Adding value to open access research data: the eBank UK Project.

Dr Liz Lyon, Director

UKOLN, University of Bath, UK

OAI4, CERN Geneva, October 2005.

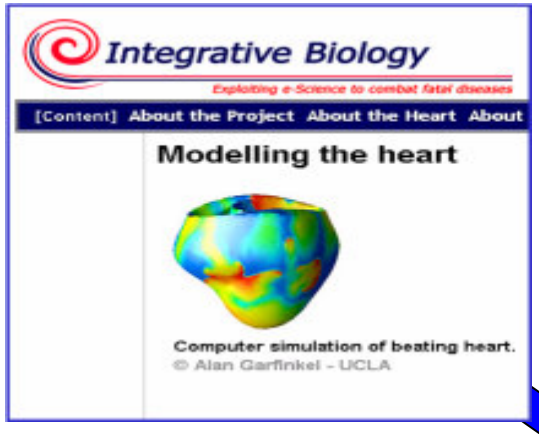
UKOLN is supported by: JISC MLA MUSEUMS LIBRARIES ARCHIVES

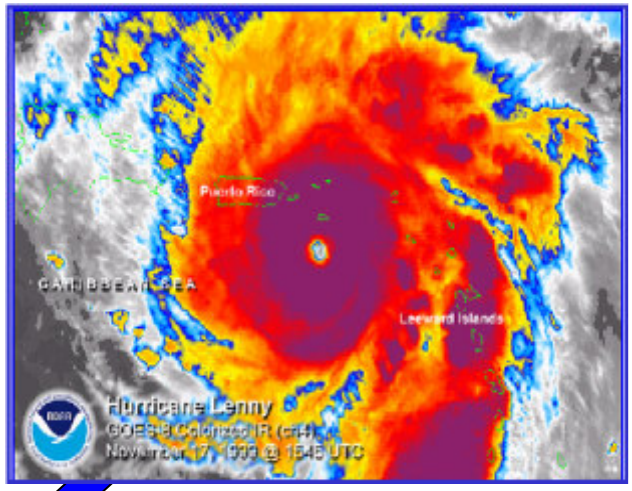
Overview

1. e-Research & data-intensive science
2. Repository services & adding value
   - Aggregation and linking: eBank UK
   - Integration and workflows
3. Looking to the longer term: digital curation and preservation

1. e-Research & data-intensive science

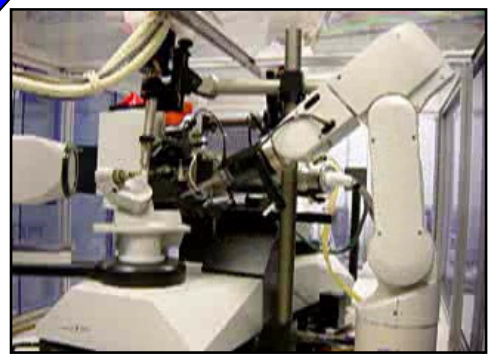
eScience - the data deluge





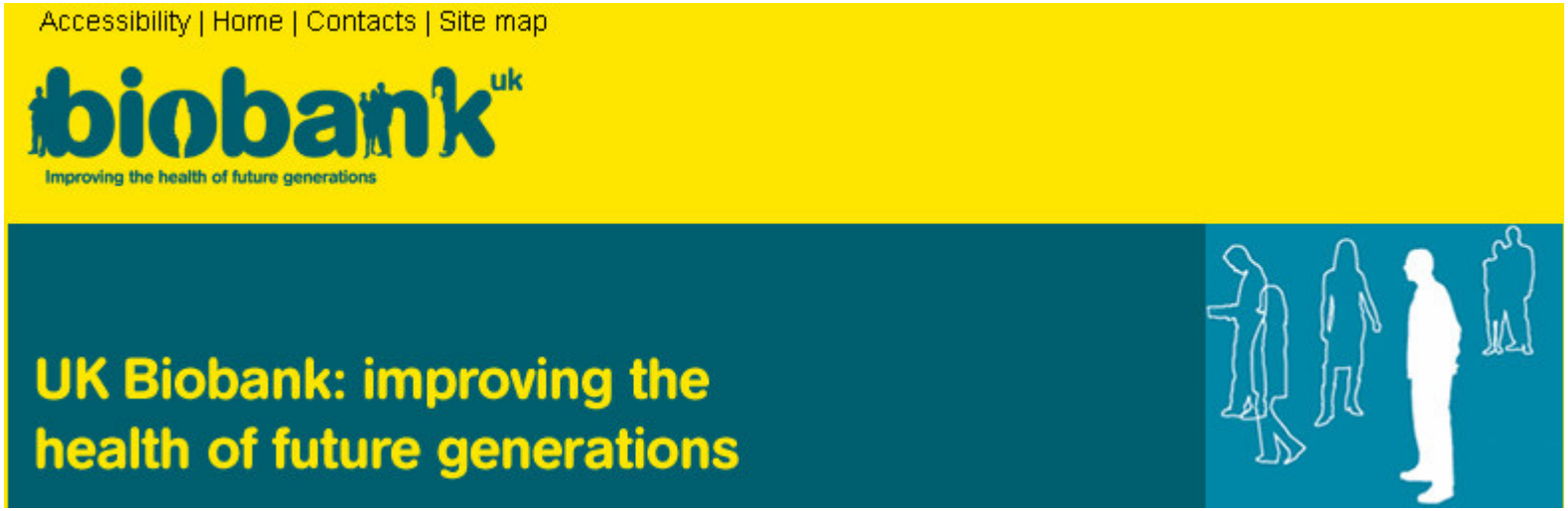
Data Overload!

How do we disseminate?



Diversity of data collections

- Very large, relatively homogeneous:
  Large-scale Hadron Collider (LHC) outputs from CERN
- Smaller, heterogeneous and richer collections:
  World Data Centre for Solar-terrestrial Physics CCLRC
- Small-scale laboratory results:
  "jumping robots" project at the University of Bath
- Population survey data: *UK Biobank*



- Highly sensitive, personal data: *patient care records*

UKOLN

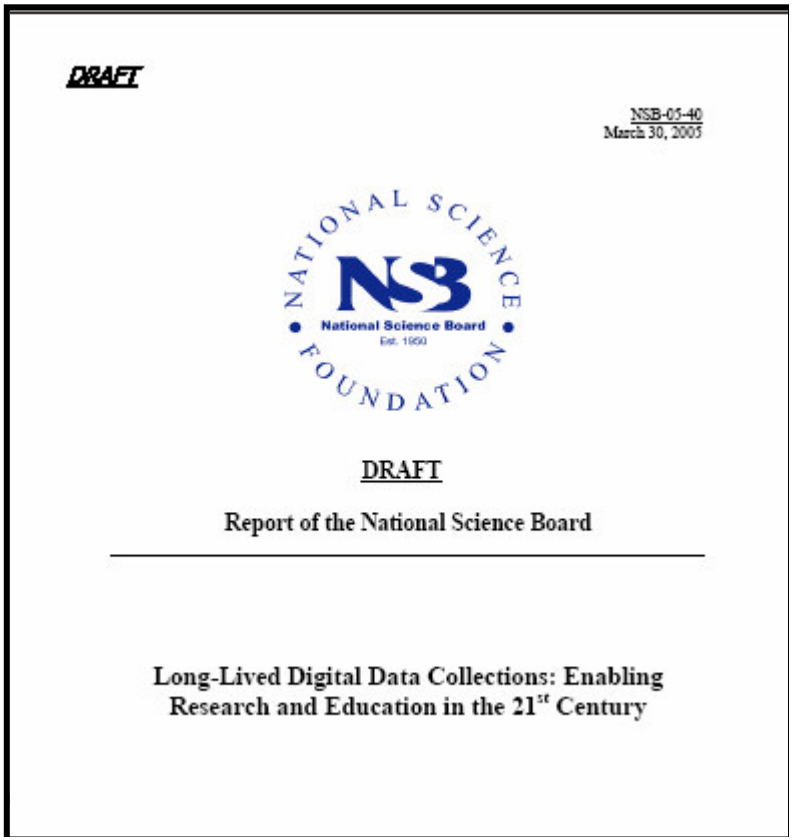
Taxonomy of data collections

- Research collections: *jumping robots*
- Community collections: *Flybase at Indiana (with UC Berkeley)*
- Reference collections: *Protein Data Bank*

Evolution……

FlyBase A Database of the Drosophila Genome

RCSB PDB PROTEIN DATA BANK


DRAFT

NSB-05-40
March 30, 2005

NATIONAL SCIENCE FOUNDATION
NSB National Science Board

DRAFT

Report of the National Science Board

Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century

Source: NSF Long-Lived Digital Data Collections

Draft report revised May 2005

UKOLN

Experience of data-sharing

- Large scale data sharing in the life sciences
  Draft Report June 2005
  Sponsored by UK research funding bodies
  MRC, BBSRC, NERC, JISC, Wellcome
- Outcomes & recommendations
  - Importance of standards and good quality metadata
  - Require a data management plan
  - Work needed on vocabularies & ontologies
  - Awareness of archiving & long term preservation
- Position of research funders and policy makers?

RESEARCH COUNCILS UK

RESEARCH COUNCILS UK Together in research

Home | What's new | Executive Group | About us | Contacts | Activities | Links

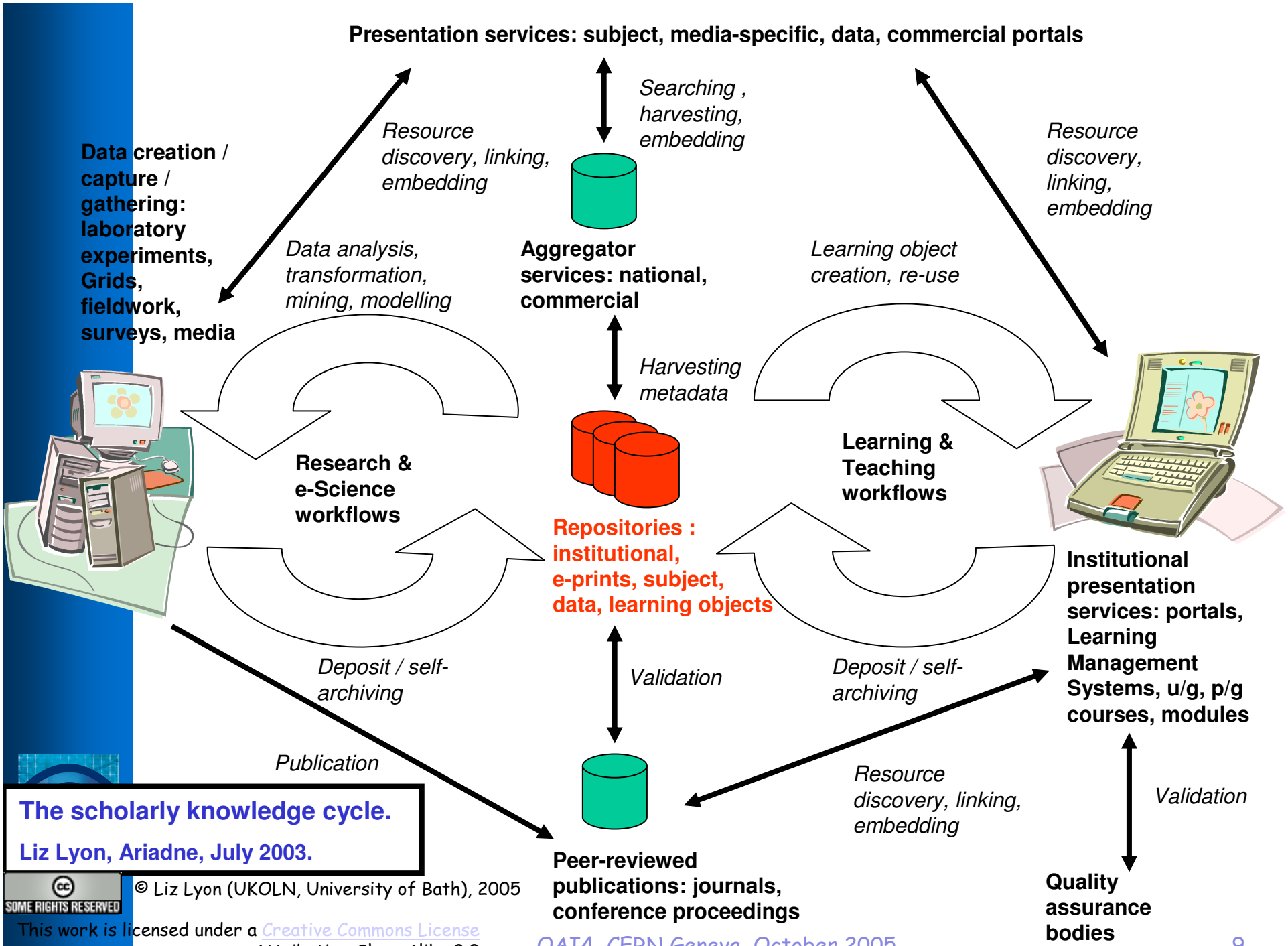
News release

28 June 2005

RCUK Announces Proposed Position on Access to Research Outputs

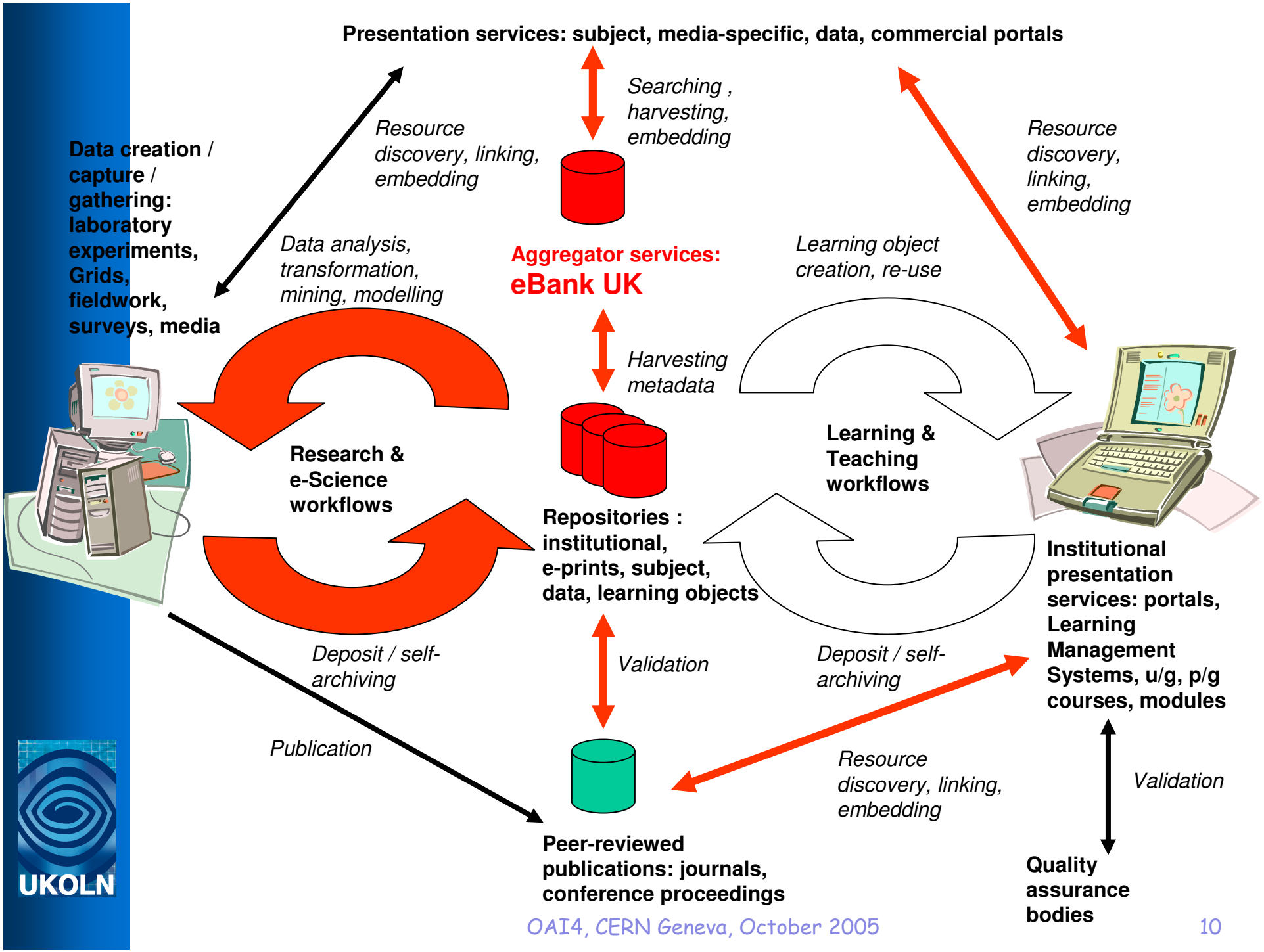
Research Data

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***



The scholarly knowledge cycle.

Liz Lyon, Ariadne, July 2003.


Presentation services: subject, media-specific, data, commercial portals
Searching , harvesting, embedding
Resource discovery, linking, embedding
Resource discovery, linking, embedding
Data creation / capture / gathering: laboratory experiments, Grids, fieldwork, surveys, media
Data analysis, transformation, mining, modelling
Aggregator services:
eBank UK
Learning object creation, re-use
Harvesting metadata
Research & e-Science workflows
Learning & Teaching workflows
Repositories : institutional, e-prints, subject, data, learning objects
Institutional presentation services: portals, Learning Management Systems, u/g, p/g courses, modules
Deposit / self-archiving
Validation
Deposit / self-archiving
Publication
Resource discovery, linking, embedding
Validation
Peer-reviewed publications: journals, conference proceedings
Quality assurance bodies

2. Repository services & adding value: the eBank UK Project

eBank UK Project

- Two key themes:
  - **Open access to datasets**
  - **Linking research data to publications and to learning**
- JISC-funded from September 2003: now in Phase 2
- UKOLN at the University of Bath (lead), University of Southampton, University of Manchester
- Exemplar: e-Science testbed ‘Combechem’
  - Grid-enabled combinatorial chemistry / crystallography
  - National Crystallography Service
- Resource Discovery Network / PSIgate physical sciences portal
- http://www.ukoln.ac.uk/projects/ebank-uk/

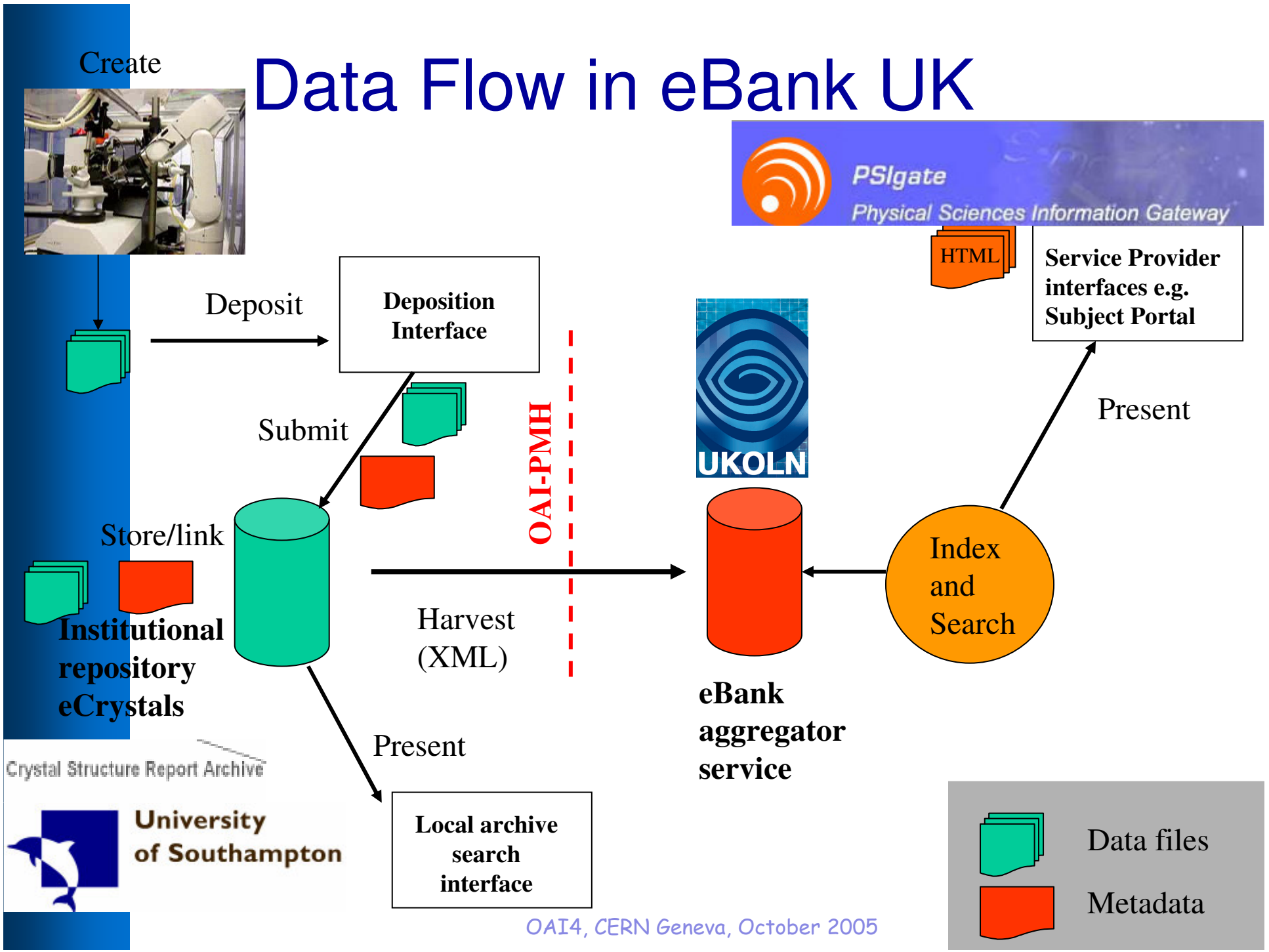
The "hybrid" project team

- **UKOLN**
- Michael Day
- Monica Duke
- Rachel Heery
- Traugott Koch
- Liz Lyon
- +
- Andy Powell

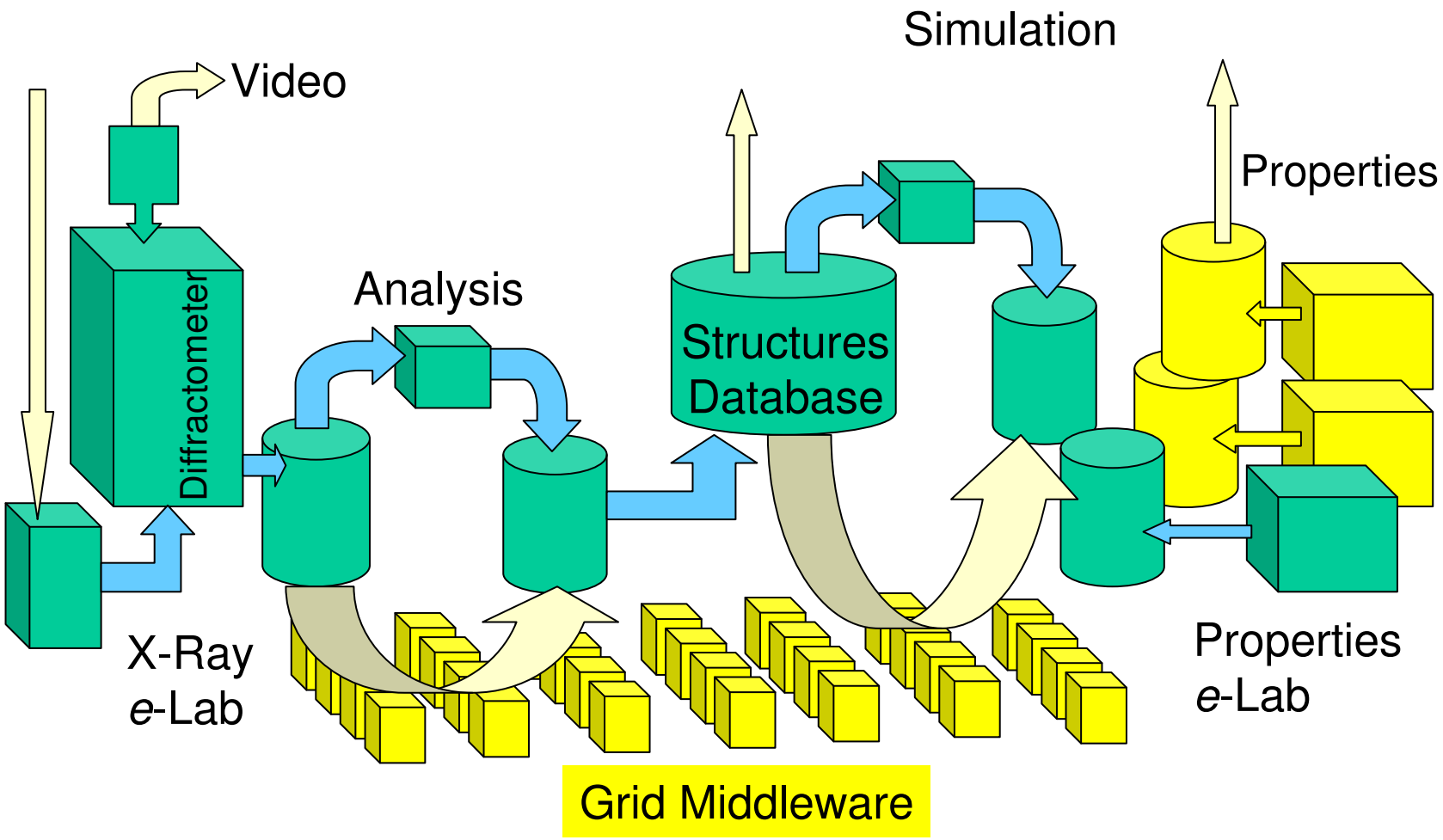
- **Southampton**
- Les Carr
- Simon Coles
- Jeremy Frey
- Chris Gutteridge
- Mike Hursthouse
- Andrew Milstead

- **Manchester**
- John Blunden-Ellis

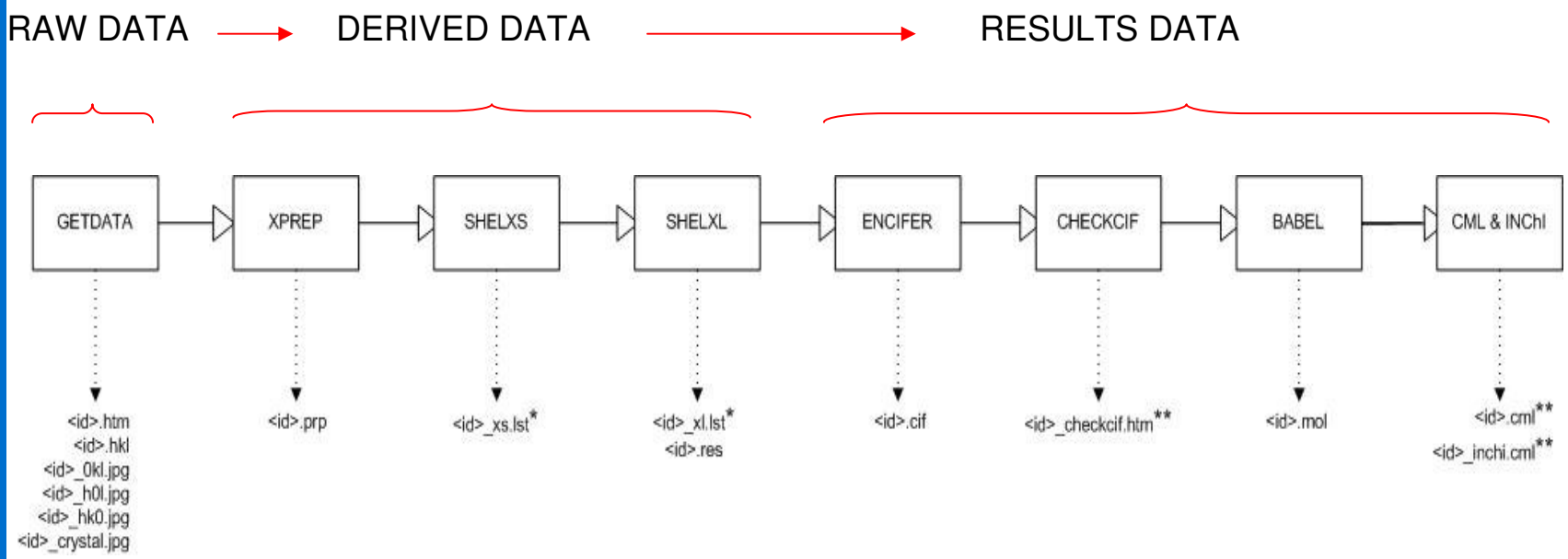
Data Flow in eBank UK



CombeChem: An EPSRC pilot project



Crystallography workflow



- **Initialisation**: mount new sample set up data collection
- **Collection**: collect data
- **Processing**: process and correct images
- **Solution**: solve structures
- **Refinement**: refine structure
- **CIF**: produce CIF (Crystallographic Information File)
- **Validation**: chemical & crystallographic checks
- **Report**: generate Crystal Structure Report

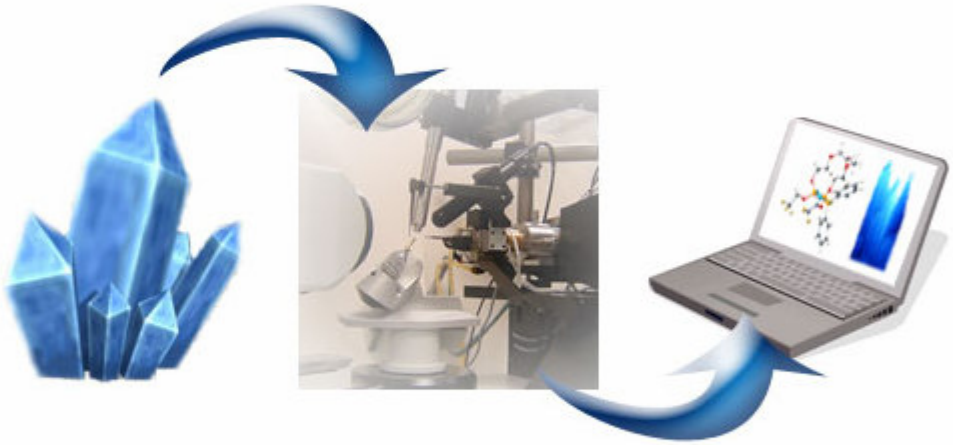
University of Southampton

Crystal Structure Report Archive

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

Welcome to Southampton EBank

Southampton eCrystals is the archive for Crystal Structures generated by the Southampton Chemical Crystallography Group and the EPSRC UK National Crystallography Service.



The information contained within each entry of this archive is all the fundamental and derived data resulting from a single crystal X-ray structure determination, but excluding the raw images. The results have not been externally refereed, but the information supplied should enable any reader to check the reliability and validity directly, since all the files provided are freely available for download. Should any error be detected, we would appreciate receiving suitable comments, and we will make any necessary amendments, and include a note to that effect. Any reader wishing to have access to the raw images is welcome to contact us, and we will make arrangements for these to be made available.

UKOLN

A data repository entry

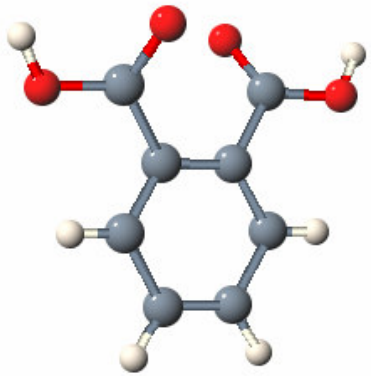
http://ecrystals.chem.soton.ac.uk/149/ Go ecrystals archive

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

UKOLN

Access to the underlying data: complex objects

Available Files

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

Final Result

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

Refinement

02sot064.RES	9k

Solution

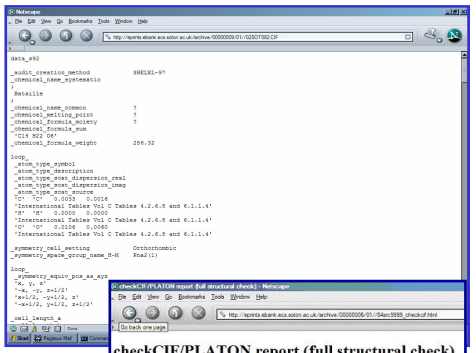
02sot064.PRP	5k

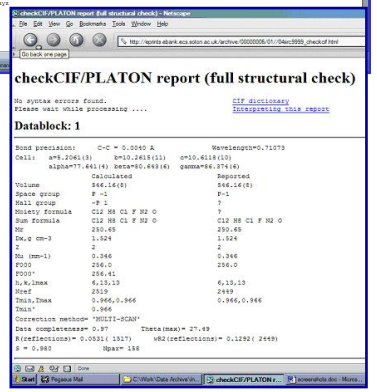
Processing

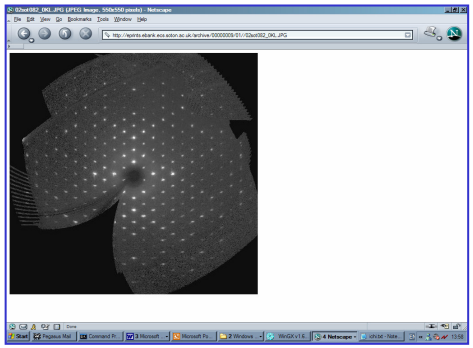
02SOT064.HTM	6k
02sot064.HKL	338k

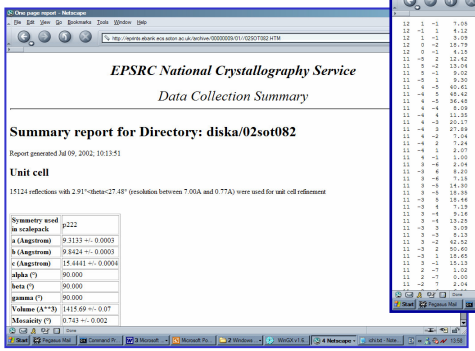
Other Files

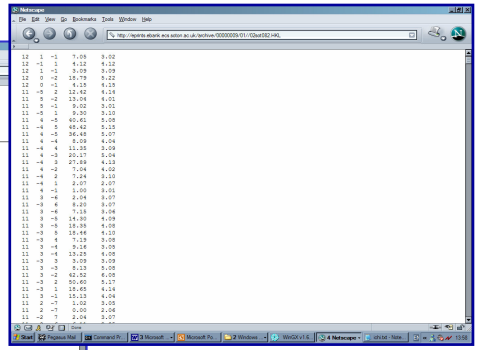
02sot064.DOC	113k
02sot064.LST	49k





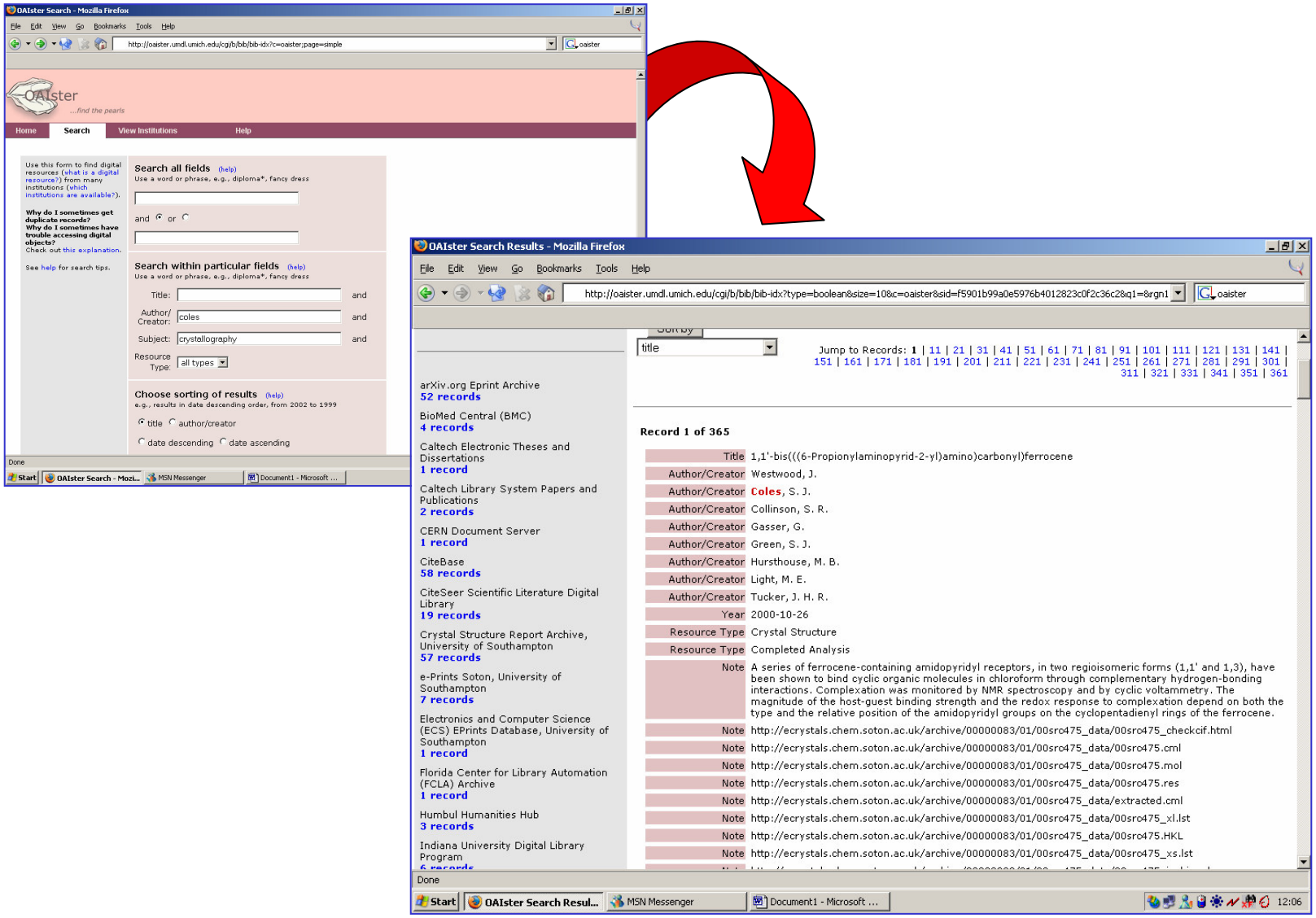




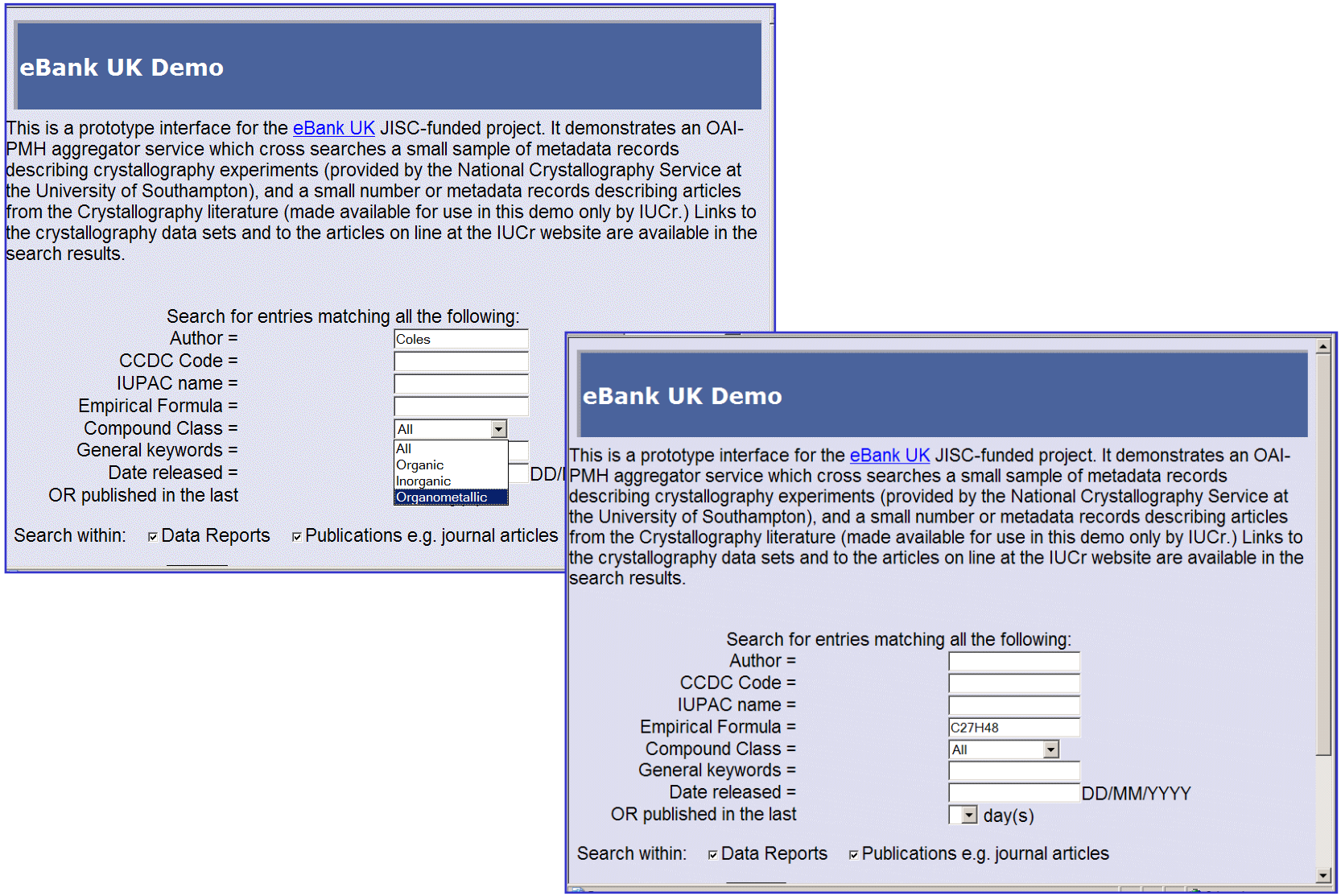


ecrystals.chem.soton.ac.uk

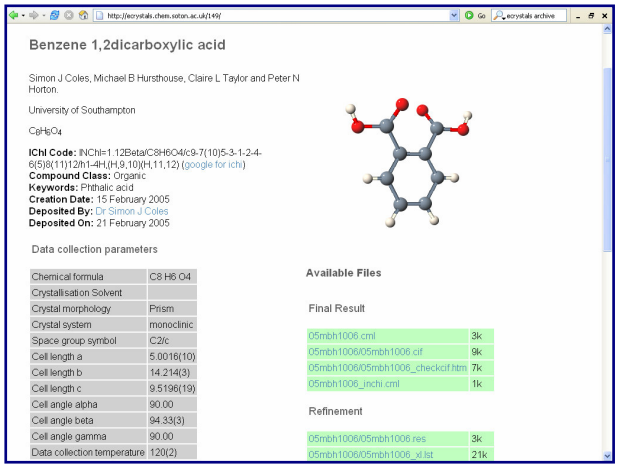
Harvesting: OAIster

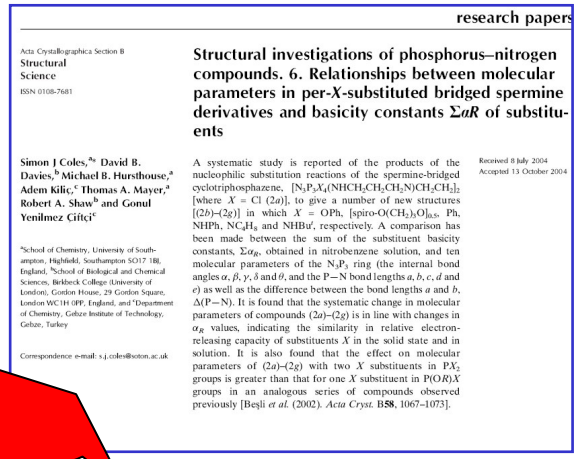


Aggregating: search & discover



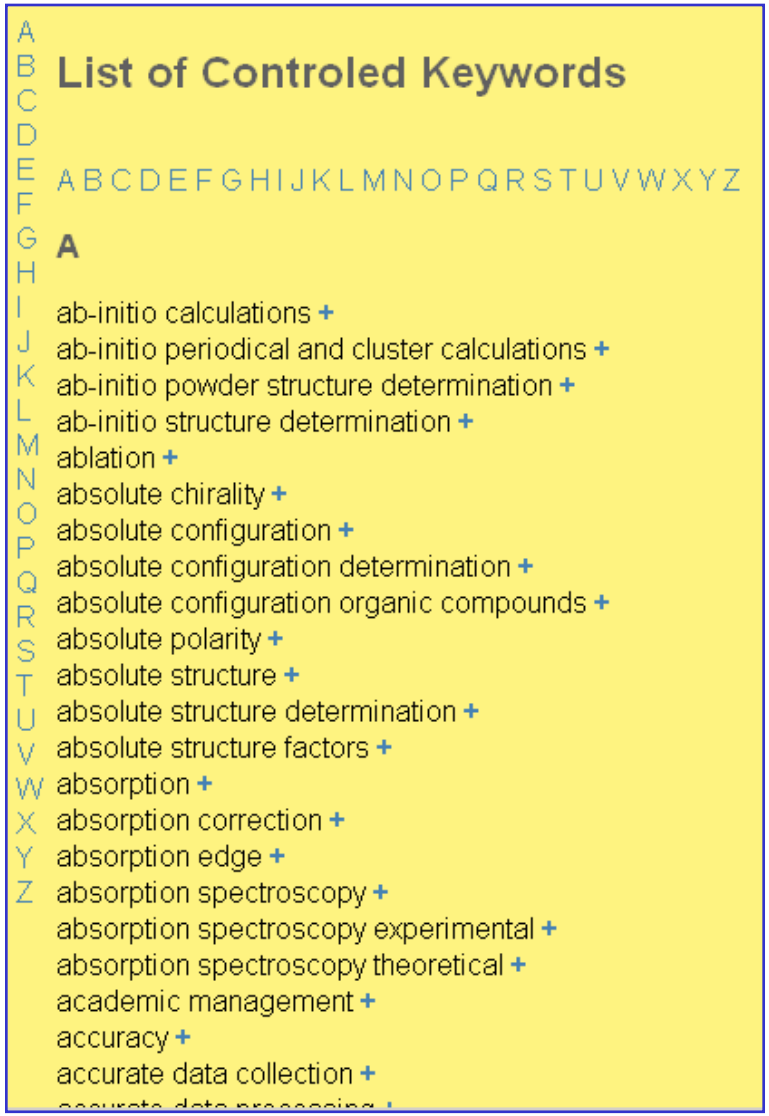
Linking data to publications





Embedding in a science portal for student learners

Ontologies for discovery in an inter-disciplinary world



- Transform the 'list' into an 'ontology'
- Embed ontology into the deposition process
- Aggregators use keywords for linking with the broader literature
- Researchers use keyword ontology in search and discovery services

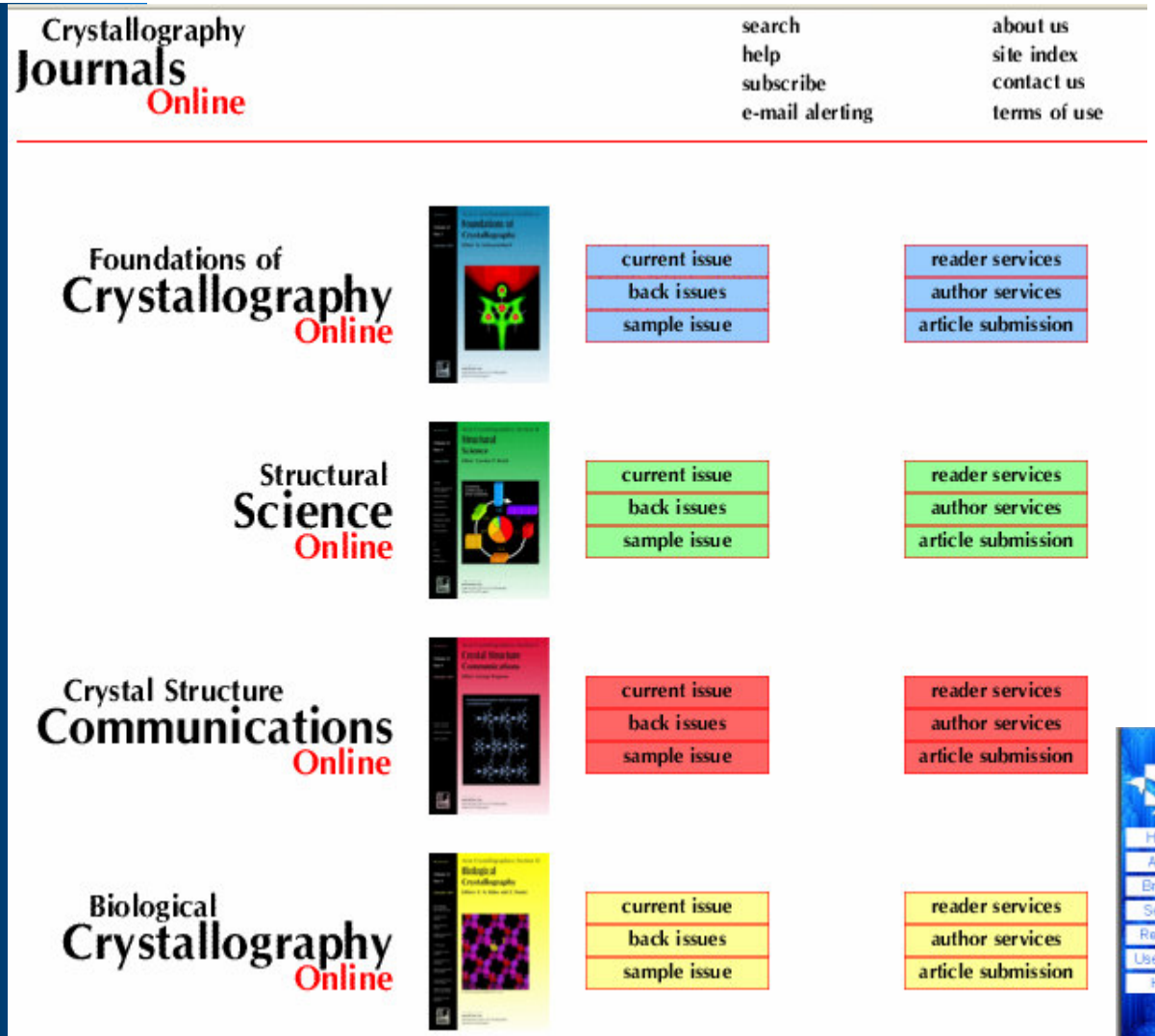
Persistent identifiers for data citation

- eBank use cases: depositor, author, service provider, reader, publisher, ?
- Schemes: DOI, Handle, ARK, PURL
- Global identification: express as http URIs
- Added value services: CrossRef, resolution service, integration (Globus), look-up service, ?
- Degree of trust or persistence
- Costs
- Future potential: political, ?
- Domain identifiers: International Chemical Identifier (InChI) codes

Publication & citation of scientific primary data project

- National Library for Science & Technology (TIB), University of Hanover, Germany
- STD-DOI Project *http://www.std-doi.de*
- DOI registry for datasets
- Data requirements: quality control, long-term curation, use DOI resolver
- Data publication agents: World Data Center Climate, GeoForschungsZentrum Potsdam
- Exemplar data citation:
  - *Kamm, H; Machon, L; Donner, S (2004): Gas chromatography (KTB Field Lab), GFZ Potsdam. doi:10.1594/GFZ/ICDP/KTB/ktb-geoch-gaschr-p*

Integration into crystallographic publishing practices

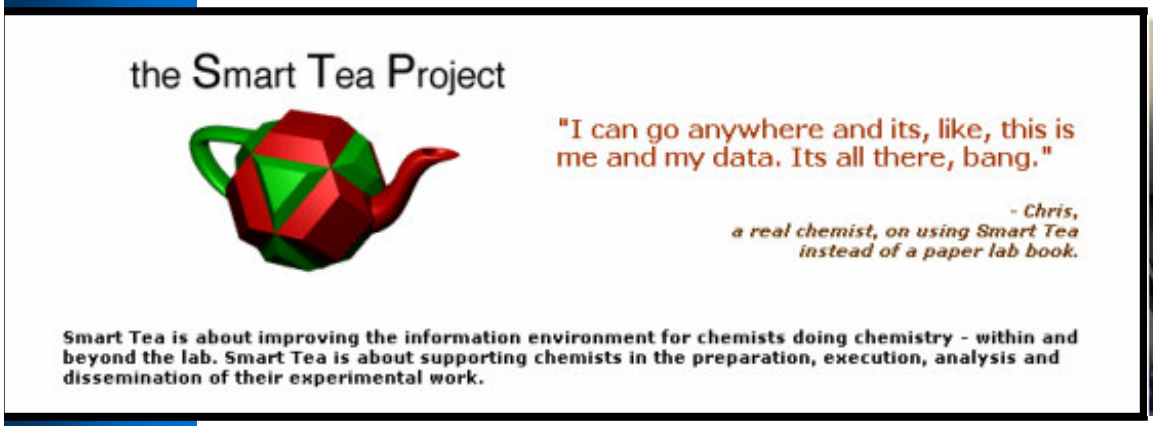


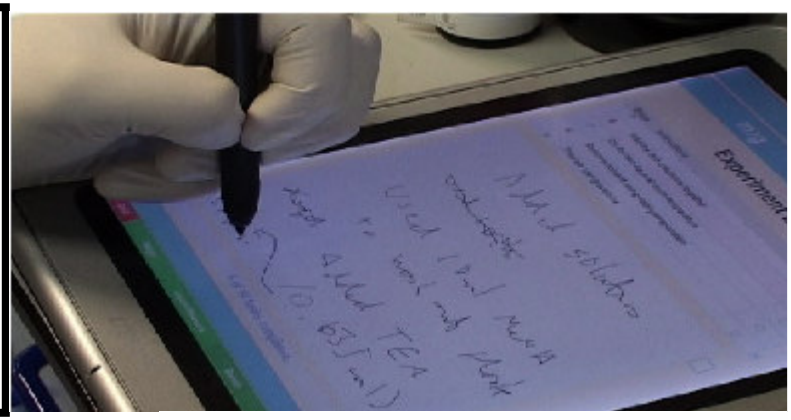
Publishers seal of approval

Integration into chemistry research workflows

R4L Repository for the Laboratory

- R4L Repository for the Laboratory Project (JISC-funded) automated data capture from instrumentation, registration of results
- SMART TEA electronic Laboratory notebook + annotations





the myTea project

- Related sub-domains of chemistry: SPECTRa Project (JISC-funded)
- Research assessment (RAE) process?

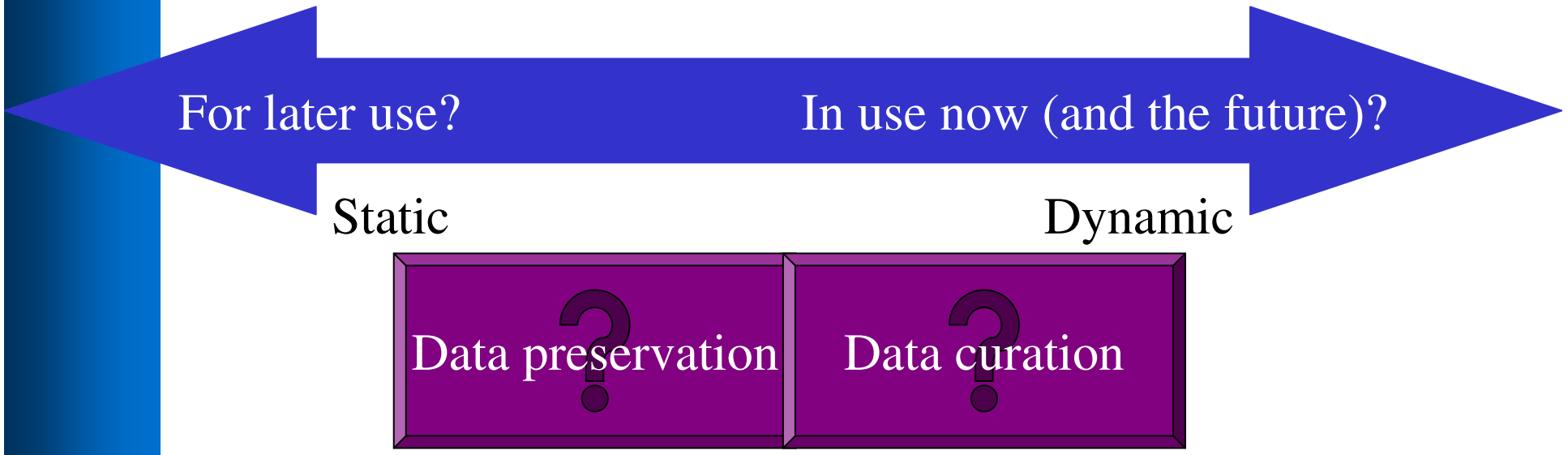
UKOLN

Integration into the curriculum and e-Learning workflows

- MChem course
- Assess role in Undergraduate Chemical Informatics courses
- Pedagogic evaluation

3. Looking to the longer term: digital curation & preservation

Repositories and digital curation



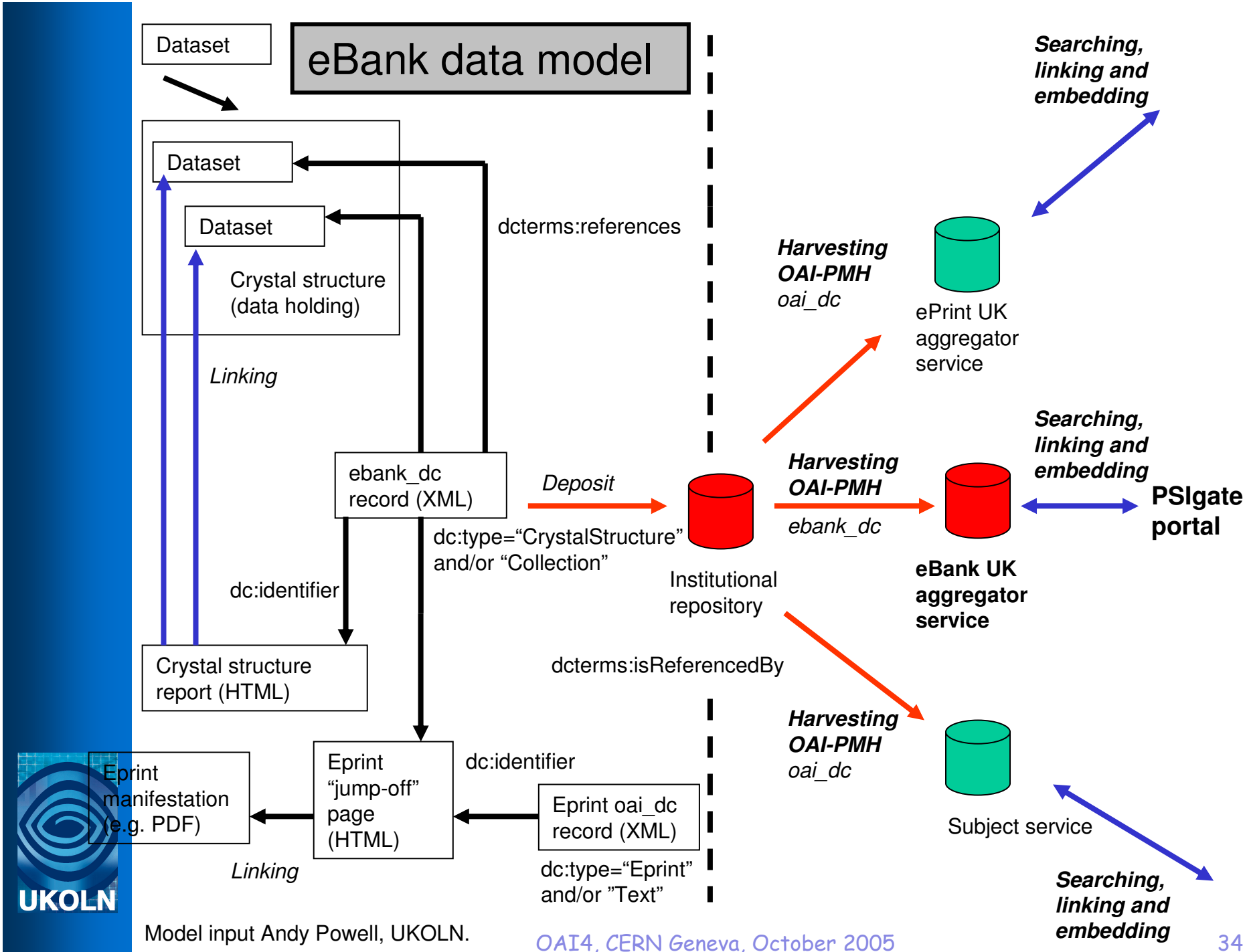
"maintaining and adding value to a trusted body of digital information for current and future use"

Assuring long term access to the research record

- Trusted digital repositories
  - Audit Checklist for Certification Draft Report
  - Research Libraries Group, August 2005
  - RLG-NARA Taskforce
  - Defined criteria under 4 categories
    - Organisation
    - Functions, processes & procedures
    - Designated community & usability
    - Technologies & technical infrastructure
- UK Digital Curation Centre *http://www.dcc.ac.uk*
  - 1st International DCC Conference presentations available
  - PV2005 Royal Society Edinburgh November 21-23 Nov

Thank you.
Questions?…..

More information: UKOLN *http://www.ukoln.ac.uk/*



Model input Andy Powell, UKOLN.