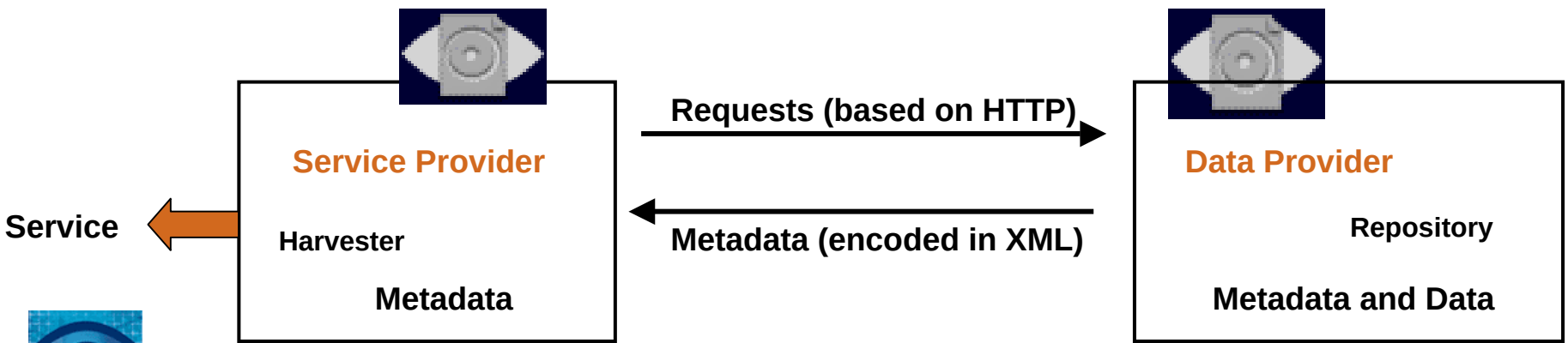
UKOLN

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)



UKOLN

ePrints UK

[home] [documents] [contacts] [workshops] [search] ...subject access to eprint archive

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

UKOLN

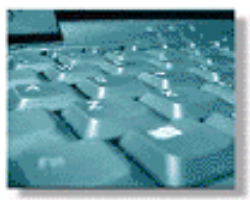
SHERPA ...assisting scholarly communicatio

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

UKOLN

open archives forum

open archives forum

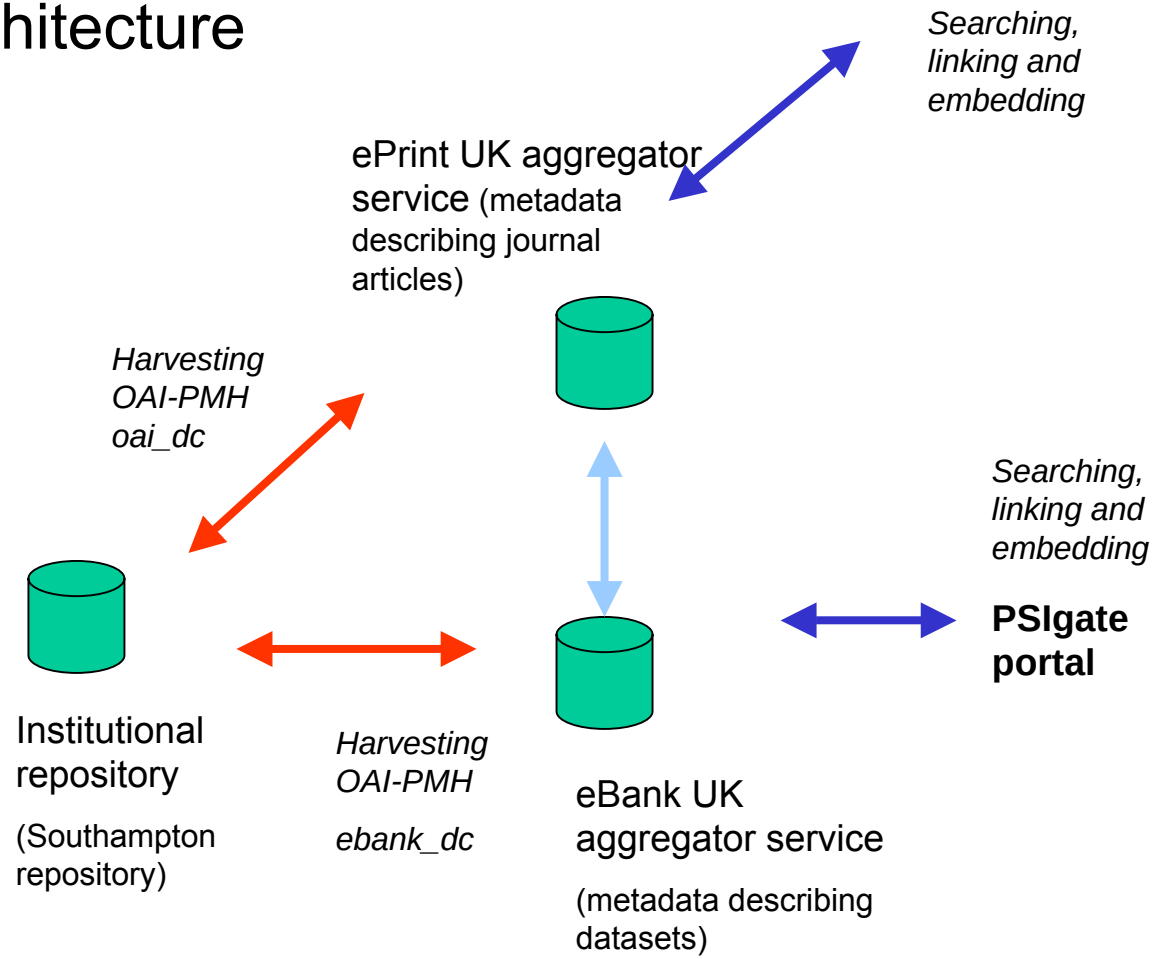
Home

eBank in a nutshell

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

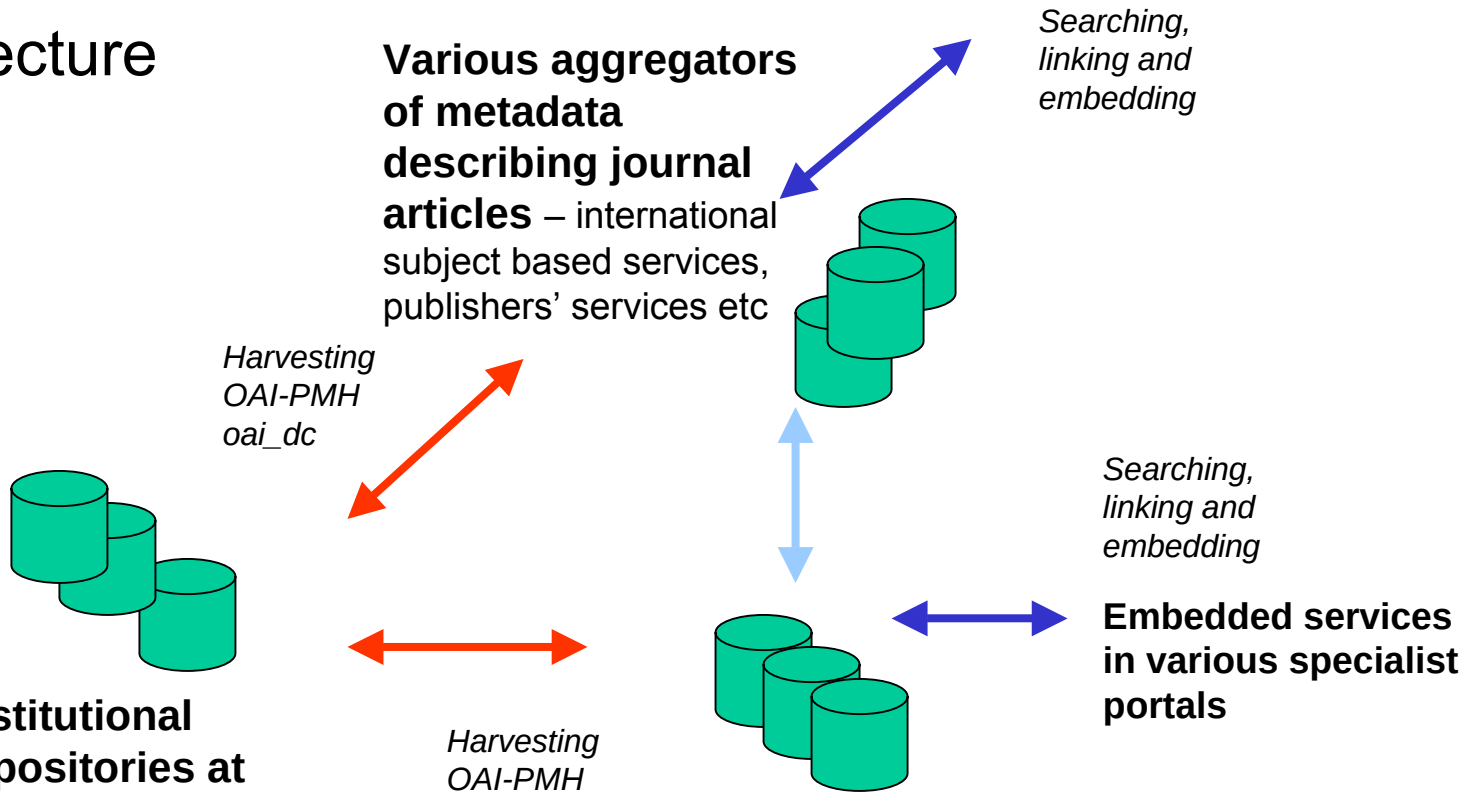
- Create institutional repository of Crystallography Data (at Southampton)
- Modify repository software to handle datasets (eprints.org at Southampton)
- Demonstrate eBank search service linked to ePrints UK, indexing harvested descriptions of datasets and journal articles (at UKOLN)
- Embed eBank service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

eBank architecture



UKOLN

Potential extended architecture

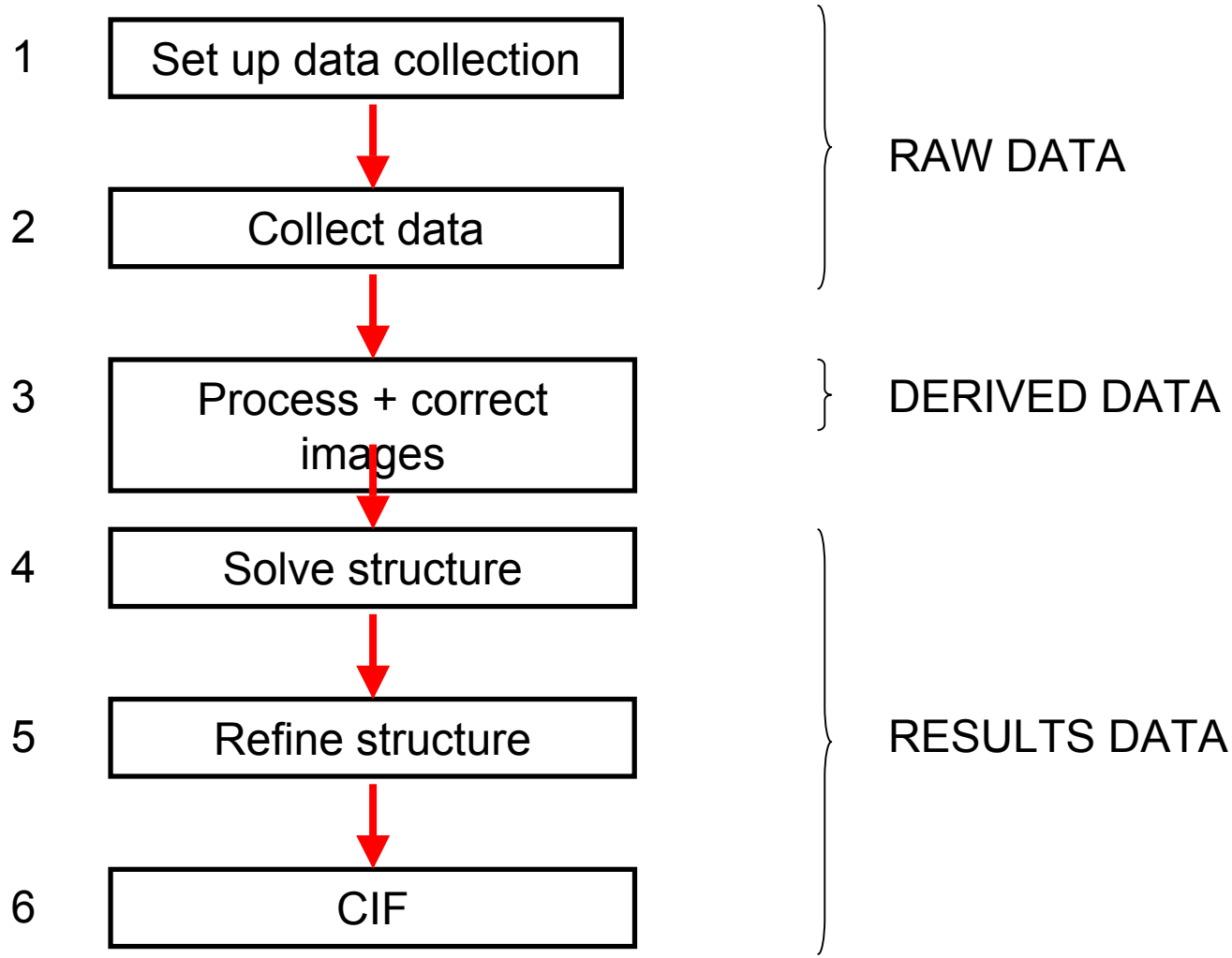


First steps: establishing common ground…

- 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

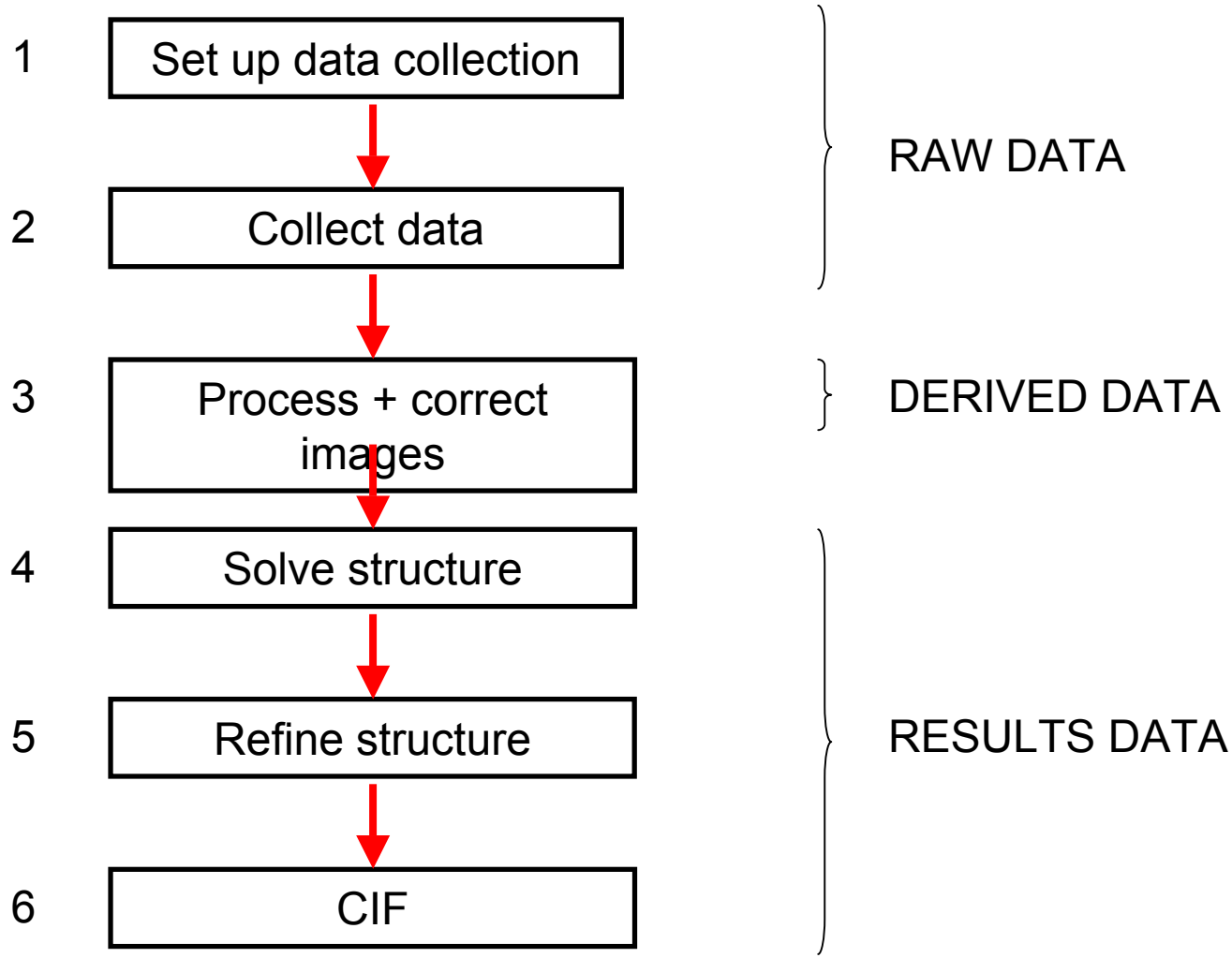
UKOLN

Crystallographic data workflow



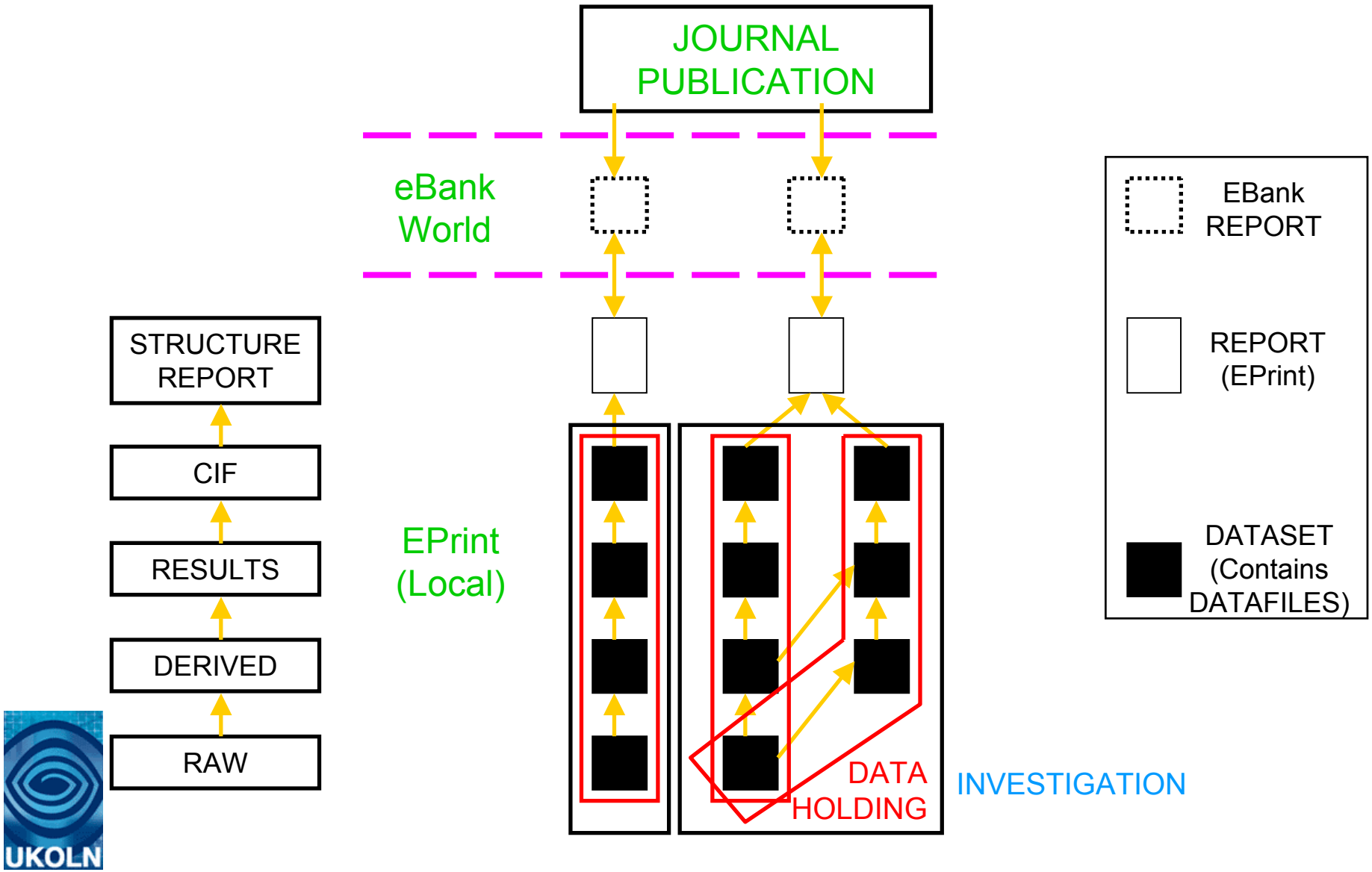
UKOLN

Crystallographic data workflow



UKOLN

Linking Crystallograpy 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	datcol.non i*.kcd .drx	ASCII BINARY ASCII	Parameters for producing i*.kcd files Unit cell determination images Unit cell	Morphology	STRING
Collection	Collect Data	collect.mat datcol.non .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	.prp 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(mu2-4,6-bis((diphenylphosphino)oxy)-5-methyl-1,3-phenylene-C,C',P,P')-tetrakis(mu2-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)

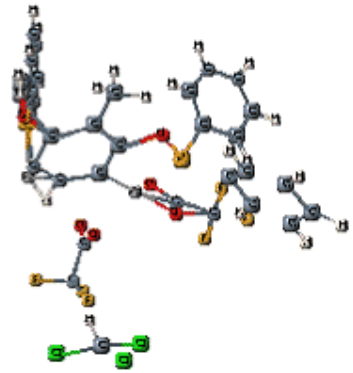
proc

- 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…

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

Available Files

Final Result

02sot064.CIF	19k
02sot064.cml	8k
02sot064_checkcif.html	14k

Refinement

02sot064.RES	9k

Solution

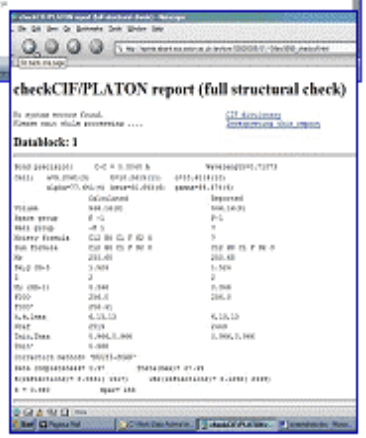
02sot064.PRP	5k

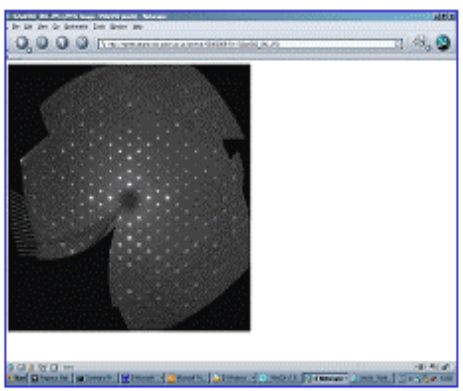
Processing

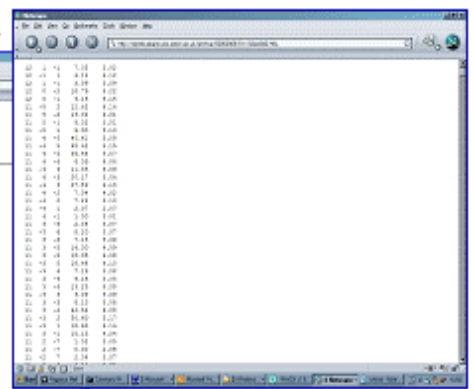
02SOT064.HTM	6k
02sot064.HKL	338k

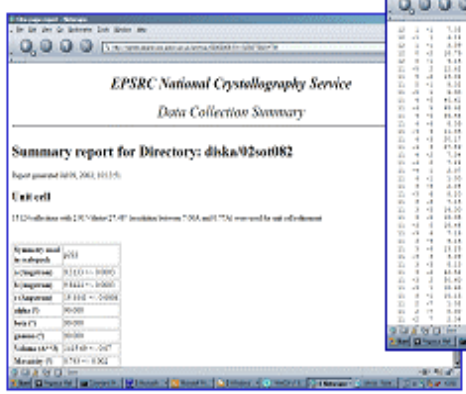
Other Files

02sot064.DOC	113k
02sot064.LST	49k



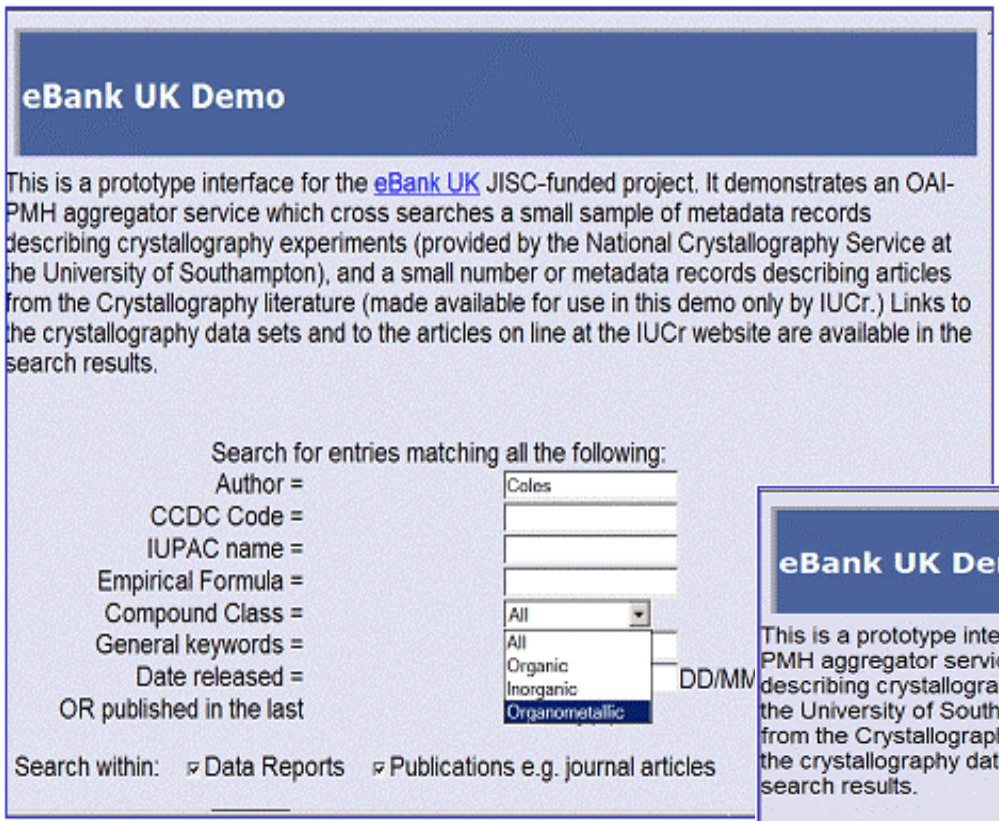


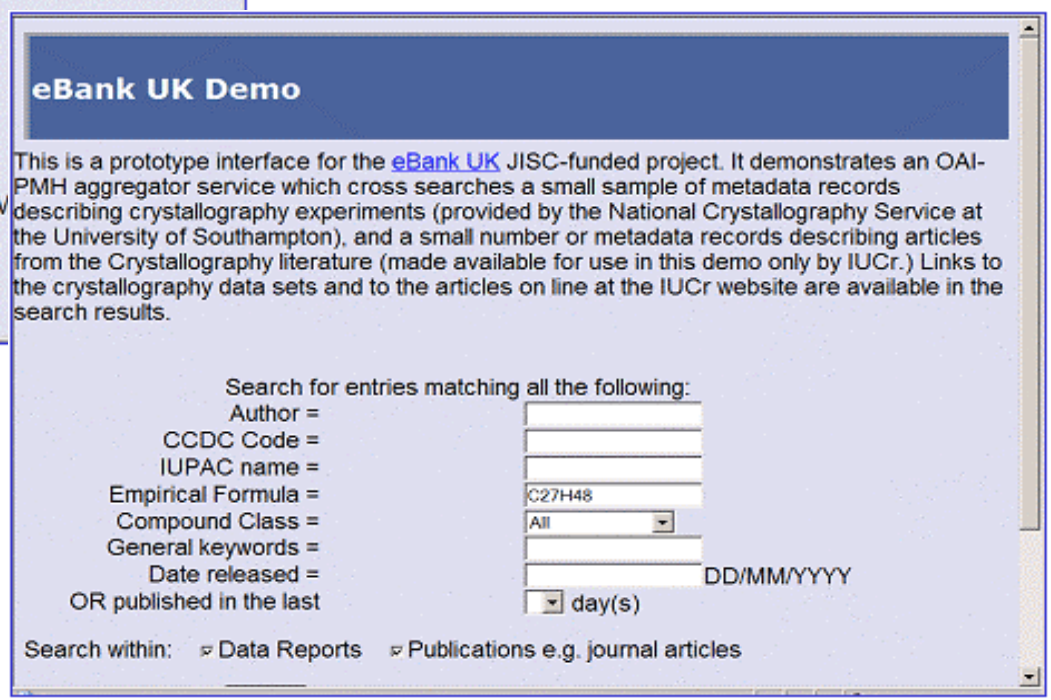




UKOLN

eBank aggregator : search





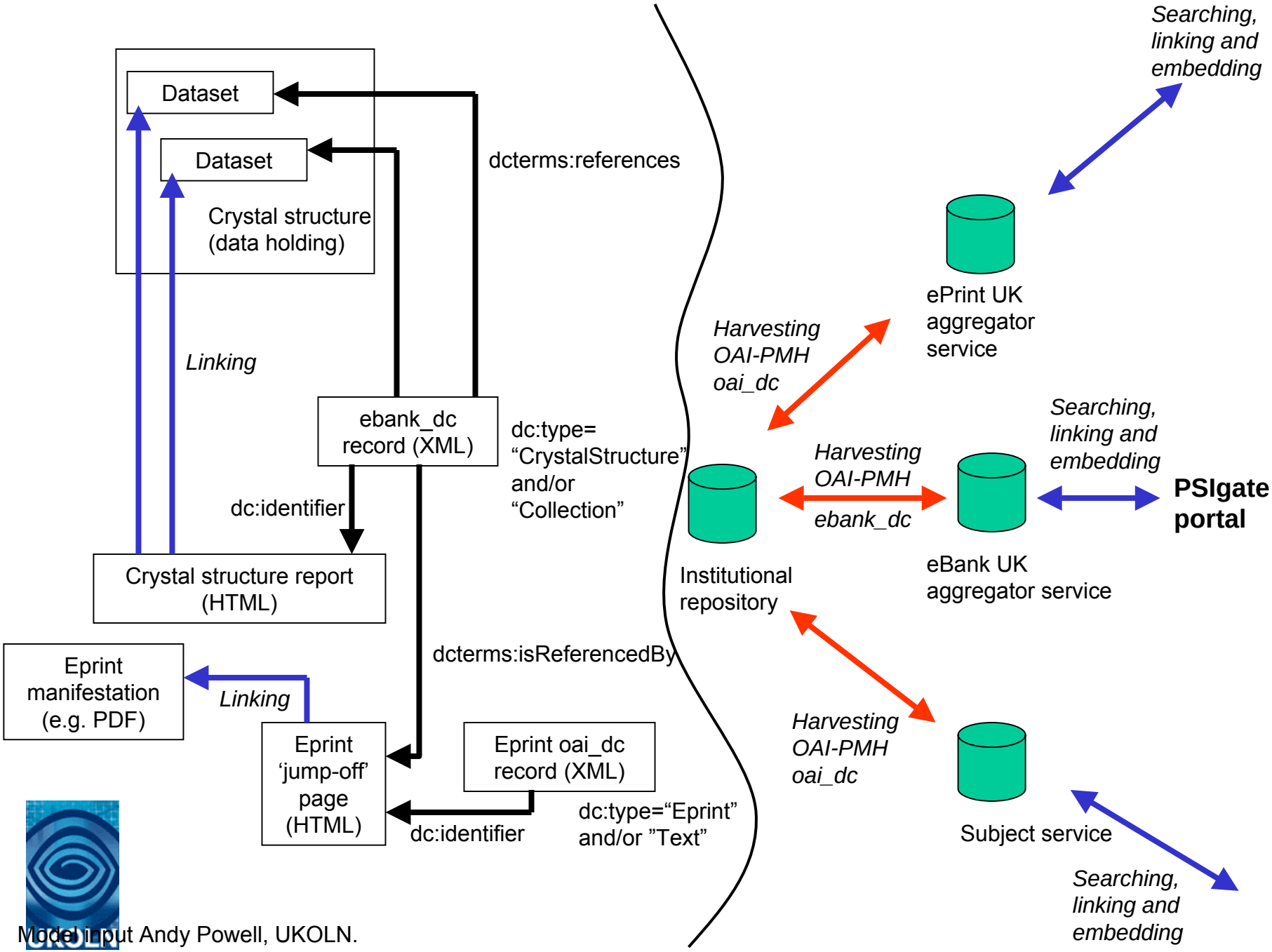
UKOLN

Ebank aggregator: browse

UKOLN

And finally…
eBank search embedded in a science portal

UKOLN



Model input Andy Powell, UKOLN.

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

UKOLN

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)

UKOLN

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

UKOLN

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

UKOLN

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/