Evolution or revolution? The changing data landscape

Dr Liz Lyon, Associate Director, UK Digital Curation Centre
Director, UKOLN, University of Bath, UK

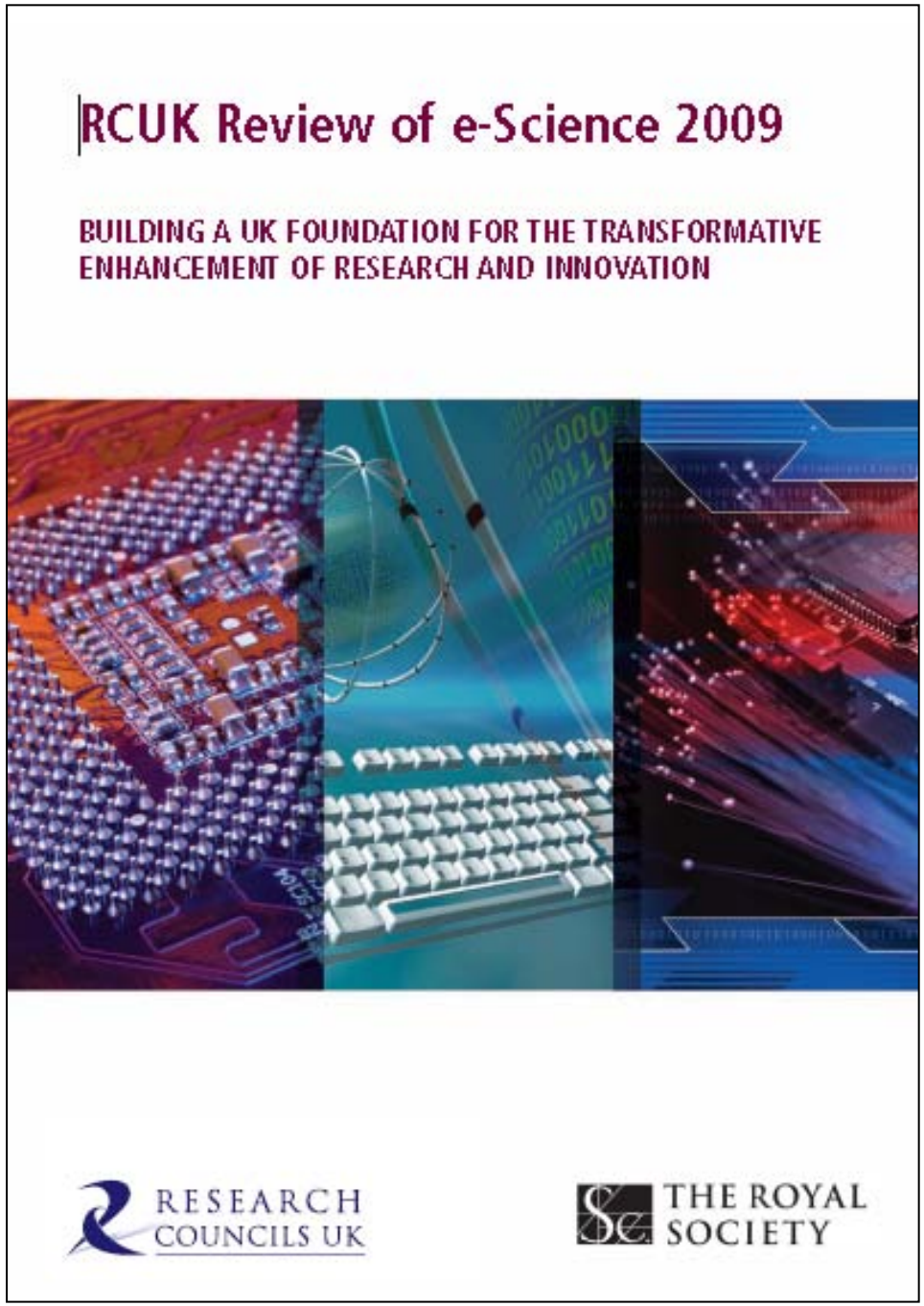
2nd DCC Regional Roadshow, Sheffield, March 2011

UKOLN is supported by: JISC DCC

www.ukoln.ac.uk

UNIVERSITY OF BATH

A centre of expertise in digital information management



"Data sets are becoming the new instruments of science"

Dan Atkins, Univ Michigan

Digital data as the new special collections?

Sayeed Choudhury, Johns Hopkins

Roman de la Rose: Lutenist and singers in a garden
British Library Harley MS 4425, f.14v

Research data : institutional crown jewels?

guardian.co.uk

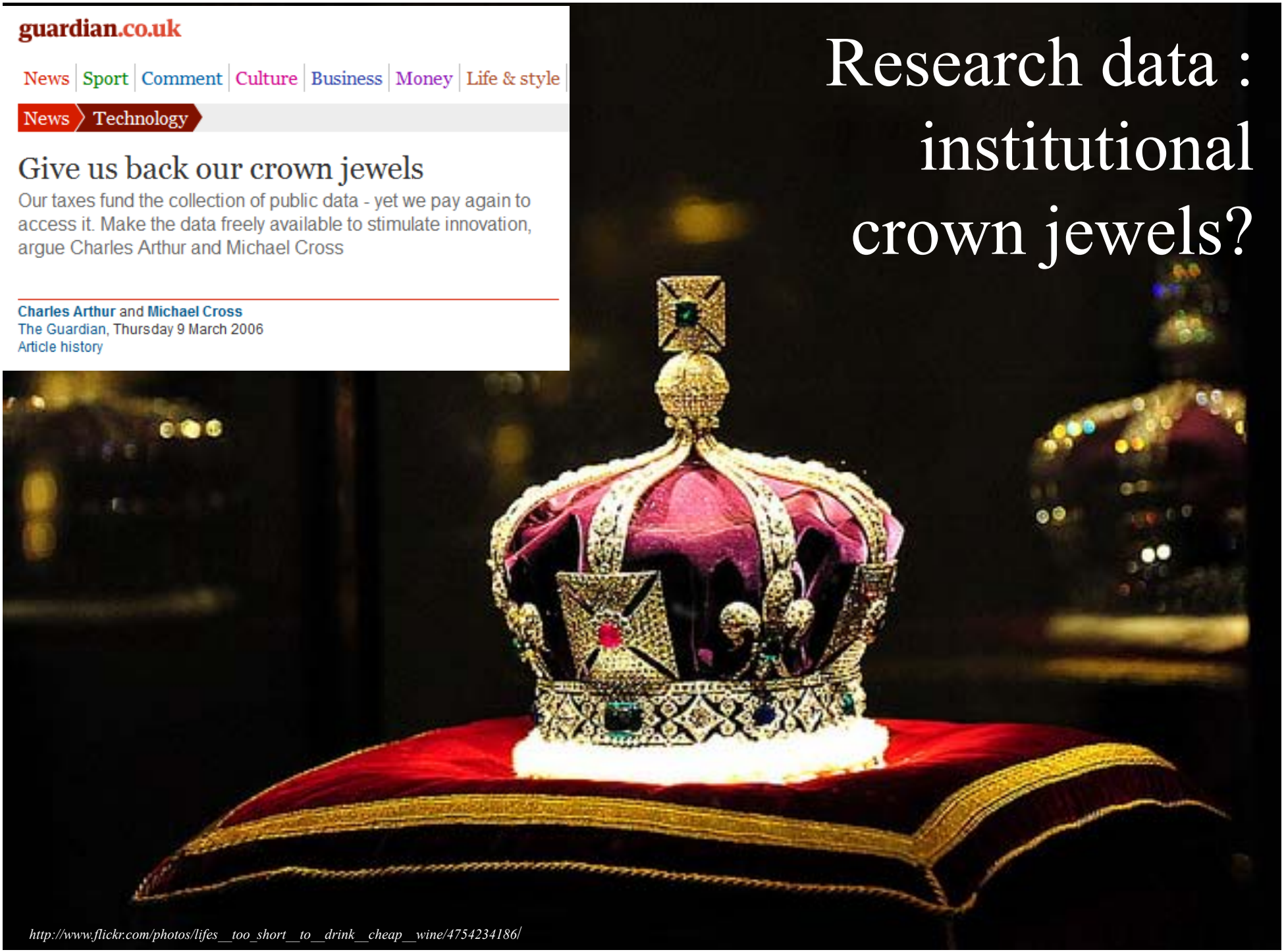
News | Sport | Comment | Culture | Business | Money | Life & style

News › Technology

Give us back our crown jewels

Our taxes fund the collection of public data - yet we pay again to access it. Make the data freely available to stimulate innovation, argue Charles Arthur and Michael Cross

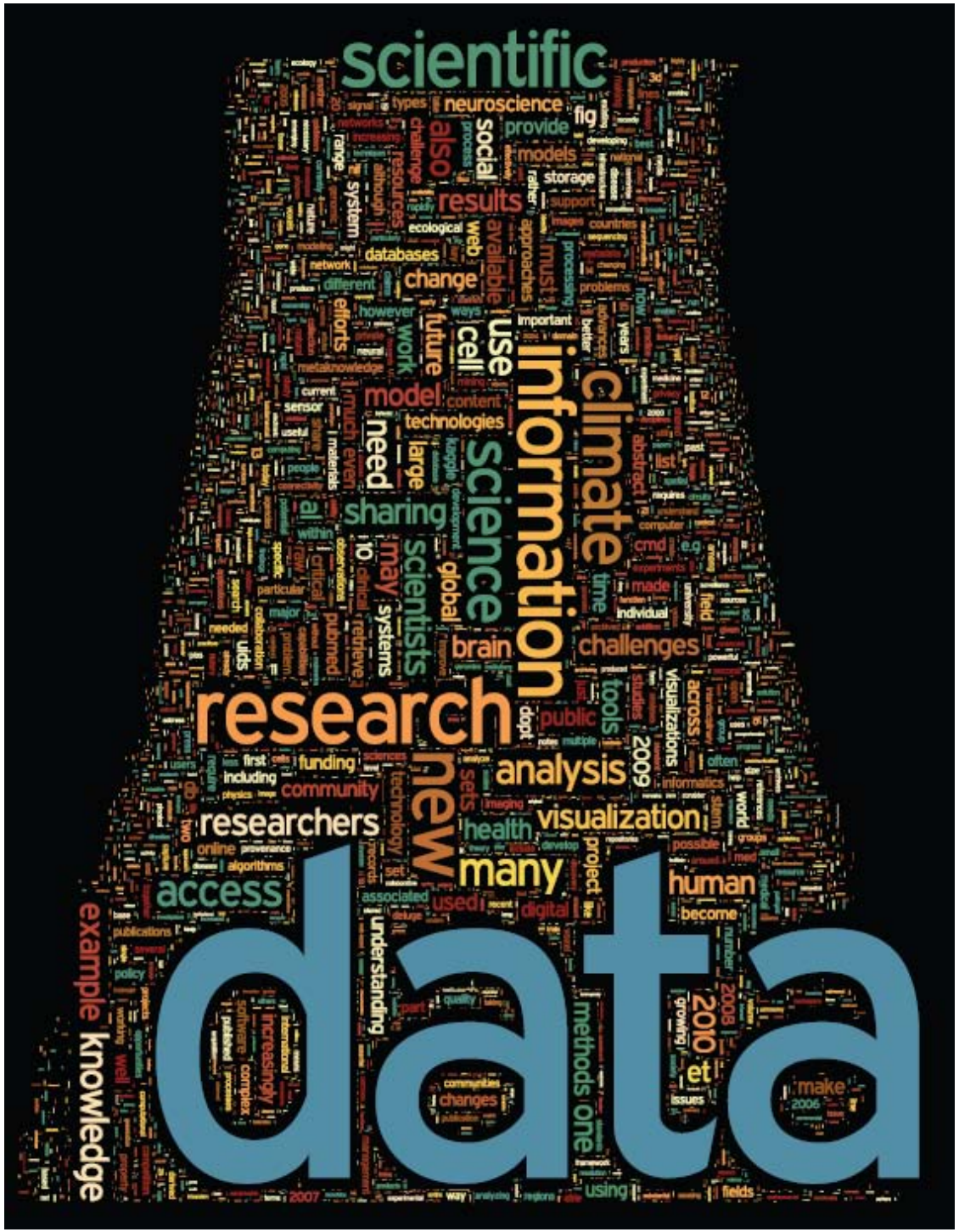
Charles Arthur and **Michael Cross**
The Guardian, Thursday 9 March 2006
Article history



http://www.flickr.com/photos/lifes__too_short__to__drink__cheap__wine/4754234186/

Perspectives

- Environmental scan
  - Scale and complexity
  - Infrastructure
  - Open science
- Policy
  - Funders
  - Institutions
  - Ethics & IP
- Practice Challenges
  - Storage
  - Incentives
  - Costs & Sustainability



“Surfing the Tsunami”

Science: 11 February 2011

"The costs of sequencing DNA has taken a nosedive...and is now dropping by 50% every 5 months".

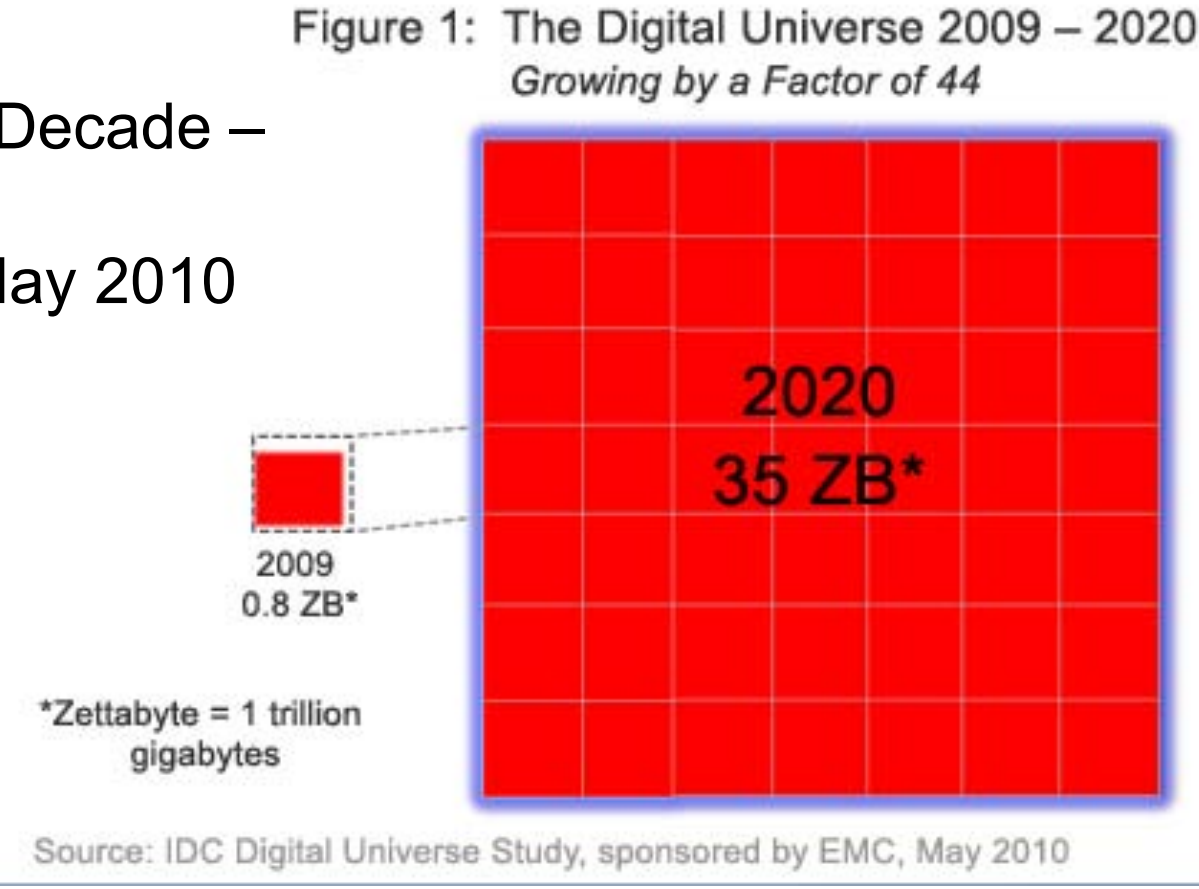
"A single sequencer can now generate in a day what it took 10 years to collect for the Human Genome Project".

"The 1000 Genomes Project generated more DNA sequence data in its first 6 months than GenBank had accumulated in its entire 21 year existence".

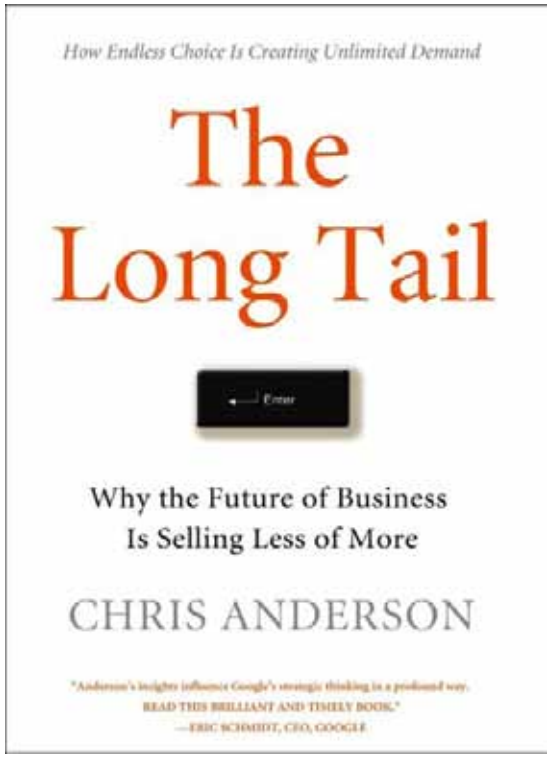
"I worry there won't be enough people around to do the analysis."
Chris Ponting, University of Oxford

“the amount of data generated worldwide...is growing by 58% per year; in 2010 the world generated 1250 billion gigabytes of data”

The Digital Universe Decade – Are You Ready?
IDCC White Paper, May 2010



Data collections



GenBank

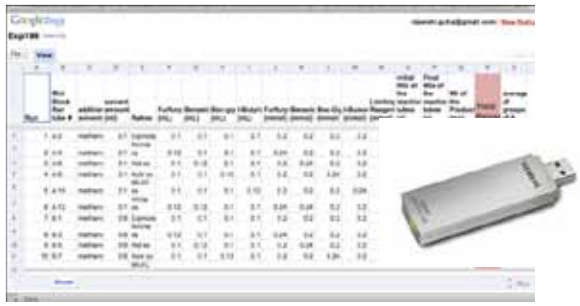
PDB

UniProt

Pfam

ChemSpider

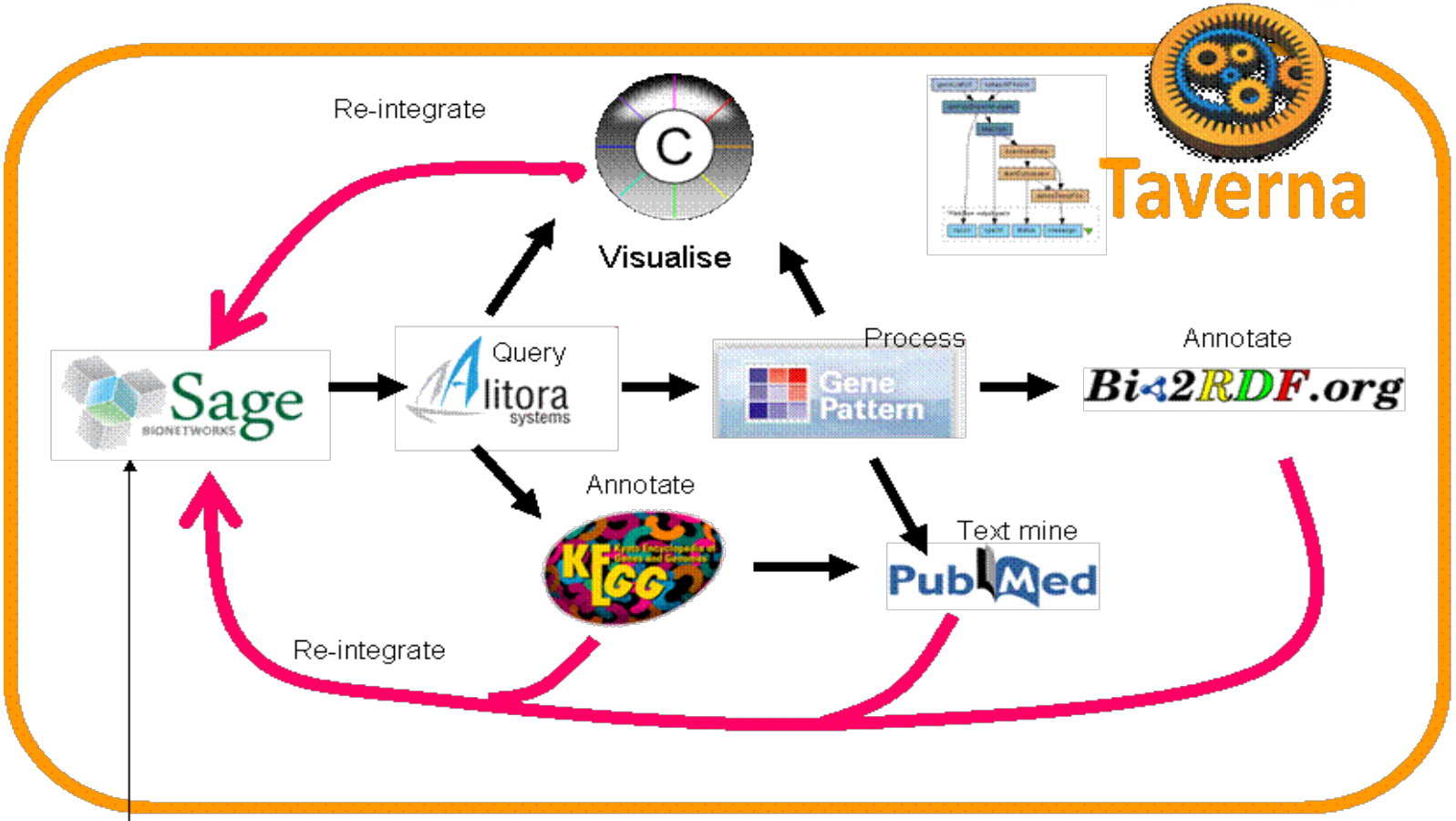
CATH, SCOP (Protein Structure Classification)



Spreadsheets, Notebooks
Local, Lost

Slide: Carole Goble

Complex data pipelines

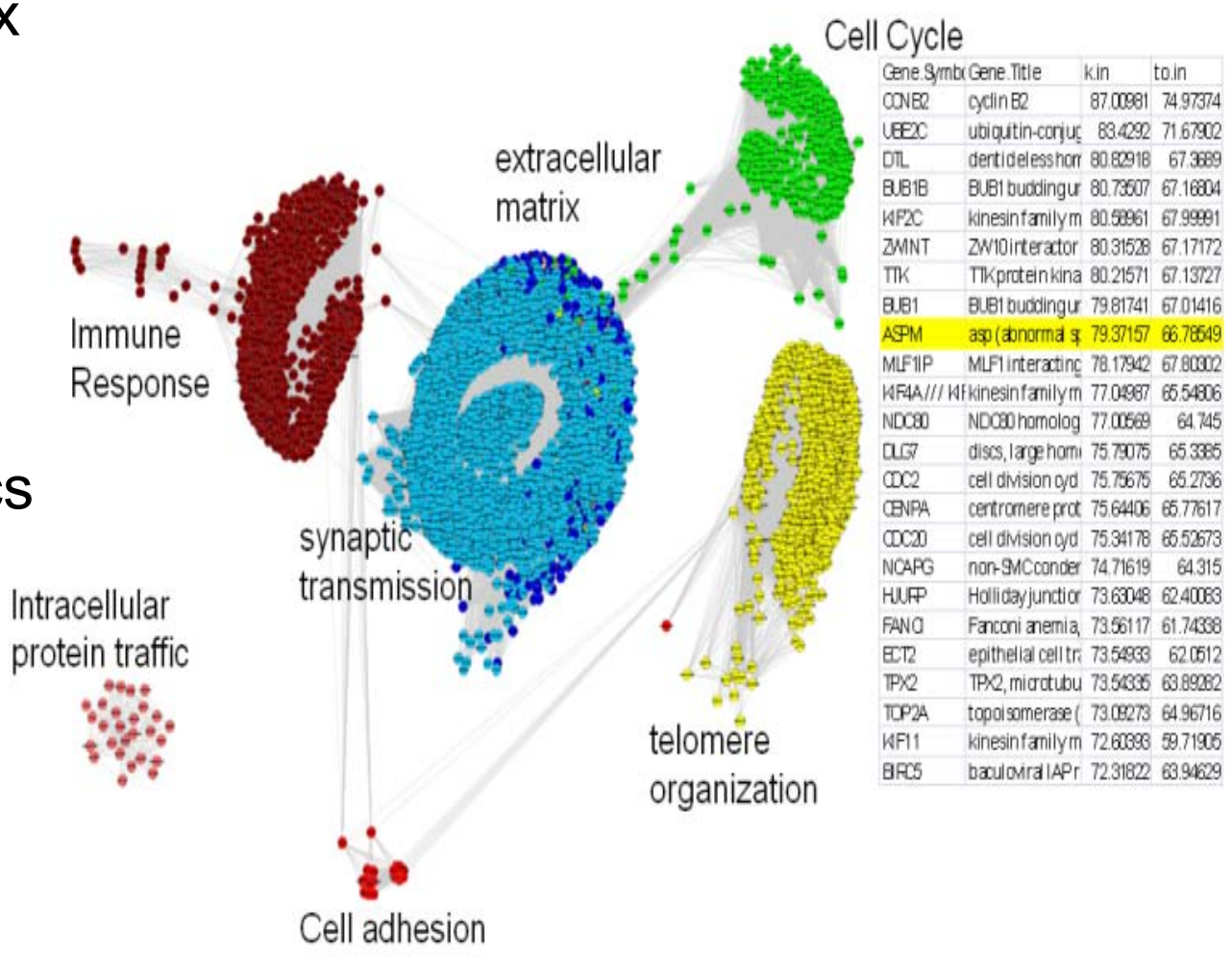


- Multiple datasets
- Visualise: Cytoscape
- Workflow: Taverna

- Distributed gene expression & clinical traits data
- Taverna workflows capture the complex model construction process
- Derive large-scale predictive network models of disease
- Integrative genomics

Sage BIONETWORKS

TCGA GBM Coexpression Network



Gene.Symb	Gene.Title	k.in	to.in
CCNB2	cyclin B2	87.00981	74.97374
UBE2C	ubiquitin-conjug	83.4292	71.67902
DTL	denticleless hom	80.82918	67.3689
BUB1B	BUB1 budding ur	80.73607	67.16804
KIF2C	kinesin family m	80.58961	67.99991
ZWINT	ZW10 interactor	80.31528	67.17172
TTK	TTK protein kina	80.21571	67.13727
BUB1	BUB1 budding ur	79.81741	67.01416
ASPM	asp (abnormal sp	79.37157	66.78549
MLF1IP	MLF1 interacting	78.17942	67.80302
KIF4A /// KIF	kinesin family m	77.04987	65.54806
NDC80	NDC80 homolog	77.00569	64.745
DLG7	discs, large homo	75.79075	65.3385
CDC2	cell division cycl	75.75675	65.2736
CENPA	centromere prot	75.64406	65.77617
CDC20	cell division cycl	75.34178	65.52673
NCAPG	non-SMC conden	74.71619	64.315
HJURP	Holliday junctior	73.63048	62.40083
FANCI	Fanconi anemia,	73.56117	61.74338
ECT2	epithelial cell tr	73.54933	62.0512
TPX2	TPX2, microtubu	73.54335	63.89282
TOP2A	topoisomerase (	73.09273	64.96716
KIF11	kinesin family m	72.60393	59.71905
BIRC5	baculoviral IAP r	72.31822	63.94629

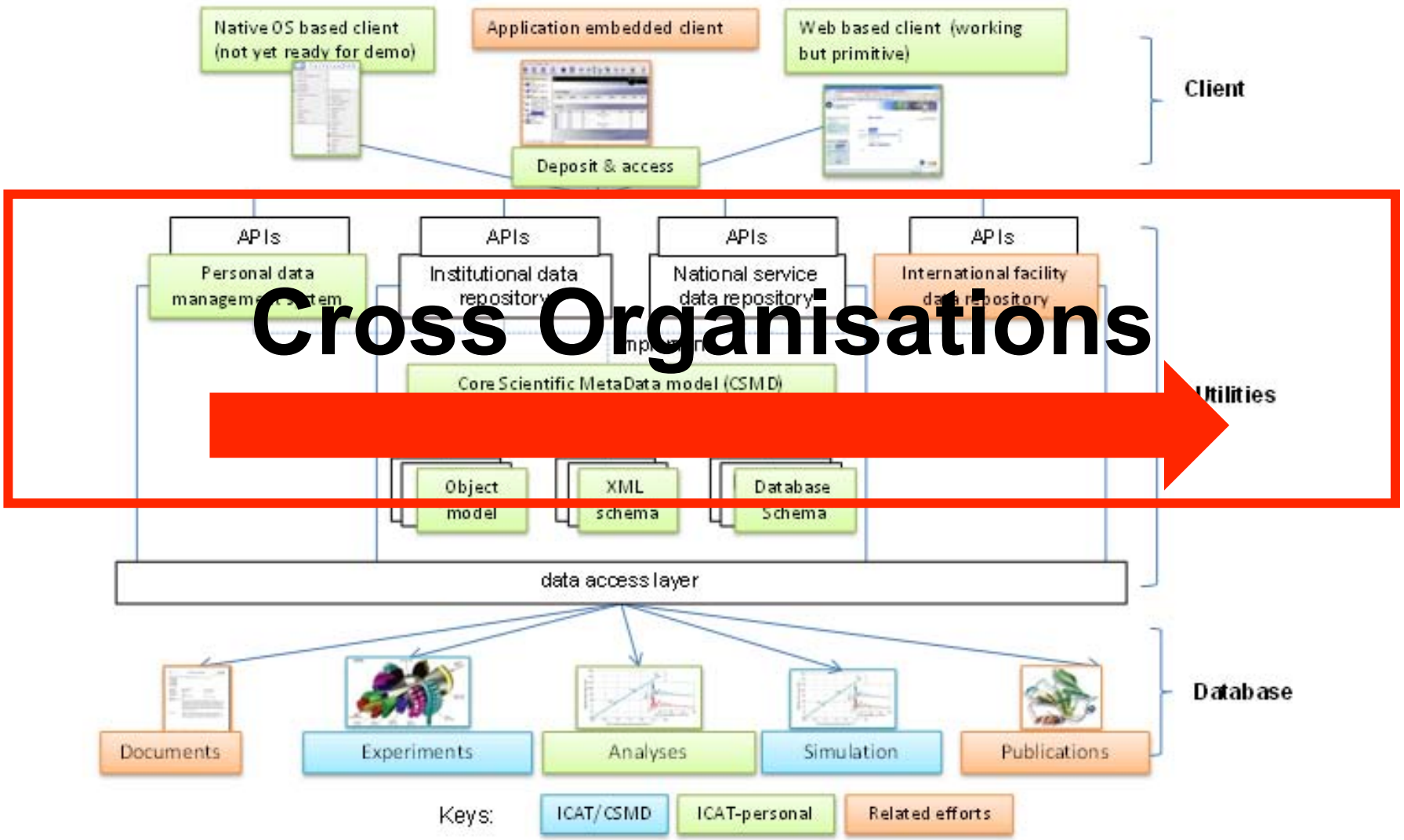
Structural Sciences Infrastructure

UNIVERSITY OF CAMBRIDGE

Department of
Earth Sciences

diamond

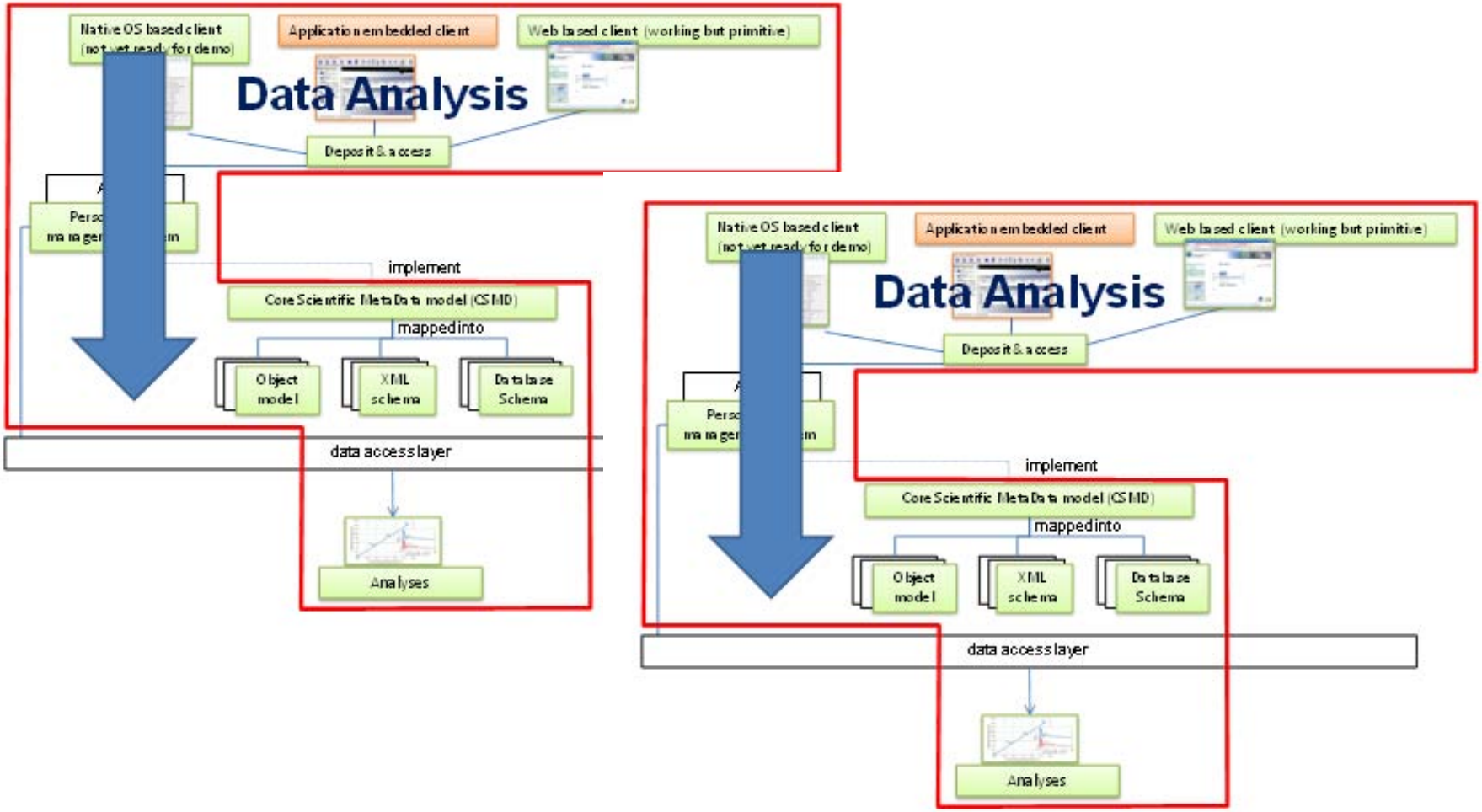
Infrastructure Roadmap



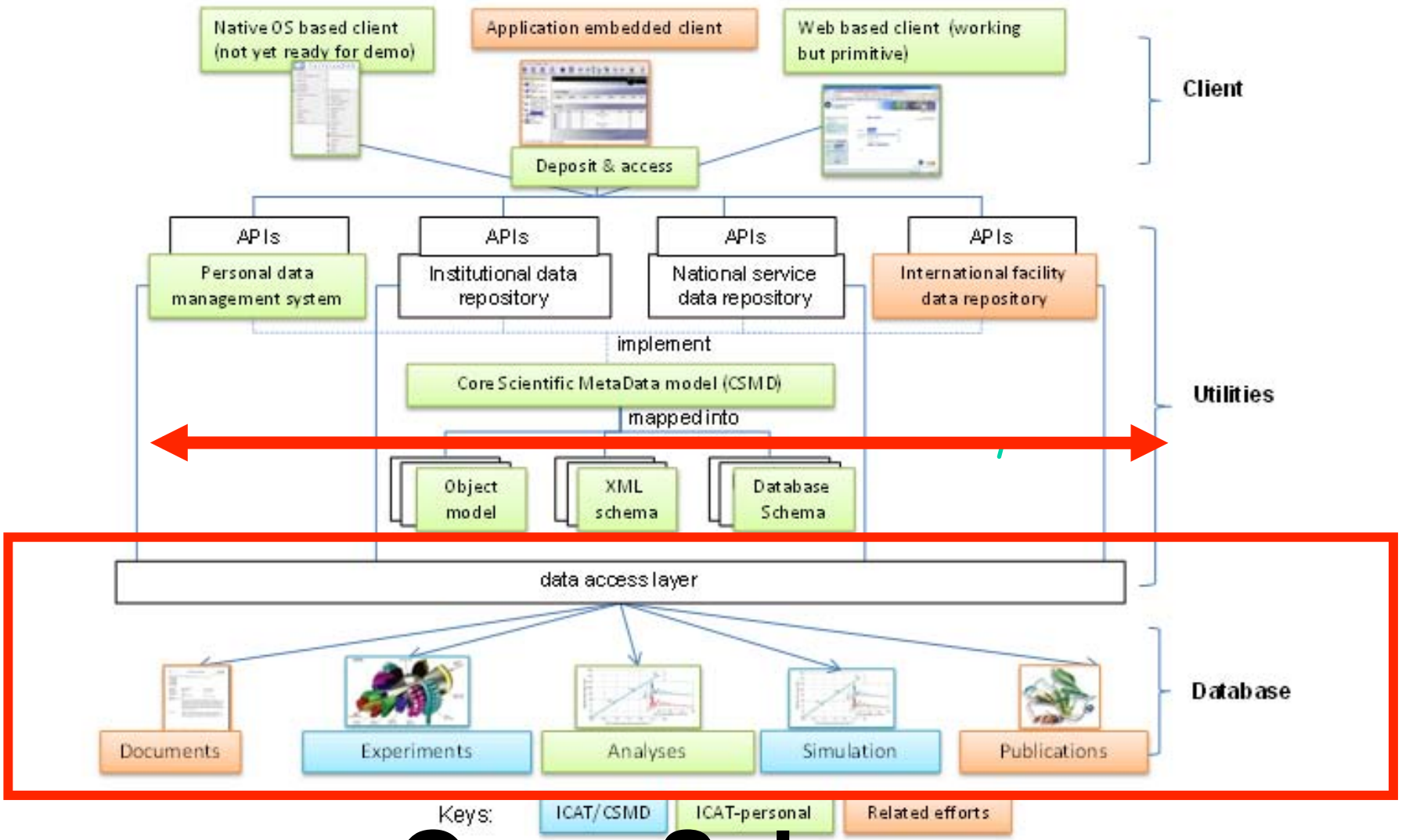
Infrastructure Roadmap

I2S2
Infrastructure for Integration in Structural Sciences

Cross Disciplines



Infrastructure Roadmap



Open Science

Panton Principles

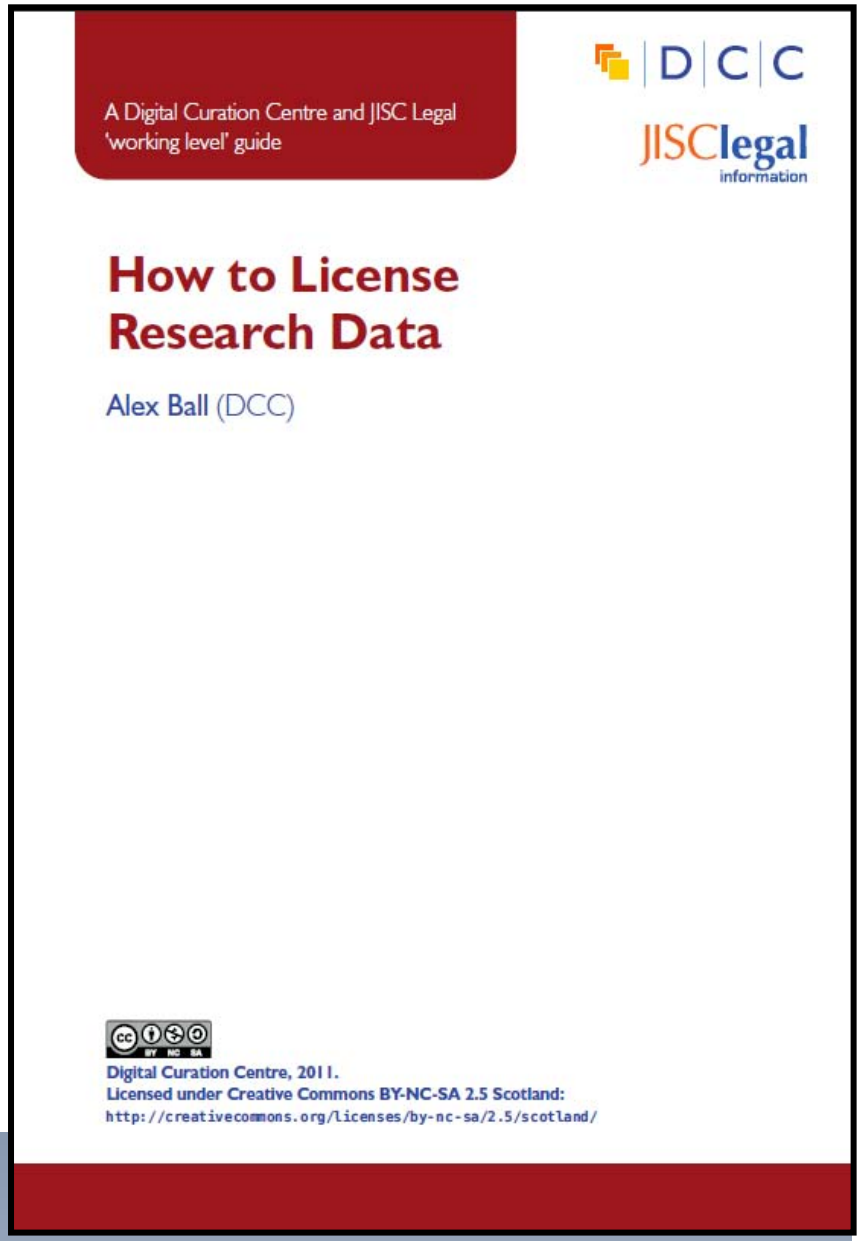
Principles for Open Data in Science

OPEN KNOWLEDGE

OPEN DATA

OPEN CONTENT

OPEN SERVICE



Open Definition

Defining the Open in Open Data, Open Content and Open Services

CITIZEN SCIENCE ALLIANCE

" The **CSA** is a collaboration of scientists, software developers and educators who collectively develop, manage and utilise **internet-based citizen science projects** in order to further science itself, and the public understanding of both science and of the scientific process. These projects use the time, abilities and energies of a **distributed community** of citizen scientists who are our collaborators "

project noah

army of citizen scientists

Yasser Ansari

Document nature with your mobile phone.

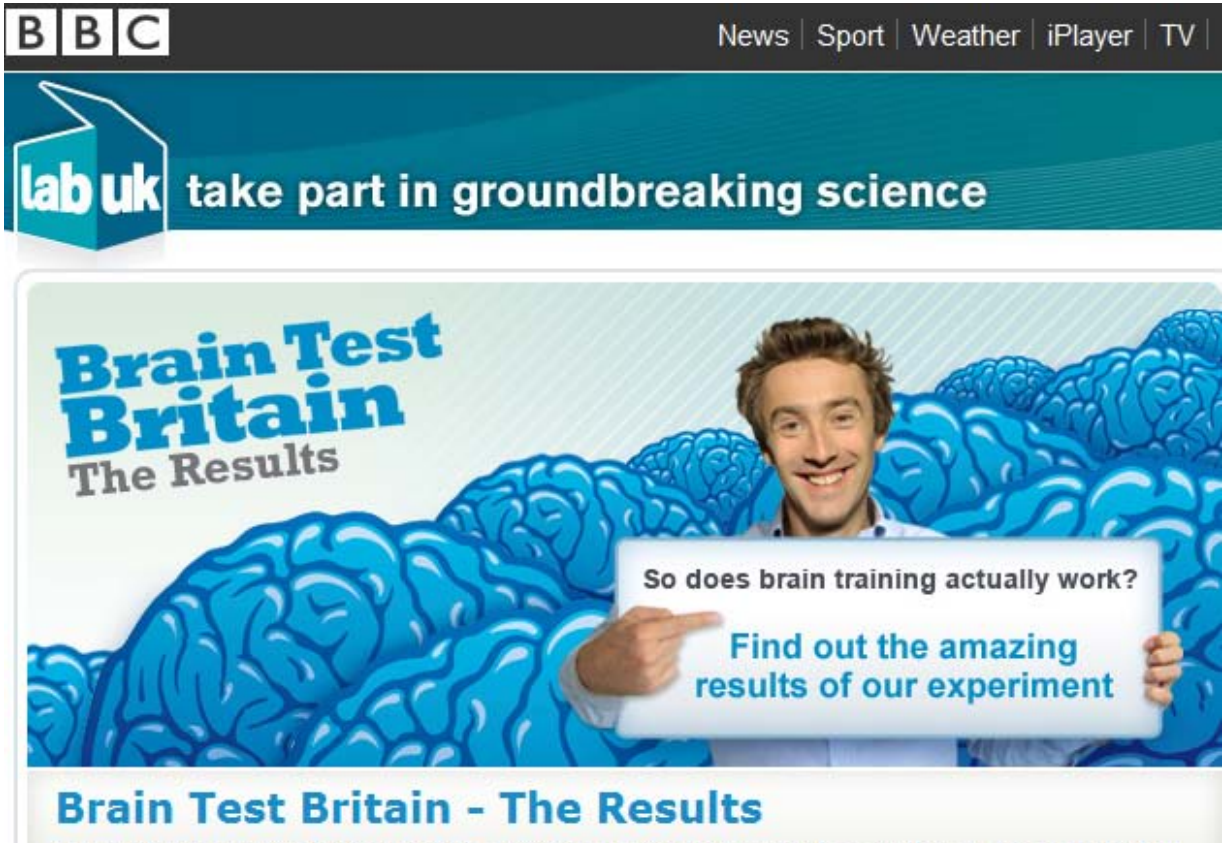
Become a top spotter!
Grab a photograph of an interesting organism and share it with the community.

Available on the App Store

Download for Android

Citizen as scientist

scienceforcitizens.net



Validate results data

Letter

Nature **465**, 775-778 (10 June 2010) | doi:10.1038/nature09042; Received 25 January 2010; Accepted 29 March 2010; Published online 20 April 2010

Putting brain training to the test

Adrian M. Owen[1], Adam Hampshire[1], Jessica A. Grahn[1], Robert Stenton[2], Said Dajani[2], Alistair S. Burns[3], Robert J. Howard[2] & Clive G. Ballard[2]

1. MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, UK
2. King's College London, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK
3. University of Manchester and Manchester Academic Health Science Centre, Manchester M13 9PL, UK

nature

Patients Participate! Project

Citizen-patients producing crowd-sourced lay summaries of UK PubMed papers

Sage Congress
San Francisco April 2010

UKOLN

JISC

BRITISH LIBRARY

amrc
ASSOCIATION OF MEDICAL RESEARCH CHARITIES

Sage BIONETWORKS

UK PUBMED CENTRAL

DCC

Policy

NERC Data Policy

This new version of the NERC Data Policy was approved by the NERC Executive Board in September 2010, and comes into force in January 2011; however, the requirement for data management plans will not be implemented until 2012, to allow NERC time to implement new grant application and review processes fully as part of the migration of grant processing to the RCUK Shared Service Centre.

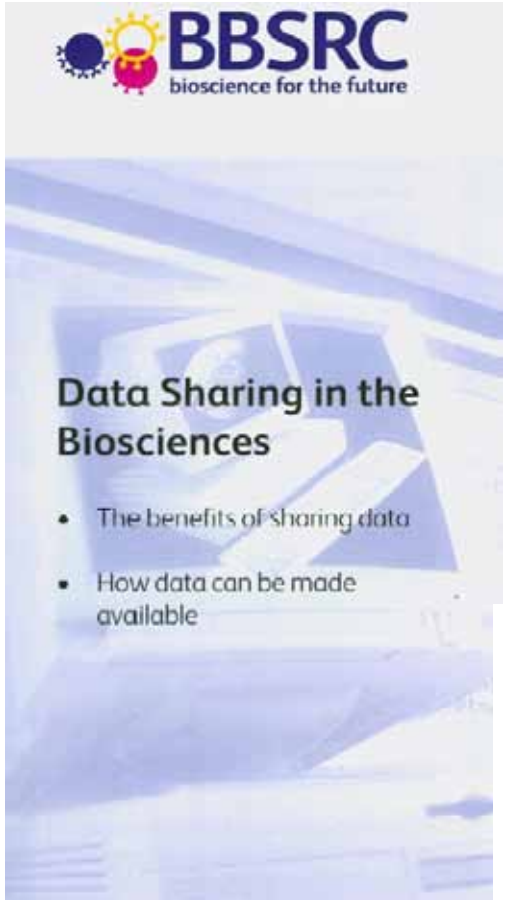
9. Working with the environmental science community NERC will maintain criteria to identify environmental data of long-term value (a Data Value Checklist). These criteria will be used to inform all decisions that NERC makes on the acceptance and disposal of data by its data centres.

Policy

NERC Data Policy

11. All applications for NERC funding must include an outline Data Management Plan, which must identify which of the data sets being produced are considered to be of long-term value, based on the criteria in NERC's Data Value Checklist. The funding application must also identify all resources needed to implement the Data Management Plan.
12. The outline data management plan will be evaluated as part of the standard NERC grant assessment process. All successful applications will be required to produce a detailed data management plan in conjunction with the appropriate NERC data centre.

Policy



wellcome trust

Policy statement

1. The Wellcome Trust expects all of its funded researchers to maximise the availability of research data with as few restrictions as possible.

2. All those seeking Wellcome Trust funding should consider their approach for managing and sharing data at the research proposal stage.In cases where the proposed research is likely to generate data outputs that will hold significant value as a resource for the wider research community, applicants will be required to submit a data management and sharing plan to the Wellcome Trust prior to an award bring made.

3. The Wellcome Trust will:

- review data management and sharing plans, and any costs involved in delivering them, as an integral part of the funding decision
- work with grant holders on an ongoing basis to support them in maximising the long-term value of key datasets resulting from their research.

National Science Foundation
WHERE DISCOVERIES BEGIN

Dissemination and Sharing of Research Results

NSF Data Sharing Policy

Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing. See Award & Administration Guide (AAG) Chapter VI.D.4.

NSF Data Management Plan Requirements

Proposals submitted or due on or after January 18, 2011, must include a supplementary document of no more than two pages labeled "Data Management Plan". This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. See Grant Proposal Guide (GPG) Chapter II.C.2.j for full policy implementation.

NSF-OCI TASK FORCE on Data and Visualization : Report coming soon….

Data management planning tool in development

A group of major research institutions is partnering to develop a flexible online tool to help researchers generate data management plans. This effort is in response to demands from funding agencies, such as the National Science Foundation (NSF) and the National Institutes of Health (NIH), that researchers plan for managing their research data.

The partners in this project include the University of California Curation Center (UC3) at the California Digital Library, the UCLA Library, the UCSD Libraries, the Smithsonian Institution, the University of Virginia Library, the University of Illinois at Urbana-Champaign, DataONE, and the United Kingdom's Digital Curation Centre (DCC).

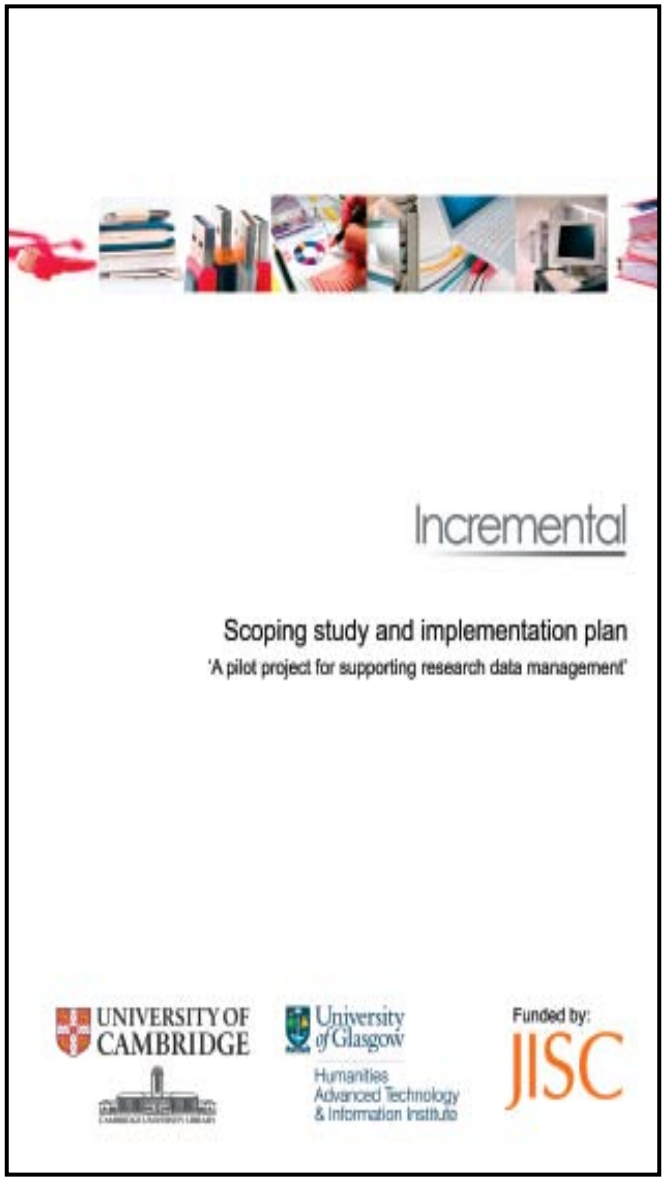
International collaboration around the DCC DMPOnline tool

DCC

Institutional perspective

- Creating & organising data
- Storage and access
- Back-up
- Preservation
- Sharing and re-use

The majority of people felt that some form of policy or guidance was needed....

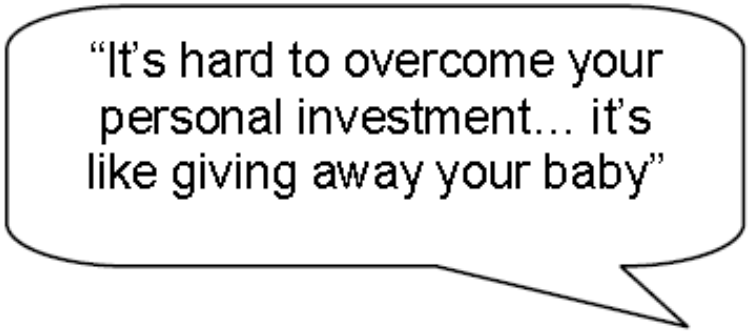


University's draft RDM policy

It shall be the University's policy that:

- Research data should be managed to the highest standards **throughout the research data lifecycle** as part of the University's commitment to research excellence.
- **The University** should provide training, support and advice, as well as mechanisms and services for storage, backup, registration, deposit and retention of research data assets in support of current and future access, during and after completion of research projects.
- **Responsibility** for research data management through a sound research data management plan during any research project or programme **lies primarily with PIs.**
- All new research proposals must include **research data management plans** or protocols that explicitly address data capture, management, integrity, confidentiality, retention, sharing and publication.
- Research data management plans must ensure that research data are **available for access and re-use** where appropriate and under appropriate safeguards.
- The legitimate interests of the subjects of research data must be protected.
- Research data of future historical interest, and all research data that represent records of the University, including data that substantiate research findings, **should be offered and assessed for deposit and retention in an appropriate national or international data service or domain repository, or a University repository.** Such research data deposited elsewhere should be registered with the University.

Jeff Haywood, RDMF V October 2010 *http://www.dcc.ac.uk/sites/default/files/documents/RDMF/RDMF5/Haywood.pdf*



"While many researchers are positive about sharing data in principle, they are almost universally reluctant in practice. ..... using these data to publish results before anyone else is the primary way of gaining prestige in nearly all disciplines."

"Data sharing was more readily discussed by early career researchers."

DCC

INCREMENTAL Project

The New York Times

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010

Alzheimer's Disease Neuroimaging Initiative:
a unique (open) $60M partnership between
NIH, FDA, universities and drug companies.

"It was unbelievable. Its not science the way most of us have practiced in our careers. But we all realised that we would never get biomarkers unless all of us parked our egos and intellectual property noses outside the door and agreed that all of our data would be public immediately."

Dr John Trojanowski, University of Pennsylvania

University told to hand over tree ring data - April 15, 2010

BBC NEWS

Posted by Daniel Cressey on April 15, 2010

Data is headline news

FOI & RESEARCH DATA: RESEARCHERS' QUESTIONS AND ANSWERS

Table of Contents *Comments by Section* *Comments by Users* *General Comments* *Login*

JISC FoI FAQ

TABLE OF CONTENTS

There are 51 comments in this document

1. Introduction (3)
2. Q1 How do I recognise a FoI or EIR request? (2)
3. Q2 What's the short answer on what I should do if asked for data? (7)
4. Q3 Why should I make my data available? (5)
5. Q4 How long have I got to respond to a request? (4)
6. Q5 I don't want to provide my data. What must I do first? (1)

The general rule is that the Data Protection Act trumps FoI/EIR. Both FoI Acts make personal data, of which the requester is the subject, exempt information (there is a similar exception under EIR). The requester should apply under the UK-wide Data Protection Act (for which different rules, timescales and fees apply). If the requester is not the subject of the personal data, the exemptions become more complicated, although our "general rule" above is likely to apply. Always discuss such c [...]

research information network

Regional workshops

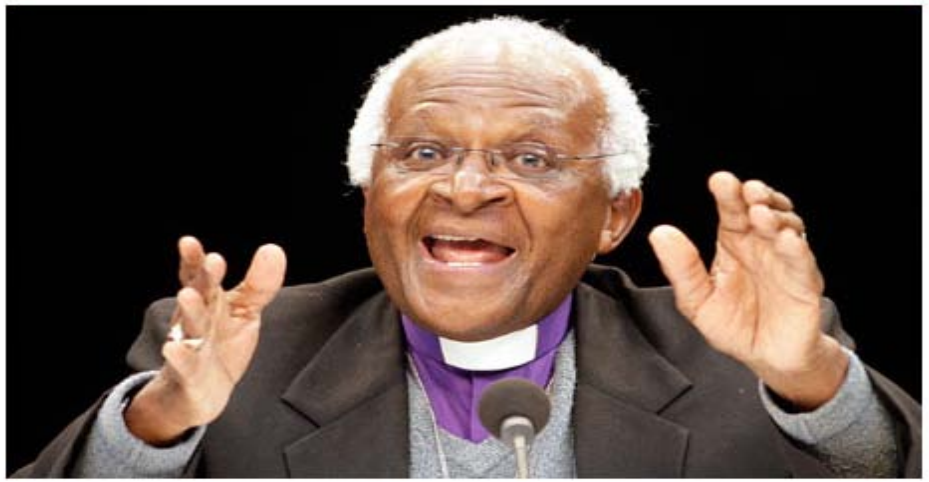
1000 Genomes Project Releases Data from Pilot Projects on Path to Providing Database for 2,500 Human Genomes

Freely available data supporting next generation of human genetic research

Desmond Tutu's genome sequenced as part of genetic diversity study

Archbishop Desmond Tutu has had his genome sequenced in research to reveal the true breadth of human genetic diversity

Ian Sample, science correspondent
guardian.co.uk, Wednesday 17 February 2010 18.02 GMT
Article history



P4 medicine: Predictive, Personalised, Preventive, Participatory.

Leroy Hood –
Institute for Systems Biology

Your genome is basis for your medical record

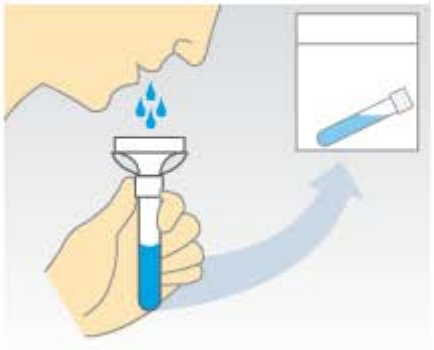
Open data and ethics

Buy a DIY kit?
Share your data?

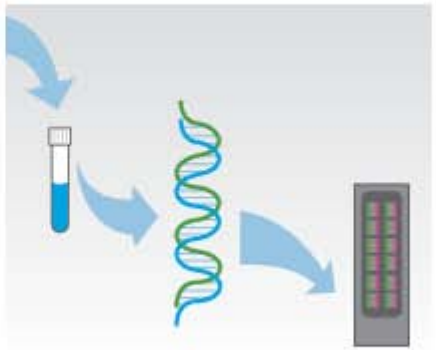
Get the latest on your DNA with $399 and a tube of saliva.

Here's what you do:

1. Order a kit ($399 USD) from our online store.



2. Claim your kit, spit into the tube, and send it to the lab.



3. Our CLIA-certified lab analyzes your DNA in 2-4 weeks.

4. Log in and start exploring your genome.

Open data and ethics

- **Bring your genes to CAL**
- UC Berkeley personalised medicine initiative
- >700 new students have submitted a genetic sample and a consent form
- Aggregate analyses for three genes related to nutrition
- Implications for UK HE students & staff?

Policy Gaps...

- Is Policy disconnected from Practice?
  - Data Sharing
  - Data Licensing
  - Ethics and Privacy
  - Citizen Science & Public Engagement
  - Data Storage, Selection & Appraisal
  - Data Citation and Attribution

"I just back everything up onto data sticks. I didn't even know you could back-up to servers".

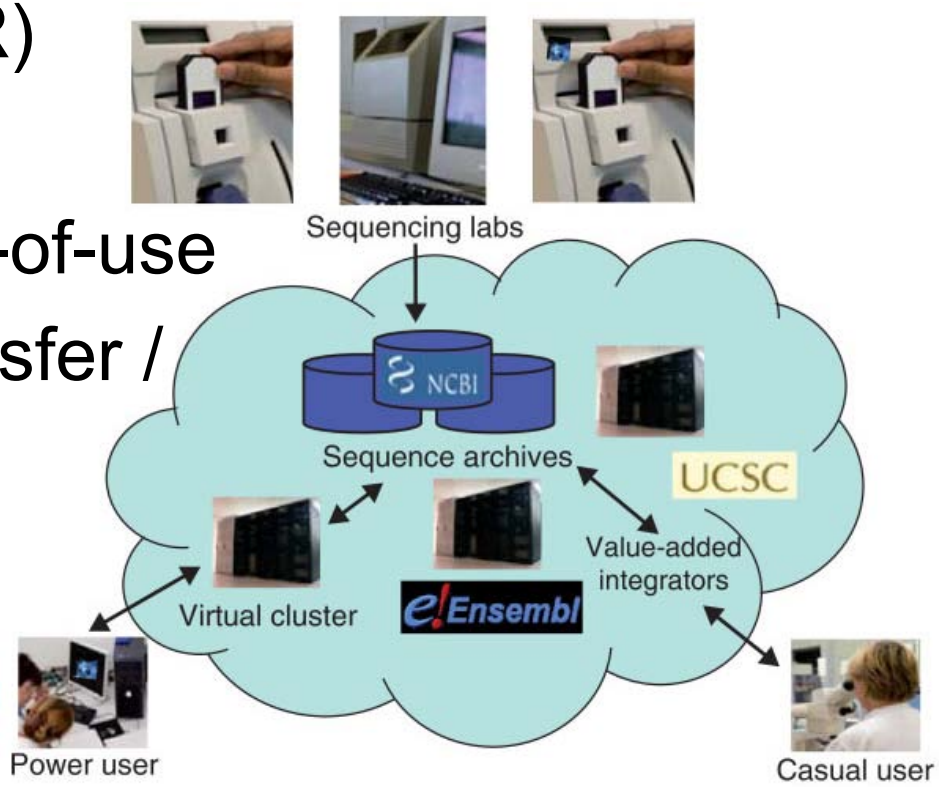
http://www.flickr.com/photos/mattimattila/3003324844/

"Departments don't have guidelines or norms for personal back-up and researcher procedure, knowledge and diligence varies tremendously. Many have experienced moderate to catastrophic data loss"

Incremental Project Report, June 2010

Data storage...

- Scaleable
- Cost-effective (rent on-demand)
- Secure (privacy and IPR)
- Robust and resilient
- Low entry barrier / ease-of-use
- Has data-handling / transfer / analysis capability



- Cloud services?

The case for cloud computing in genome informatics. Lincoln D Stein, May 2010

Genome Biology

The New York Times

When American and European Ideas of Privacy Collide

By ADAM LIPTAK
Published: February 27, 2010

WORLD PRIVACY FORUM

Privacy in the Clouds:
Risks to Privacy and Confidentiality from Cloud Computing

Prepared by Robert Gellman
for the World Privacy Forum

February 23, 2009

European Data Privacy Restrictions Slow Cloud Computing Spread

Posted by Lora Bentley 21-Sep-2010 16:00:17

Your data in the cloud

Cloud Computing for Research

The *Window* Conference Centre, London, Tuesday 20 July 2010

RESEARCH COUNCILS UK

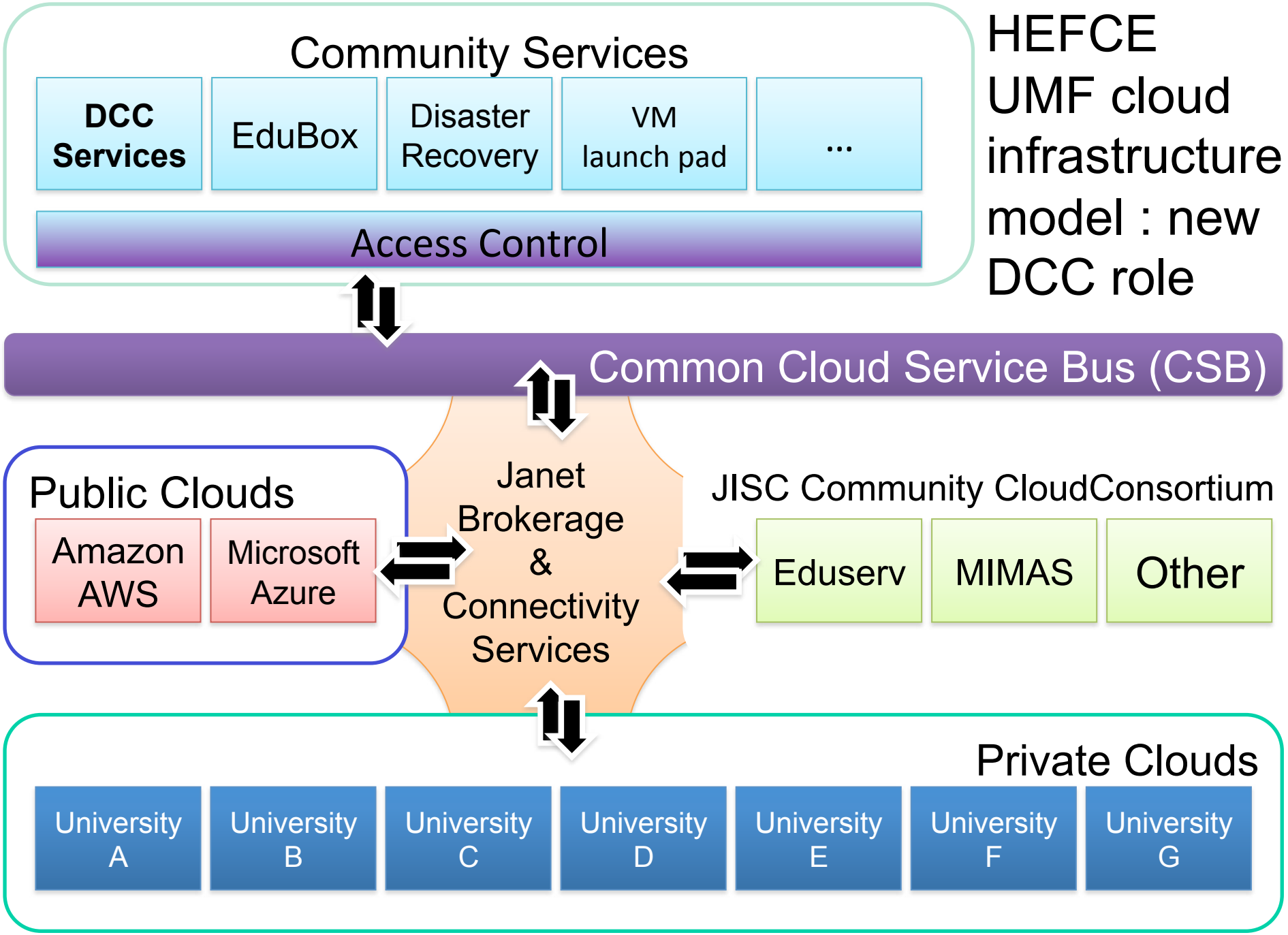
THE ROYAL SOCIETY

Research Councils UK Digital Economy Transforming Business and Society

Cloud Matters: Ethics and Policy in the Digital Age

6th July 2010, Royal Society

REPORT


Community Services
DCC Services
EduBox
Disaster Recovery
VM launch pad
...
Access Control
HEFCE UMF cloud infrastructure model : new DCC role
Common Cloud Service Bus (CSB)
Public Clouds
Amazon AWS
Microsoft Azure
Janet Brokerage & Connectivity Services
JISC Community CloudConsortium
Eduserv
MIMAS
Other
Private Clouds
University A
University B
University C
University D
University E
University F
University G

Incentivising data management

Editorial

Nature Cell Biology **11**, 1273 (2009)
doi:10.1038/ncb1109-1273a

nature cell biology

Sharing data

Reference datasets should be accessible independently of scientific papers in a citable form, allowing attribution.

nature

OPINION

Let's make science metrics more scientific

To capture the essence of good science, stakeholders must combine forces to create an open, sound and consistent system for measuring all the activities that make up academic productivity, says **Julia Lane**.

PLoS COMPUTATIONAL BIOLOGY

Scholar Factor (SF)

Philip E. Bourne, J. Lynn Fink

Correspondence

Nature Biotechnology **27**, 984 - 985 (2009)
doi:10.1038/nbt1109-984b

Accreditation and attribution in data sharing

Gudmundur A Thorisson[1]

1. Department of Genetics, Univer

nature biotechnology

Credit where credit is overdue

A universal tagging system that links data sets with the author(s) that generated them is essential to promote data sharing within the proteomics and other research communities.

Beyond the PDF Workshop, January 2011

- Concept of “reproducibility”
- Executable papers
- Data papers
- Links to data, workflows, analyses (GenePattern) within a document
- Post-publication peer review
- Alternative impact metrics : downloads, slide reuse, data citation, YouTube views etc.
- La Jolla Manifesto : guiding principles for digital scholarship

Jodi Schneider, Ariadne, Issue 66, January 2011

SageCite

Sage BIONETWORKS · MANCHESTER 1824 · UKOLN · BRITISH LIBRARY

DCC

Process
(Taverna workflow)

Research Object

Citation Chains

DataCite
International Data Citation

Citation Framework

Data Commons
(Sage)

Credit & Attribution

Publication
(Nature PG, PLoS)

Sage BIONETWORKS

PLoS COMPUTATIONAL BIOLOGY
An official journal of the International Society for Computational Biology

nature genetics

SageCite

Slide : Peter Li

Taverna workflow

The relationships between data via DataCite DOIs with tools are captured by the provenance (OPM) produced by Taverna

1 Produces

Data

Workflow metadata

Generate landing page for data

5

2 Register

6

DataCite API

DataCite

BRITISH LIBRARY

3 Mint

DOIs

For referring to data reported in the provanance?

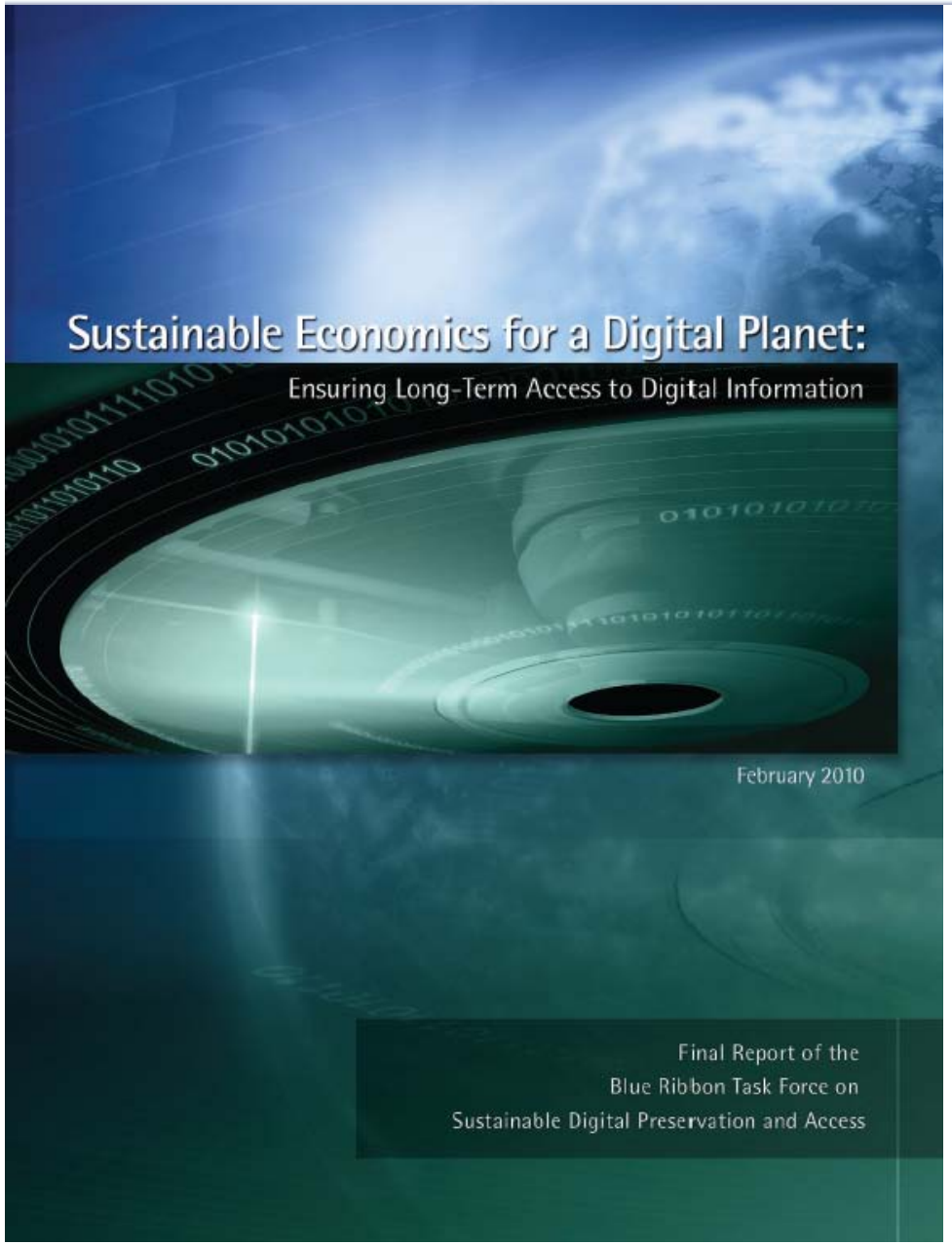
4 Resolve to landing page

Google API

sagecitedemorepository

Google sites

Blue Ribbon Task Force
on Sustainable Digital Preservation and Access



Sustainability:

Who owns?
Who benefits?
Who selects?
Who preserves?
Who pays?

4 Domain areas

Scholarly Discourse	Research Data
Commercially-Owned Cultural Content	Collectively-Produced Web Content

Keeping Research Data Safe Factsheet

Cost issues in digital preservation of research data

This factsheet illustrates for institutions, researchers, and funders some of the key findings and recommendations from the JISC-funded Keeping Research Data Safe (KRDS1) and Keeping Research Data Safe 2 (KRDS2) projects. Further information on the research and findings can be found in the final reports.

What Costs Most?

Acquisition and ingest costs most. The costs of archival storage and preservation activities are consistently a very small proportion of the overall costs and significantly lower than the costs of acquisition/ingest or access activities for all our case studies. Note we believe early preservation action during ingest or pre-ingest produces lower costs over the lifecycle as a whole. (KRDS1, p.25; KRDS2, pp.31-52)

Activity Costs for the Archaeology Data Service		
Outreach/ Acquisition/ Ingest	Archival Storage and Preservation	Access
c. 55%	c. 15%	c. 31%

Recommendation to Funders

From our research, it is likely that the largest potential cost efficiencies will come from future tool development supporting automation of ingest and access activities for curation and preservation. (KRDS2, p.83)

Impact of Fixed Costs

- The costs of long-term data curation/preservation are dominated by fixed costs that do not vary with the size of the collections;
- Staff are the major cost component overall and there is a minimum base-level of staff cover, skills and equipment required for any service;
- Activities characterised by significant fixed costs can reduce the per-unit cost of long-term preservation by leveraging economies of scale.

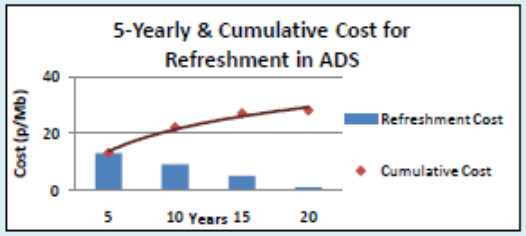
(KRDS2, pp.32-34, 79-80)

Recommendation to Institutions

Repositories should take advantage of economies of scale, using multi-institutional collaboration and outsourcing as appropriate. Once core capacity is in place additional content can be added at increasing levels of efficiency and lower cost. (KRDS1, pp.77-78)

Declining Costs over Time

We found a trend of relatively high preservation costs in the early years reducing substantially over time for data collections. An example is the preservation costs projected for the Archaeology Data Service (ADS) based on their experience of the first 10 years of operating the data service. (KRDS1, pp.4-6)



Costs for archival storage and preservation ("refreshment") decline to a minimal level over 20 years

Recommendation to Funders and Institutions

The implications of these factors and projection for sustainability of data archives e.g. via archive charges to project budgets, are notable and worthy of more extensive study and testing. (KRDS1, pp.5-6)

KRDS


Charles Beagrie

KEEPING RESEARCH DATA SAFE 2

Neil Beagrie, Brian Lavoie and Matthew Woollard

with contributions by the Universities of Cambridge, Oxford, and Southampton, the Archaeology Data Service, OCLC Research, UK Data Archive, and University of London Computer Centre.

Final Report - April 2010

Prepared by:

Charles Beagrie Limited

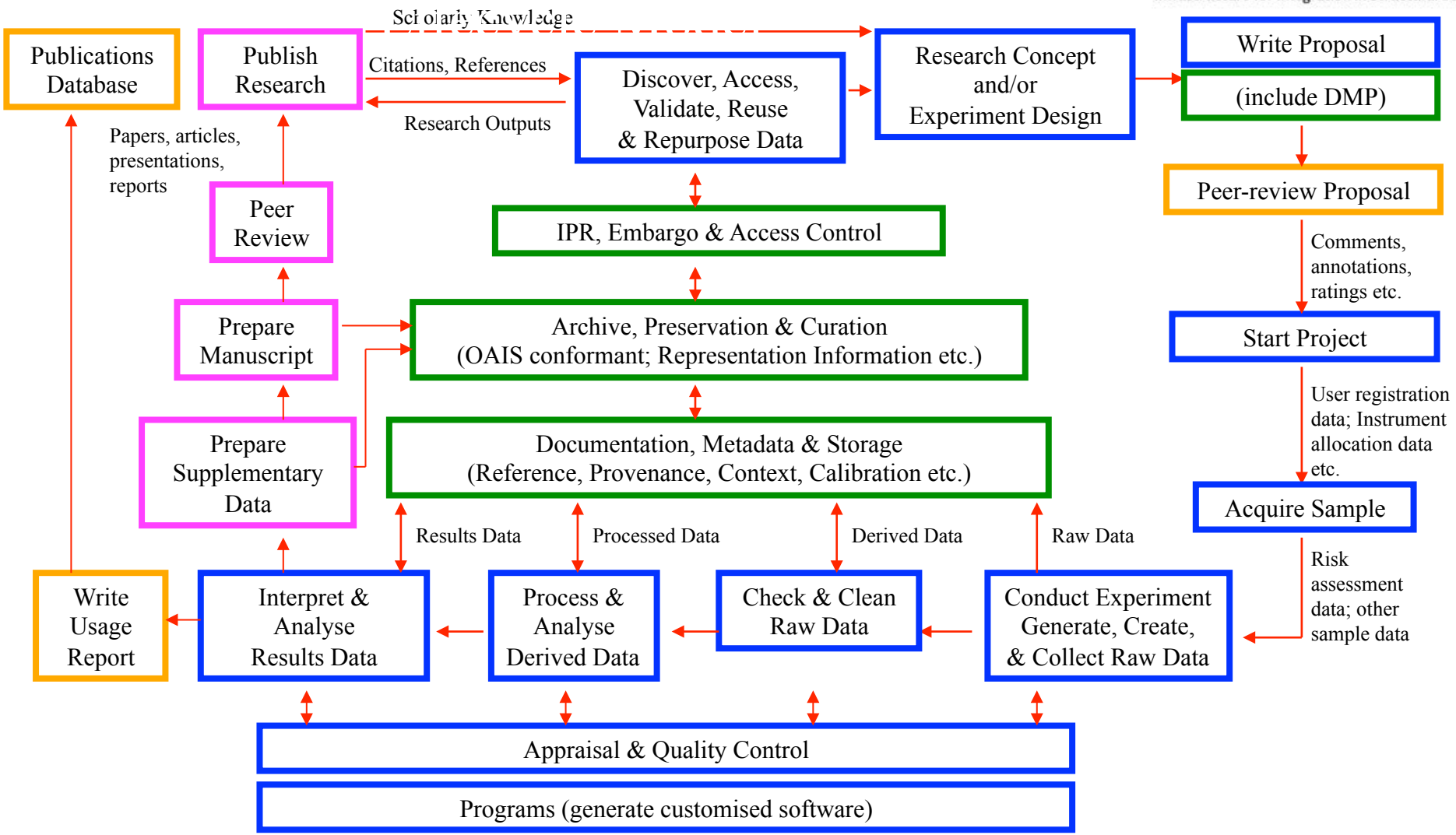
www.beagrie.com

A study funded by

JISC

With support from OCLC Research and the UK Data Archive

An Idealised Scientific Research Data Lifecycle Model



KRDS Activity Model Benefits & Metrics

Use Case 1 : National Crystallography Service

Use Case 2 : Researcher in the lab

Activity e.g. Write proposal, conduct experiment, analyse raw data

Key resources to measure e.g. Time spent, staff salary

Metrics e.g. Time savings for researcher, times savings for facility, increased output, % data available for re-use

Qualitative impacts e.g. Research quality, knowledge transfer to industry

I2S2

Infrastructure for Integration in Structural Sciences

Research Data Management Forum 6

Planning for Research Data Management: "Meeting funder imperatives"

5-6 May 2011, University of Leicester

Data Infrastructure Challenges: working across scale, disciplinary and institutional boundaries,
5th May 2011, University of Leicester

I2S2
Infrastructure for Integration in Structural Sciences

University of Leicester

DCC because good research needs good data