

**Adding value to open access research data: reflections on the process of data curation**

Dr Liz Lyon,  
DCC Associate Director Outreach  
Director, UKOLN, University of Bath, UK

3<sup>rd</sup> European Conference on Research Infrastructures

Funded by JISC

Digital | Curation | Centre

### What is digital curation?

For later use? In use now (and the future)?

Static Dynamic

Data preservation Data curation

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

Digital | Curation | Centre

### (Very simple) e-Research Cycle and Data Curation

*(New) knowledge extraction: data mining, modelling, analysis, synthesis*

*Formulate hypothesis / ideas, test, experiment, observe: data creation, collection & capture*

*Adding value: Data linking, annotation, visualisation, simulation*

*Data management storage & validation: description, deposit, self-archiving, preservation, certification*

*Scholarly communications: data disclosure, publication, citation, discovery, re-use*

**e-Infrastructure  
Open access  
Collaboration**

Digital | Curation | Centre

### (Very simple) e-Research Cycle and Data Curation

*(New) knowledge extraction: data mining, modelling, analysis, synthesis*

*Formulate hypothesis / ideas, test, experiment, observe: data creation, collection & capture*

*Adding value: Data linking, annotation, visualisation, simulation*

*Data management storage & validation: description, deposit, self-archiving, preservation, certification*

*Scholarly communications: data disclosure, publication, citation, discovery, re-use*

**e-Infrastructure  
Open access  
Collaboration**

Digital | Curation | Centre

### Curation issues 1: Data capture & integration into research workflows

**R4L** Repository for the Laboratory

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

the myTea project

Digital | Curation | Centre

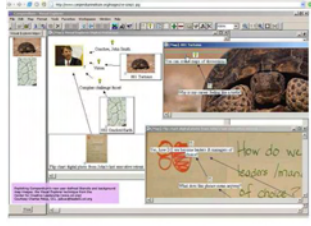
Access Grid  
Collaborative telematic art  
Modify spaces for performers  
Interplay: Hallucinations

### Art on the Grid

Arctic Region Supercomputing Center

HPC-UK

### Human discourse : supporting "persistent conversations"?

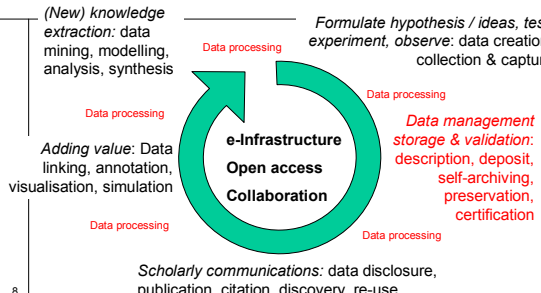


- MEMETIC Project
- JISC-funded
- Virtual Research Environments Programme
- Compendium software + Access Grid

7

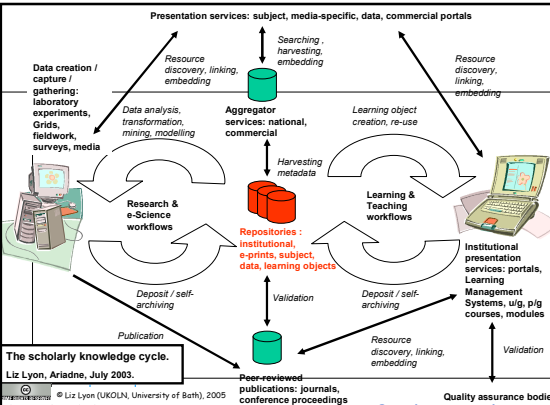
Digital | Curation | Centre

### (Very simple) e-Research Cycle and Data Curation



8

Digital | Curation | Centre



The scholarly knowledge cycle.  
Liz Lyon, Ariadne, July 2003.

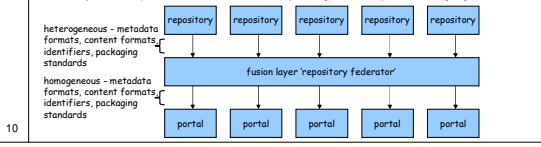
© Liz Lyon (UKOLN, University of Bath), 2005  
This work is licensed under a Creative Commons License Attribution-ShareAlike 2.0

### Federated repository architectures & repository services

- Global
- Inter-disciplinary
- Cross-sectoral
- Multiple format types

- Data, eprints, images.....
- e-Framework: JISC & DEST
- Defining common services + domain-specific services

From Andy Powell: <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/presentations/jisc-jcs-2005/>




10

Digital | Curation | Centre

### eBank UK Project

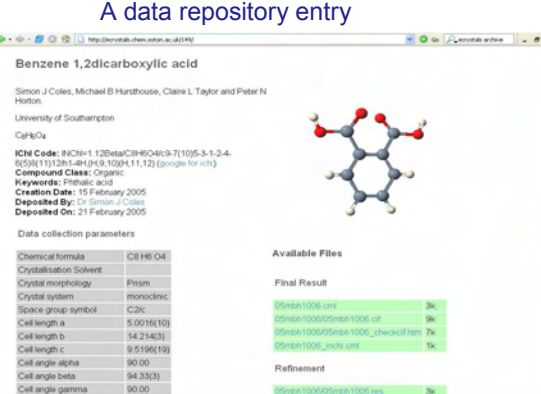
<http://www.ukoln.ac.uk/projects/ebank-uk/>

- Two key themes:
  - Open access to datasets
  - Linking research data to publications and to learning
- UKOLN, University of Southampton, University of Manchester
- e-Science application 'Combechem': Grid-enabled combinatorial chemistry + National Crystallography Service
- Resource Discovery Network / PSLgate physical sciences portal



11

### A data repository entry



Benzene 1,2dicarboxylic acid

Simon J. Cole, Michael B. Hursthouse, Claire L. Taylor and Peter N. Horton  
University of Southampton  
Cdk/Cs

IChI Code: #C8H6O4=C1=125DataC8H6O4=C9-71105-3-1-2-4-059(11)12H-4H(JH,9,10)H,11,12 (google for ichi)

Compound Class: Organic  
Keywords: Dicarboxylic acid  
Creation Date: 15 February 2005  
Deposited By: Simon J. Cole  
Deposited On: 21 February 2005

Data collection parameters

Chemical formula	C8H6O4
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	
05mbr1006.cml	3k
05mbr100605mbr1006.cif	5k
05mbr100605mbr1006_checkcif.htm	7k
05mbr1006_sch.cml	1k
Refinement	
05mbr100605mbr1006.res	3k
05mbr100605mbr1006.sad	21k

### Access to the underlying data: complex objects

13

[ecrystals.chem.soton.ac.uk](http://ecrystals.chem.soton.ac.uk)

### Curation issues 2: describing data

- Validation, publication & discovery of data models & schema
- Managing complex objects
- Metadata packaging standards
  - METS
  - MPEG 21 DIDL
- Semantic descriptions
  - Formal controlled vocabularies
  - High-level and domain ontologies
  - Inter-disciplinary discovery
- Informal approaches Web 2.0 “folksonomies”

List of Controlled Keywords

A  
B  
C  
D  
E  
F  
G  
H  
I  
J  
K  
L  
M  
N  
O  
P  
Q  
R  
S  
T  
U  
V  
W  
X  
Y  
Z

abs-inhio calculations +  
abs-inhio periodical and cluster calculations +  
abs-inhio powder structure determination +  
abs-inhio structure determination +  
absorption +  
absorption chirality +  
absorption configuration +  
absorption configuration determination +  
absorption configuration organic compounds +  
absorption polarity +  
absorption structure +  
absorption structure determination +  
absorption structure factors +  
absorption +  
absorption correction +  
absorption edge +  
absorption spectroscopy +  
absorption spectroscopy experimental +  
absorption spectroscopy theoretical +  
academic management +  
accuracy +  
accurate data collection +

14

[del.icio.us](http://del.icio.us)  
social bookmarks

### JISC PALS Dictate project

15

### (Very simple) e-Research Cycle and Data Curation

16

Digital | Curation | Centre

### Curation issues 3: Persistent identifiers for data citation

- Identify 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
- Domain identifiers: e.g. International Chemical Identifier (InChI) codes
- Google molecules using InChIs demo: Peter Murray-Rust, Uni Cambridge

17

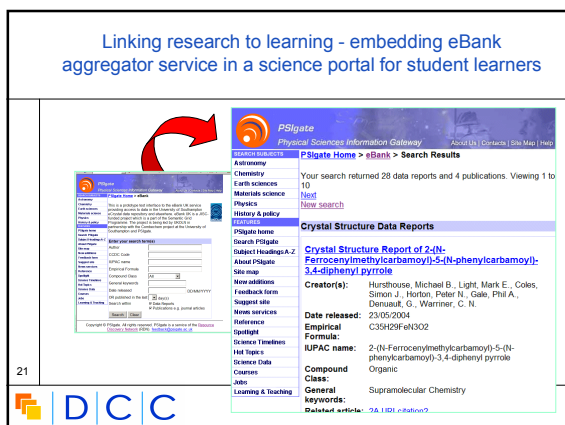
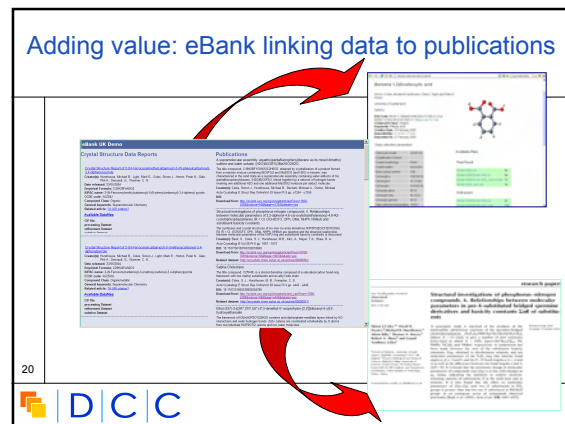
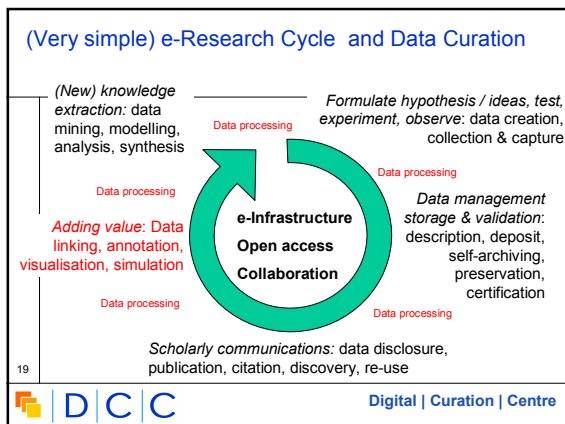
Digital | Curation | Centre

### One approach to data citation using DOIs

- 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 publication agents: World Data Center Climate, GeoForschungsZentrum Potsdam
- Data requirements: quality control, long-term curation, use DOI resolver
- 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

18

Digital | Curation | Centre



### DCC Digital Curation Centre

- Delivering services
- Development activities
- Research agenda
- Outreach Programme

<http://www.dcc.ac.uk/>

DCC

### Adding value through annotation

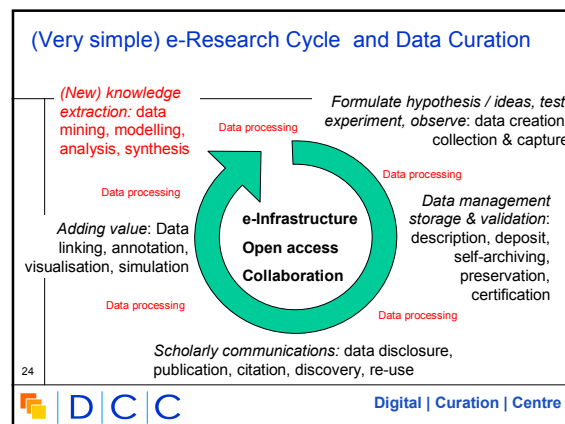
#### DCC Research Agenda at the University of Edinburgh

- Databases: Annotation scoping report
- AstroDAS distributed annotation servers
- New annotation model + prototype: top-ranked demonstration at recent DB conference

Annotation graphic:

22

DCC



**Modelling the Effects of Global Warming on Hurricane Frequency and Intensity Using Remote Sensing Data**

**Integrative Biology**  
Exploiting e-science to combat fatal diseases  
[Content] About the Project About the Heart About

**Modelling the heart**

Computer simulation of beating heart.  
© Alan Garfinkel - UCLA

**Sloan Digital Sky Survey**

**Curation issues 5: workforce development, capacity building & achieving cultural change**

- DCC Outreach & Services:
  - [HELPDESK@dcc.ac.uk](mailto:HELPDESK@dcc.ac.uk) (legal - technical guidance)
  - Curation Manual
  - Workshops, Information Days
  - 2<sup>nd</sup> International Conference November 2006
- NSF Report : "Data scientist"
- Develop hybrid skills
- Embed in u/g, p/g curriculum
- *Facilitate collaboration: researchers, data centres, digital libraries & archives communities*

D | C | C

Thank you.

[e.lyon@ukoln.ac.uk](mailto:e.lyon@ukoln.ac.uk)  
Join the DCC Associates Network at  
[www.dcc.ac.uk](http://www.dcc.ac.uk)

D | C | C Digital | Curation | Centre