Building Capacity and Capability for Data : Requirements, Challenges, Opportunities

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

Horizon2020 Workshop Brussels, May 2012

UKOLN is supported by:

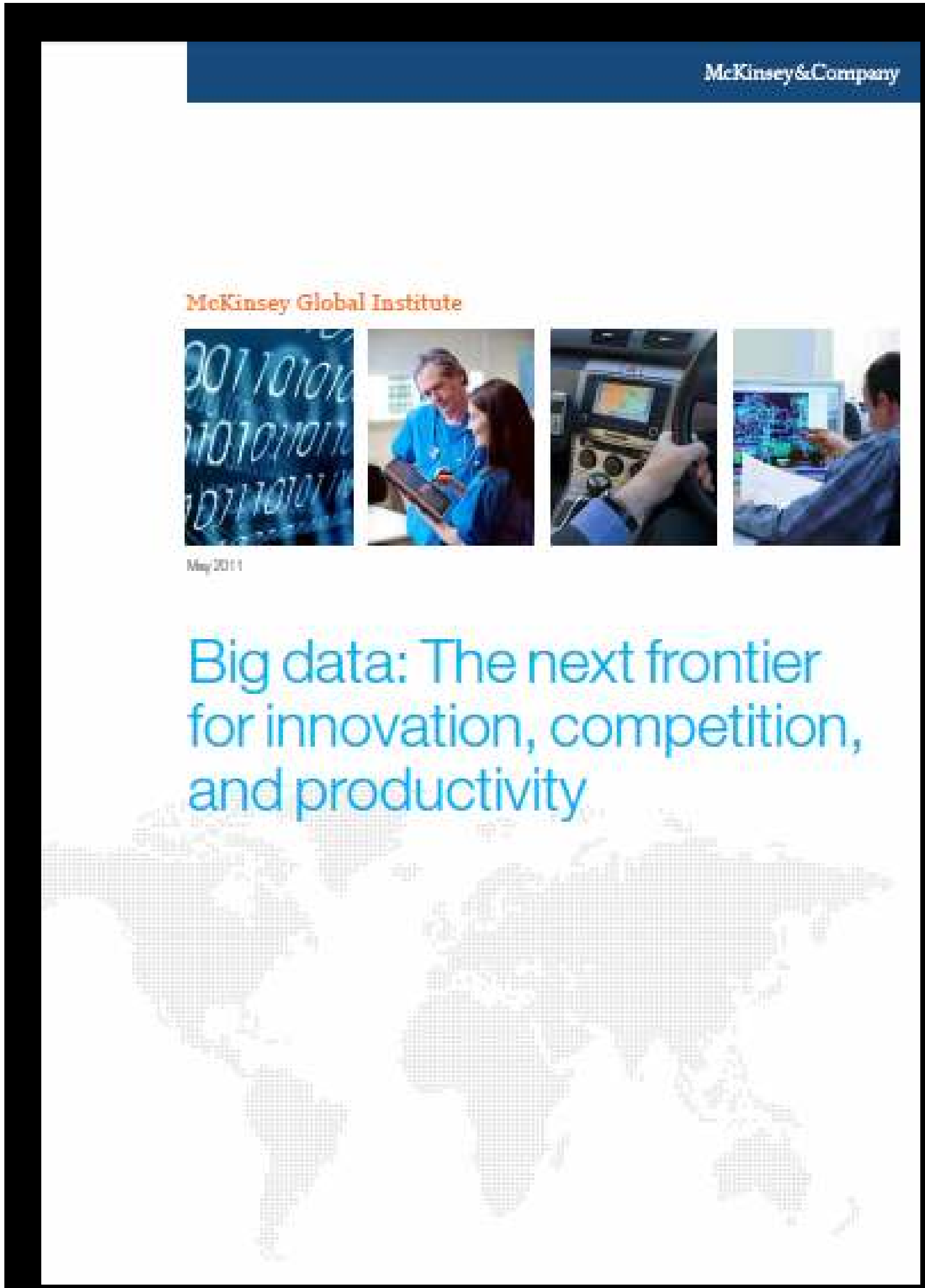
JISC

Running order…..

- Data landscape snapshots
- Roles and responsibilities
- Skills and competencies
- Gaps and opportunities

“The ability to take data - to be able to understand it, to process it, to extract value from it, to visualise it, to communicate it - that’s going to be a hugely important skill in the next decades.”

Hal Varian, Chief Economist, Google



McKinsey&Company

McKinsey Global Institute

May 2011

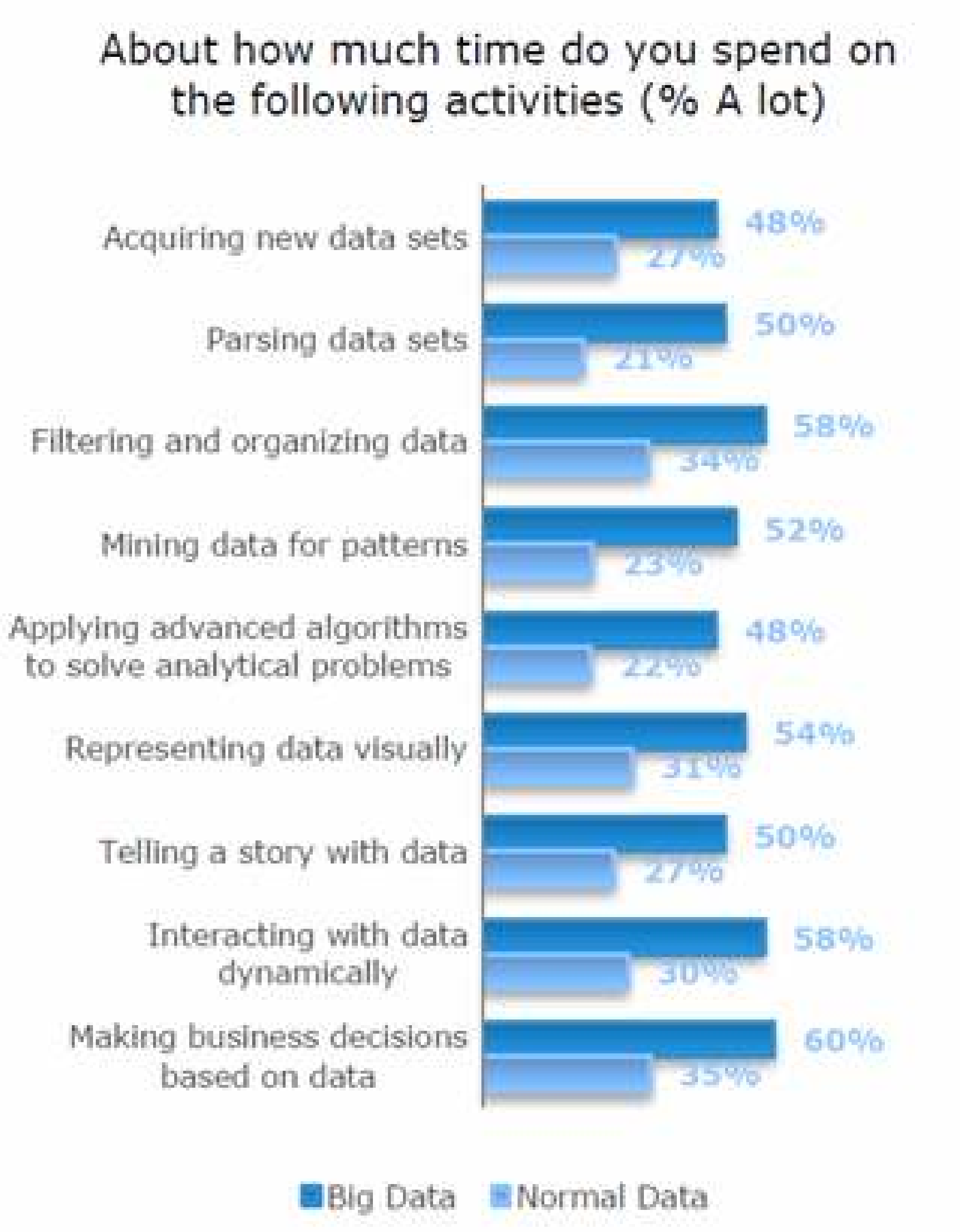
Big data: The next frontier for innovation, competition, and productivity

http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation

Implications of "Big Data" and data science for organisations in all sectors

Predicts a shortage of 190,000 data scientists by 2019

“Big Data”
Data scientist



About how much time do you spend on the following activities (% A lot)

EMC²

Data Science Revealed community survey

http://www.emc.com/collateral/about/news/emc-data-science-study-wp.pdf

Forbes

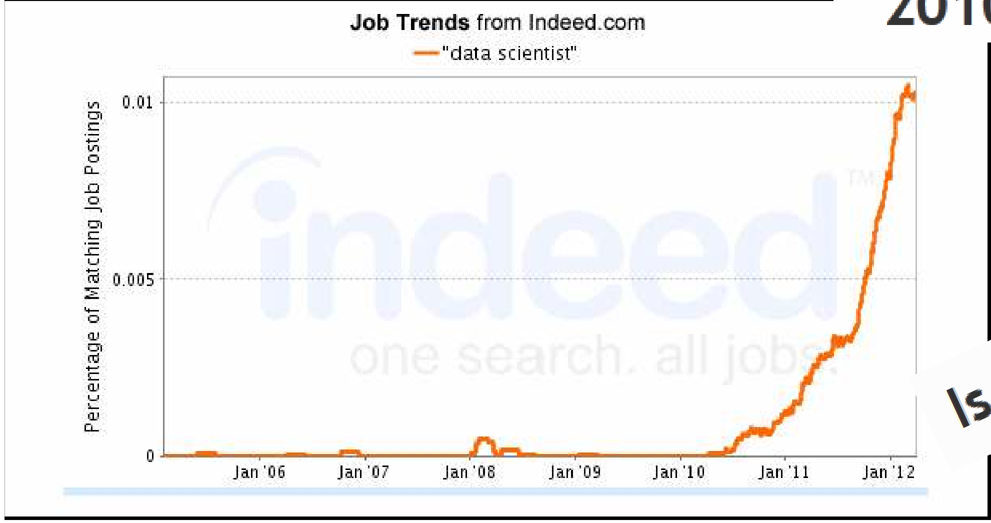
Big Data Needs Data Scientists, Or Quants, Or Excel Jockeys

Data Scientist = Rock Star, Really?

The Term "Data Scientist" is Still New

CMS WiRE

"Data Scientist" Jobs = Near Zero Until 2010



Is There a Shortage of Data Scientists?

DATABLOG

Facts are sacred

What is a data scientist?

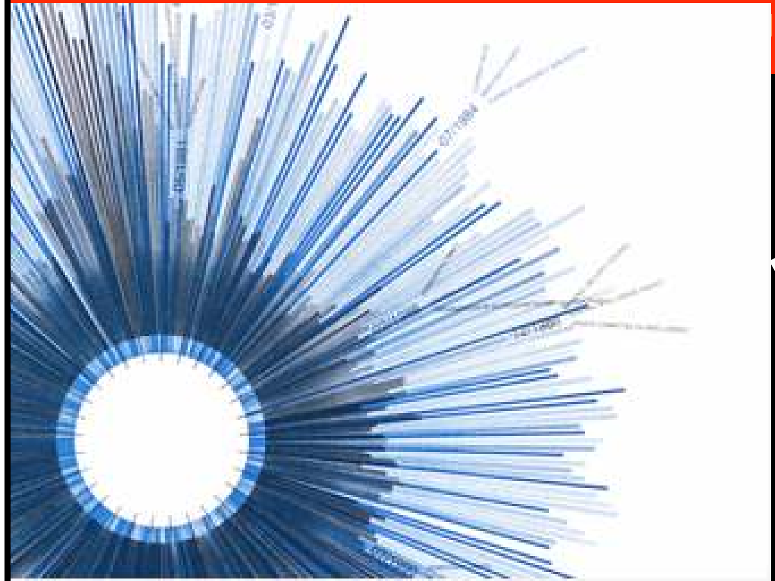
It's the job of the moment. But what exactly is a data scientist?

Other data-related roles?

Data journalist

New York Times Data Artist in Residence, Jer Thorp Joins Stellar Cast of Speakers at TEDxVancouver 2011

1 · No Comments »



Jer Thorp: Hope / Crisis, NYT Word Frequency

The New York Times

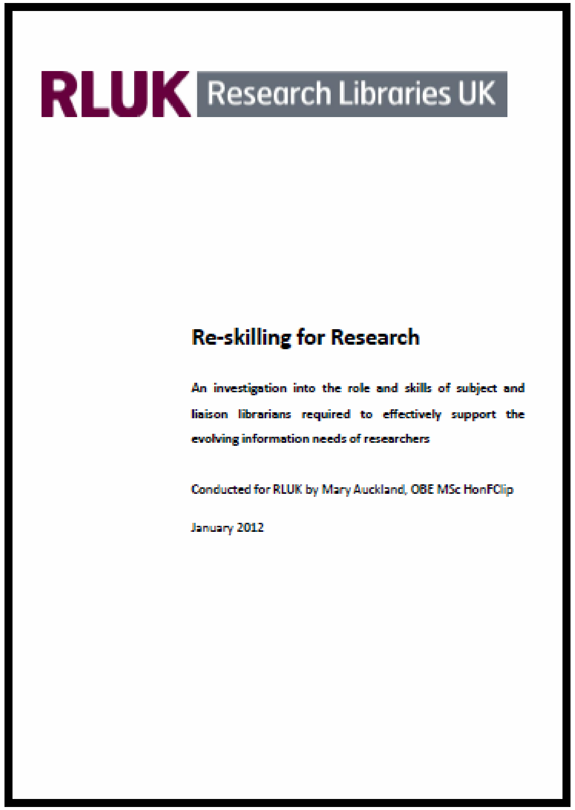
Position	Location
Science Data Librarian	Stanford
Data Management Librarian	Oregon State
Social Sciences Data Librarian	Brown
Data Curation Librarian	Northeastern
Data Librarian	New South Wales
Research Data Management Co-ordinator	Sydney
Research Data & Digital Curation Officer	Cambridge
Data Services Librarian	Iowa
Data Analyst	ANDS
Institutional Data Scientist	Bath

MANAGING RESEARCH DATA

EDITED BY GRAHAM PRYOR

RLUK/Mary Auckland:
Reskilling for Research
9 areas are skill gaps
for subject librarians

Sheila Corrall: Libraries,
Librarians and Data
Many action exemplars



RLUK Research Libraries UK

Re-skilling for Research

An investigation into the role and skills of subject and liaison librarians required to effectively support the evolving information needs of researchers

Conducted for RLUK by Mary Auckland, OBE MSc HonFClip

January 2012

2012: Libraries in review

Skill gap	**2-5 years**	**Now**
Preserving research outputs	49%	10%
Data management & curation	48%	16%
Comply with funder mandates	40%	16%
Data manipulation tools	34%	7%
Data mining	33%	3%
Metadata	29%	10%
Preservation of project records	24%	3%
Sources of research funding	21%	8%
Metadata schema, discipline standards, practices	16%	2%

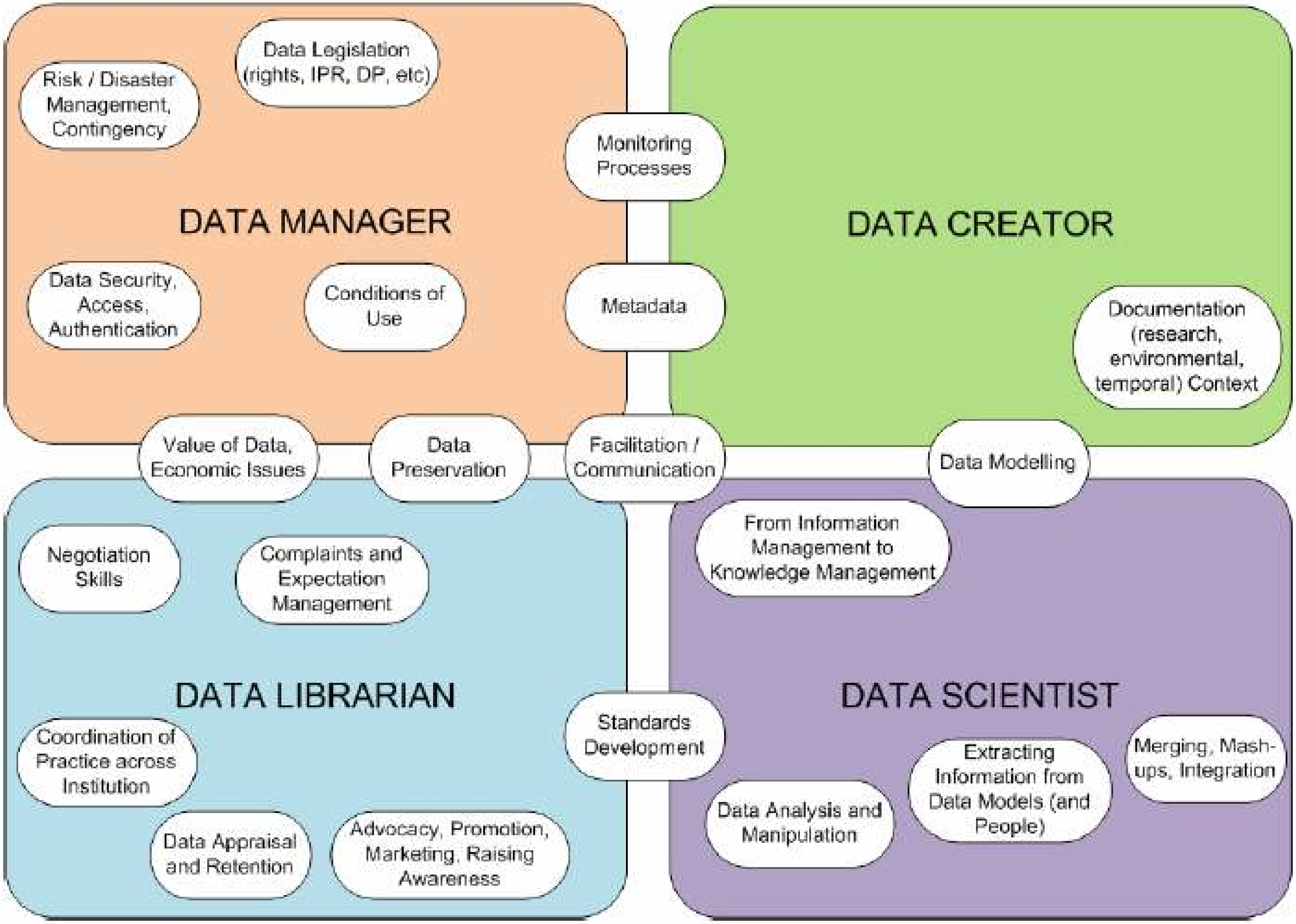
Data from RLUK/Mary Auckland: Reskilling for Research 2012

“Very few librarians are likely to have specialist scientific or medical knowledge - if you train as a research scientist or a medic, you probably won’t become a librarian.”

RLUK/Mary Auckland: Reskilling for Research 2012

CORE SKILLS FOR DATA MANAGEMENT

A follow-up from the second DCC Research Data Management Forum (November 2008)



DATA MANAGER

Risk / Disaster Management, Contingency

Data Legislation (rights, IPR, DP, etc)

Data Security, Access, Authentication

Conditions of Use

Monitoring Processes

Metadata

DATA CREATOR

Documentation (research, environmental, temporal) Context

Value of Data, Economic Issues

Data Preservation

Facilitation / Communication

Data Modelling

DATA LIBRARIAN

Negotiation Skills

Complaints and Expectation Management

Coordination of Practice across Institution

Data Appraisal and Retention

Advocacy, Promotion, Marketing, Raising Awareness

Standards Development

DATA SCIENTIST

From Information Management to Knowledge Management

Data Analysis and Manipulation

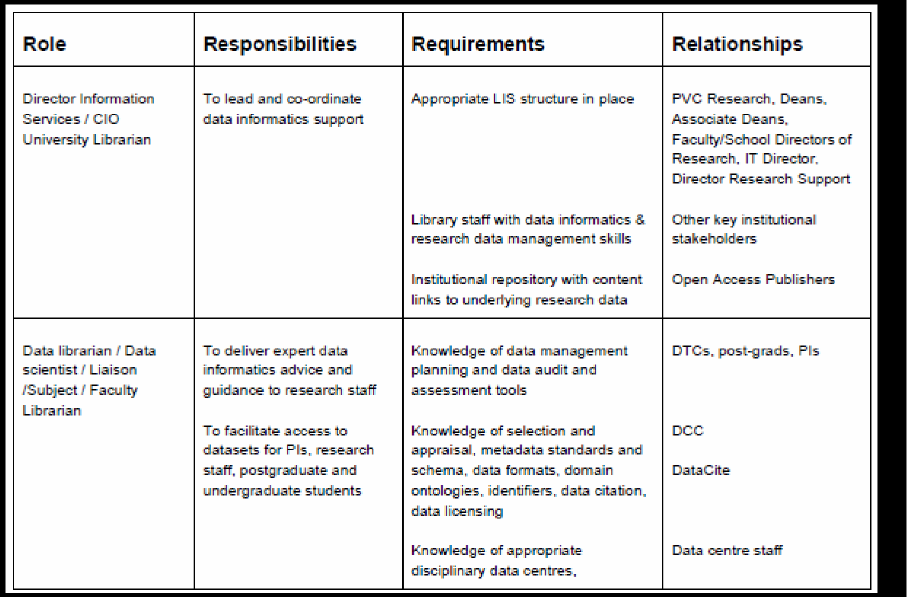
Extracting Information from Data Models (and People)

Merging, Mash-ups, Integration

- Leadership & co-ordination
- Strategy and planning
- Policy
- Legal and ethical (FoI, Data Protection)
- Advocacy (data informatics)
- Data repositories
- Data storage
- Data analysis
- Data visualisation
- Data mining
- Data modelling
- Data licensing
- Training….

University data roles?

- •Roles (7 listed)
- •Responsibilities
- •Requirements
- •Relationships



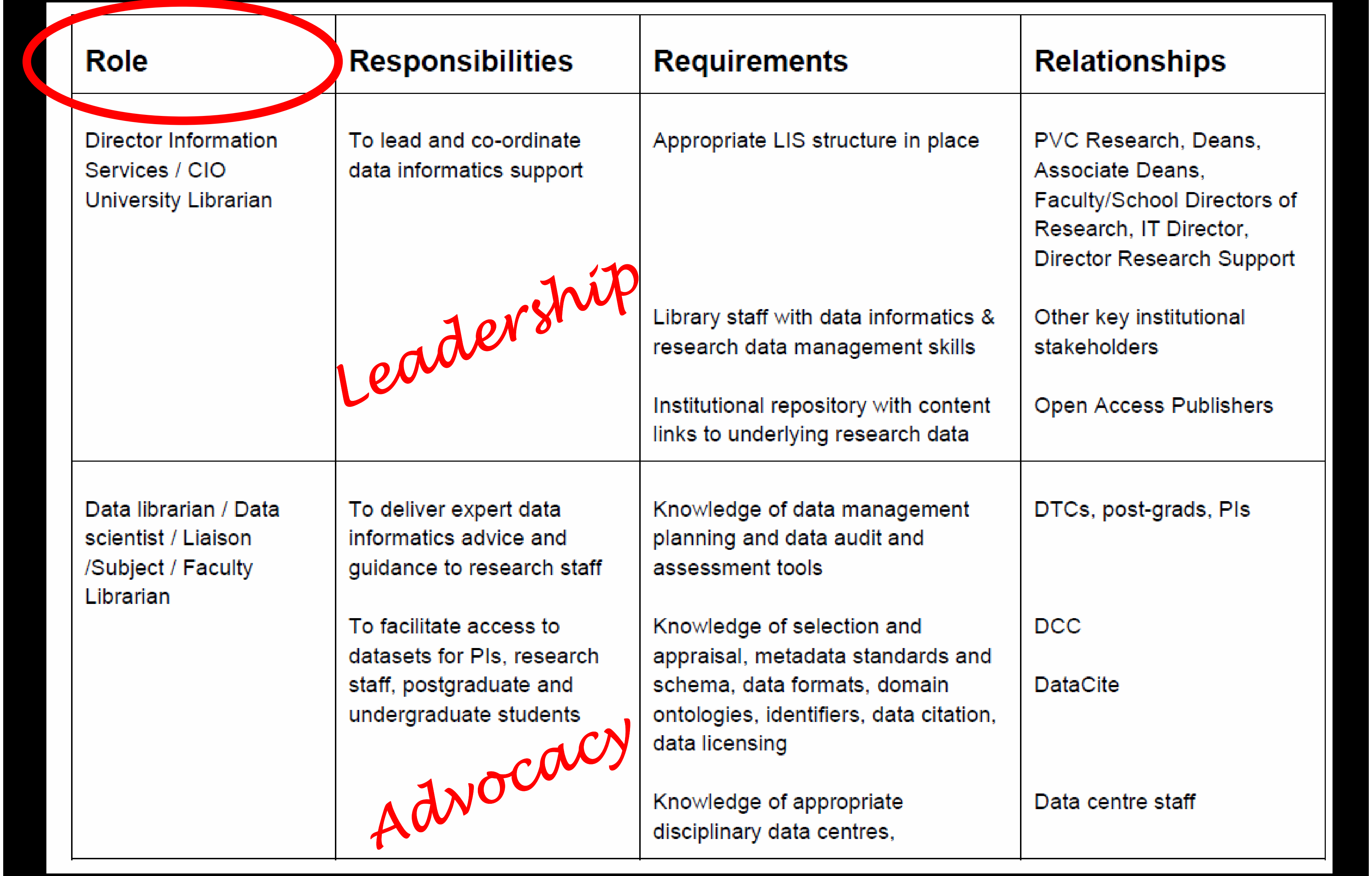
Role	Responsibilities	Requirements	Relationships
Director Information Services / CIO University Librarian	To lead and co-ordinate data informatics support	Appropriate LIS structure in place Library staff with data informatics & research data management skills Institutional repository with content links to underlying research data	PVC Research, Deans, Associate Deans, Faculty/School Directors of Research, IT Director, Director Research Support Other key institutional stakeholders Open Access Publishers
Data librarian / Data scientist / Liaison /Subject / Faculty Librarian	To deliver expert data informatics advice and guidance to research staff To facilitate access to datasets for PIs, research staff, postgraduate and undergraduate students	Knowledge of data management planning and data audit and assessment tools Knowledge of selection and appraisal, metadata standards and schema, data formats, domain ontologies, identifiers, data citation, data licensing Knowledge of appropriate disciplinary data centres,	DTCs, post-grads, PIs DCC DataCite Data centre staff

Liz Lyon, Informatics Transform, IJDC Current Issue, 2012

1. Director IS/CIO/University Librarian
2. Data librarians /data scientist /liaison/subject/faculty librarians
3. Repository managers
4. IT/Computing Services
5. Research Support/Innovation Office
6. Doctoral Training Centres
7. PVC Research
8. *+ Public Engagement Office*

Liz Lyon, Informatics Transform,
IJDC Current Issue, 2012

Data roles



Role	Responsibilities	Requirements	Relationships
Director Information Services / CIO University Librarian	To lead and co-ordinate data informatics support	Appropriate LIS structure in place Library staff with data informatics & research data management skills Institutional repository with content links to underlying research data	PVC Research, Deans, Associate Deans, Faculty/School Directors of Research, IT Director, Director Research Support Other key institutional stakeholders Open Access Publishers
Data librarian / Data scientist / Liaison /Subject / Faculty Librarian	To deliver expert data informatics advice and guidance to research staff To facilitate access to datasets for PIs, research staff, postgraduate and undergraduate students	Knowledge of data management planning and data audit and assessment tools Knowledge of selection and appraisal, metadata standards and schema, data formats, domain ontologies, identifiers, data citation, data licensing Knowledge of appropriate disciplinary data centres,	DTCs, post-grads, PIs DCC DataCite Data centre staff

Leadership

Advocacy

Full mapping : Informatics Transform, IJDC Current issue, 2012

April 2011 - EPSRC Letter to VCs

EPSRC

Pioneering research and skills

Engineering and Physical Sciences Research Council

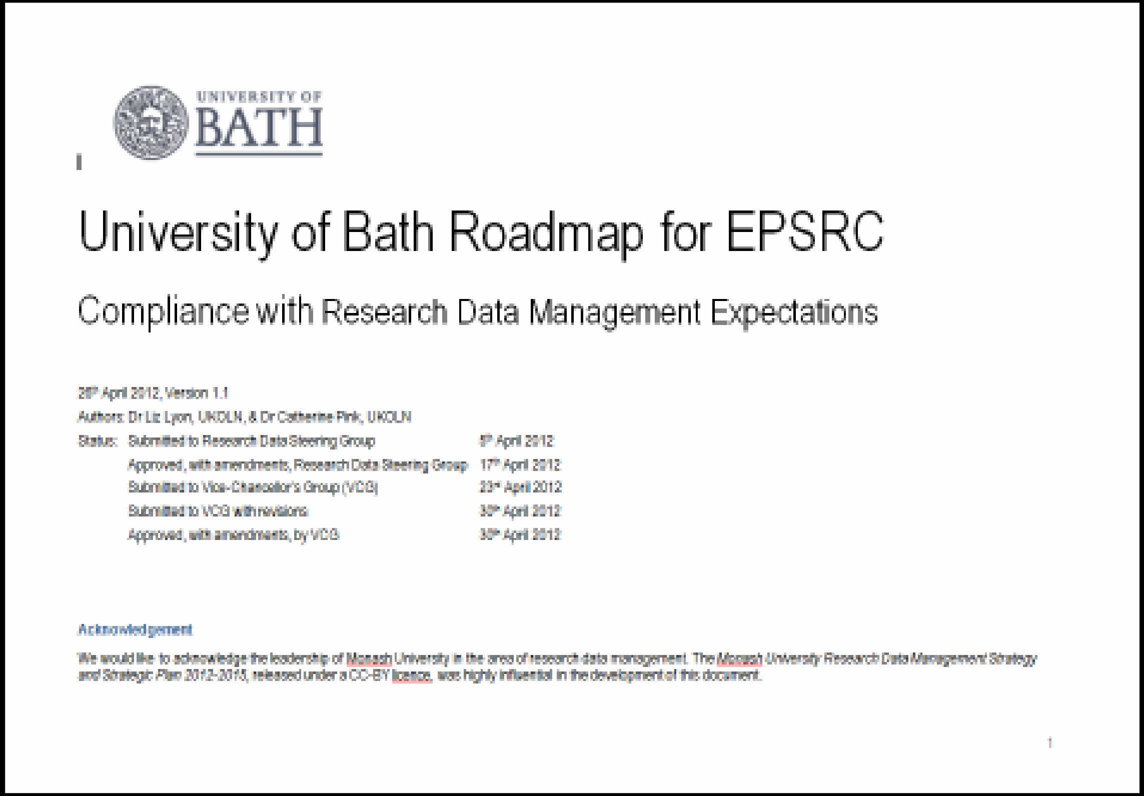
EPSRC expects all those institutions it funds

- to develop **a roadmap** that aligns their policies and processes with EPSRC's **expectations** by **1st May 2012**;
- to be fully compliant with these **expectations** by **1st May 2015**.

http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx

- Awareness of regulatory environment
- Data access statement
- Data policies and processes
- Data storage
- Structured metadata descriptions
- DOIs for data
- Data securely preserved for a minimum of 10 years

- Leadership
- Co-ordination
- Pan-institutional perspective
- Operational plan
- Wider strategic alignment



UNIVERSITY OF BATH

University of Bath Roadmap for EPSRC

Compliance with Research Data Management Expectations

26th April 2012, Version 1.1

Authors: Dr Liz Lyon, UKOLN, & Dr Catherine Pink, UKOLN

Status:	Submitted to Research Data Steering Group	5th April 2012
	Approved, with amendments, Research Data Steering Group	17th April 2012
	Submitted to Vice-Chancellor's Group (VCG)	23rd April 2012
	Submitted to VCG with revisions	30th April 2012
	Approved, with amendments, by VCG	30th April 2012

Acknowledgement

We would like to acknowledge the leadership of Monash University in the area of research data management. The *Monash University Research Data Management Strategy and Strategic Plan 2012-2015*, released under a CC-BY licence, was highly influential in the development of this document.

Advocacy and support

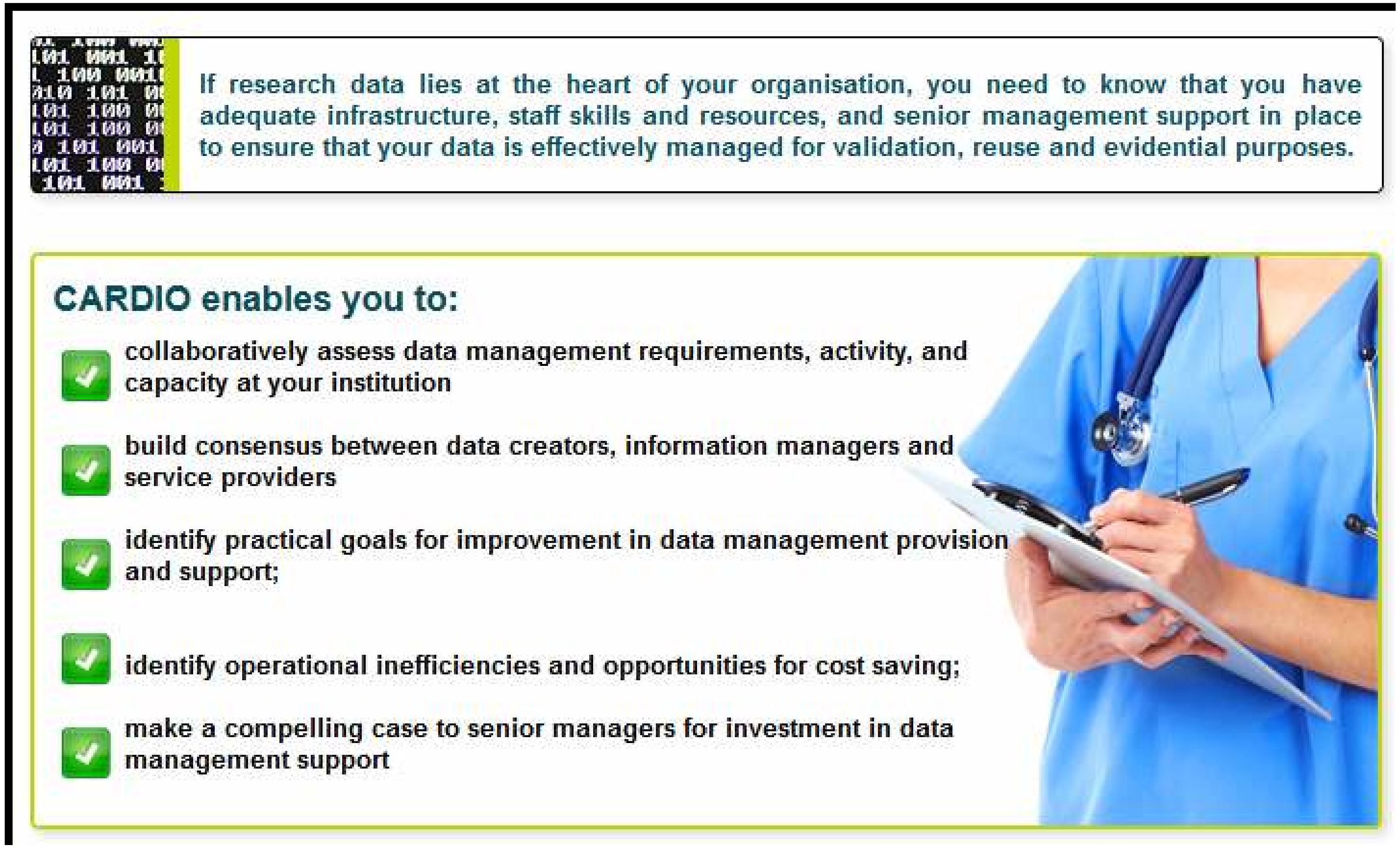
- ***Data requirements:*** legacy data
- ***Data management plans***: tools
- ***Informatics***: disciplinary metadata schema, standards, formats, identifiers, ontologies
- ***Citation:*** links to publications
- ***Reuse***: tracking your data

Understanding Data Requirements

Data Asset Framework

CARDIO

collaborative assessment of research data infrastructure and objectives



If research data lies at the heart of your organisation, you need to know that you have adequate infrastructure, staff skills and resources, and senior management support in place to ensure that your data is effectively managed for validation, reuse and evidential purposes.

CARDIO enables you to:

- ✓ collaboratively assess data management requirements, activity, and capacity at your institution
- ✓ build consensus between data creators, information managers and service providers
- ✓ identify practical goals for improvement in data management provision and support;
- ✓ identify operational inefficiencies and opportunities for cost saving;
- ✓ make a compelling case to senior managers for investment in data management support

DCC

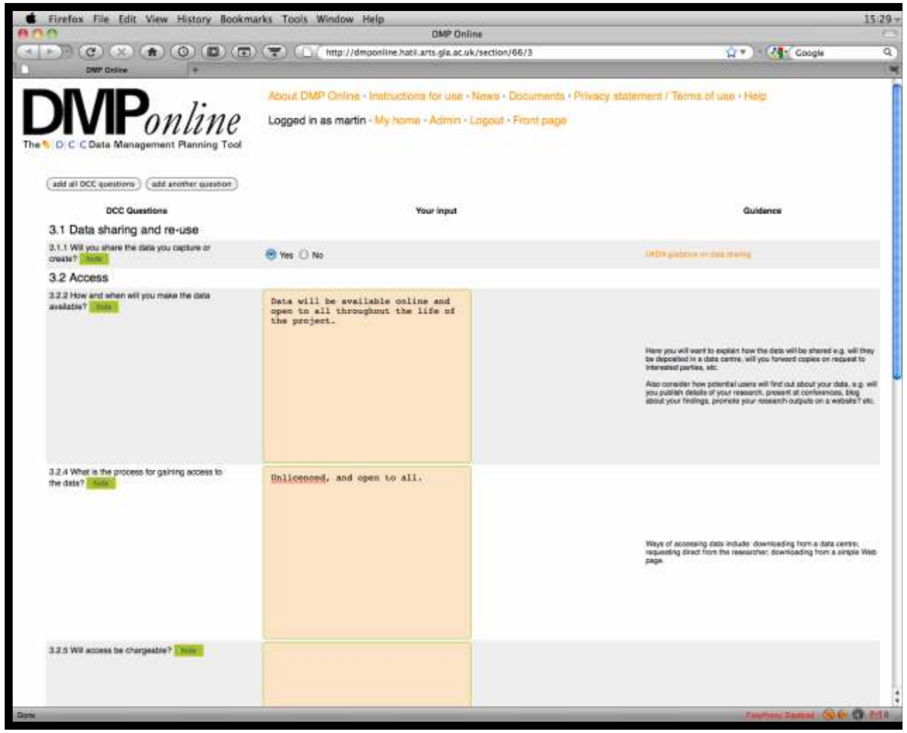
http://www.dcc.ac.uk/

DCC

Data management plans

DMPonline

The DCC Data Management Planning Tool


Firefox File Edit View History Bookmarks Tools Window Help
15:29
DMP Online
http://dmponline.hatii.arts.gla.ac.uk/section/66/3
Google
DMP Online
About DMP Online · Instructions for use · News · Documents · Privacy statement / Terms of use · Help
My home · Admin · Logout · Front page
add all DCC questions
add another question
hide
Done

DMPonline

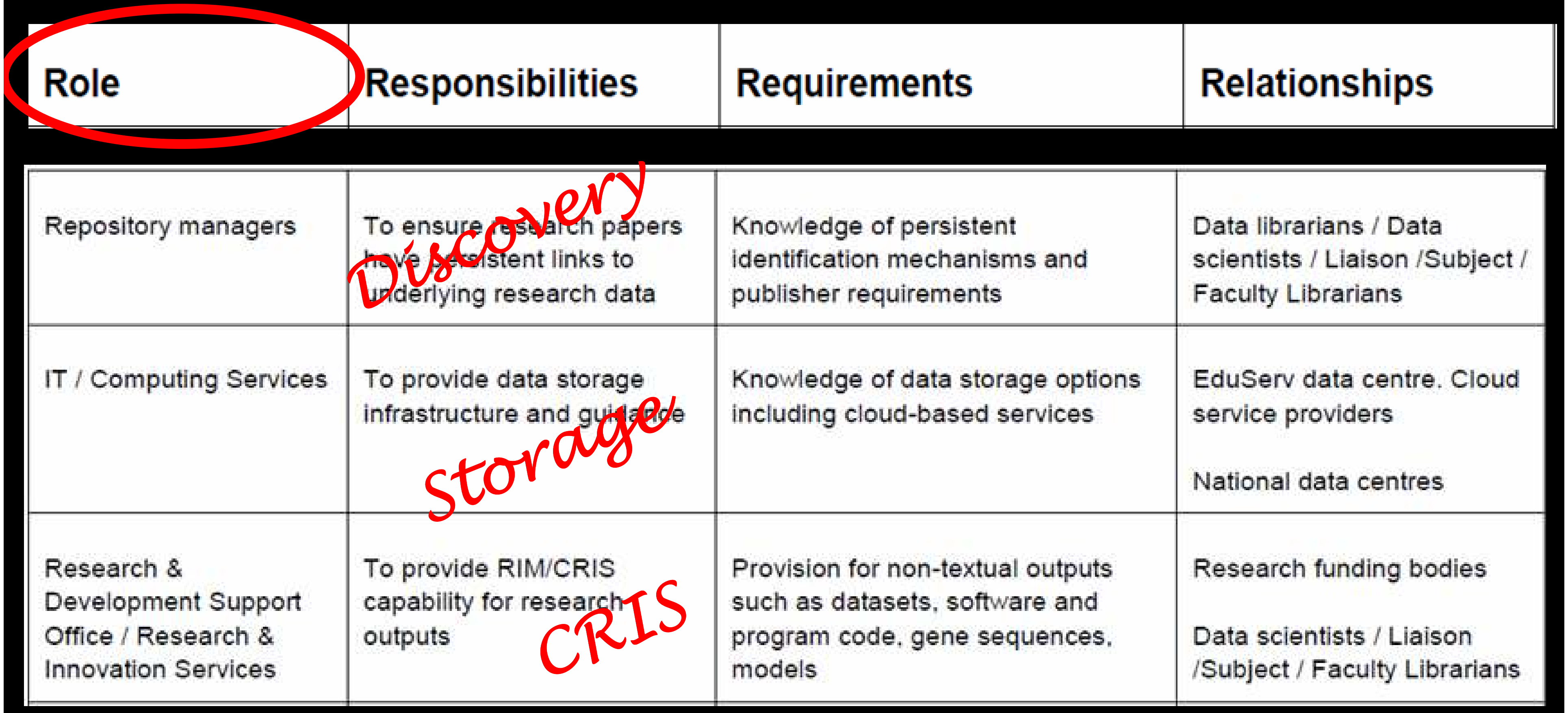
The DCC Data Management Planning Tool

Logged in as martin

DCC Questions	Your input	Guidance
3.1 Data sharing and re-use		
3.1.1 Will you share the data you capture or create?	● Yes ○ No	UKDA guidance on data sharing
3.2 Access		
3.2.2 How and when will you make the data available?	Data will be available online and open to all throughout the life of the project.	Here you will want to explain how the data will be shared e.g. will they be deposited in a data centre, will you forward copies on request to interested parties, etc. Also consider how potential users will find out about your data, e.g. will you publish details of your research, present at conferences, blog about your findings, promote your research outputs on a website? etc.
3.2.4 What is the process for gaining access to the data?	Unlicenced, and open to all.	Ways of accessing data include: downloading from a data centre; requesting direct from the researcher; downloading from a simple Web page.
3.2.5 Will access be chargeable?		

DMPTool

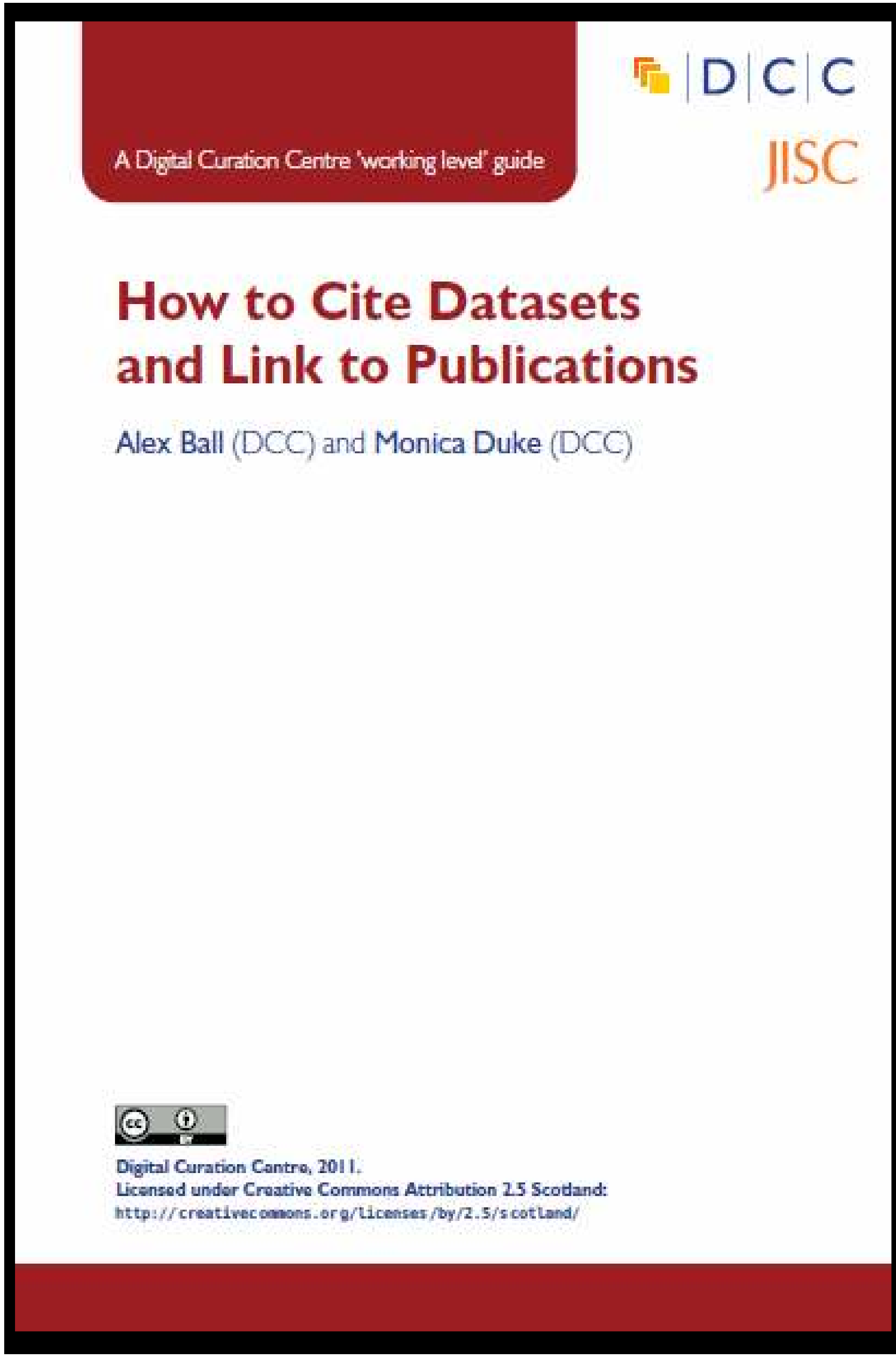
Guidance and Resources for your Data Management Plan



Role	Responsibilities	Requirements	Relationships
Repository managers	To ensure research papers have persistent links to underlying research data Discovery	Knowledge of persistent identification mechanisms and publisher requirements	Data librarians / Data scientists / Liaison /Subject / Faculty Librarians
IT / Computing Services	To provide data storage infrastructure and guidance Storage	Knowledge of data storage options including cloud-based services	EduServ data centre. Cloud service providers National data centres
Research & Development Support Office / Research & Innovation Services	To provide RIM/CRIS capability for research outputs CRIS	Provision for non-textual outputs such as datasets, software and program code, gene sequences, models	Research funding bodies Data scientists / Liaison /Subject / Faculty Librarians

Full mapping : Informatics Transform, IJDC Current issue, 2012

How to cite data



A Digital Curation Centre 'working level' guide

DCC

JISC

How to Cite Datasets and Link to Publications

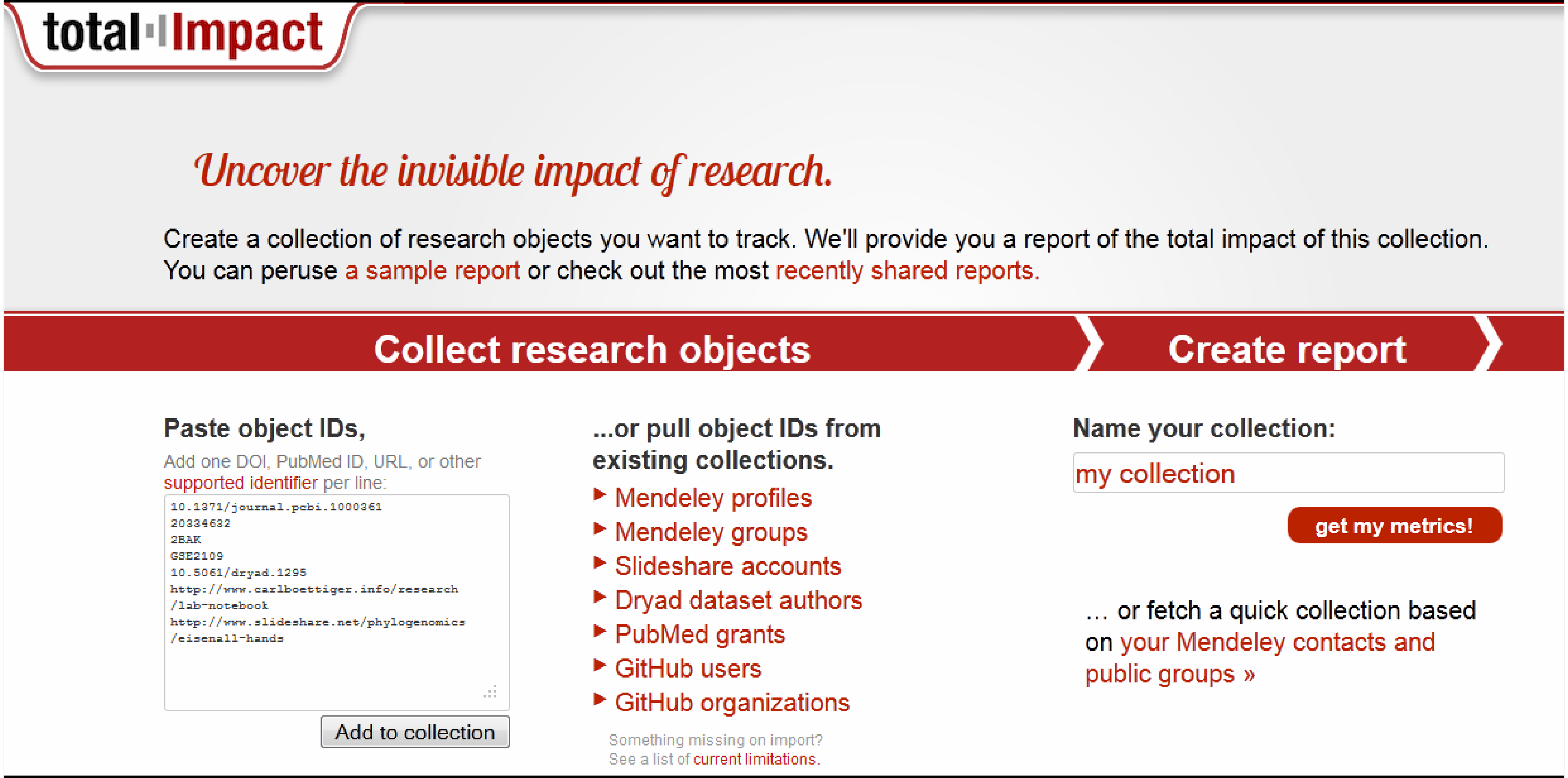
Alex Ball (DCC) and Monica Duke (DCC)

DataCite

Helping you to find, access, and reuse data

Using DOIs

How to track impact



Uncover the invisible impact of research.

Create a collection of research objects you want to track. We'll provide you a report of the total impact of this collection. You can peruse a sample report or check out the most recently shared reports.

Collect research objects

Create report

Paste object IDs,

Add one DOI, PubMed ID, URL, or other supported identifier per line:

```
10.1371/journal.pcbi.1000361
20334632
2BAK
GSE2109
10.5061/dryad.1295
http://www.carlboettiger.info/research
/lab-notebook
http://www.slideshare.net/phylogenomics
/eisenall-hands
```

...or pull object IDs from existing collections.

- Mendeley profiles
- Mendeley groups
- Slideshare accounts
- Dryad dataset authors
- PubMed grants
- GitHub users
- GitHub organizations

Something missing on import?
See a list of current limitations.

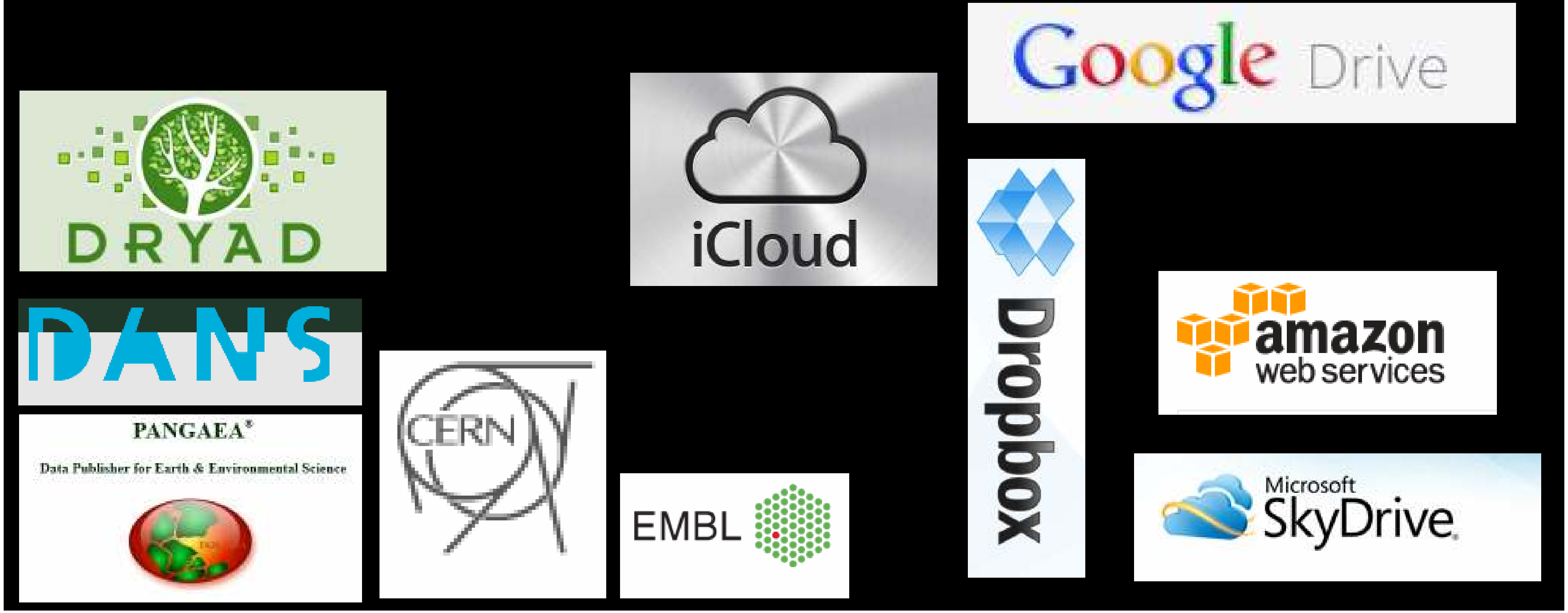
Name your collection:

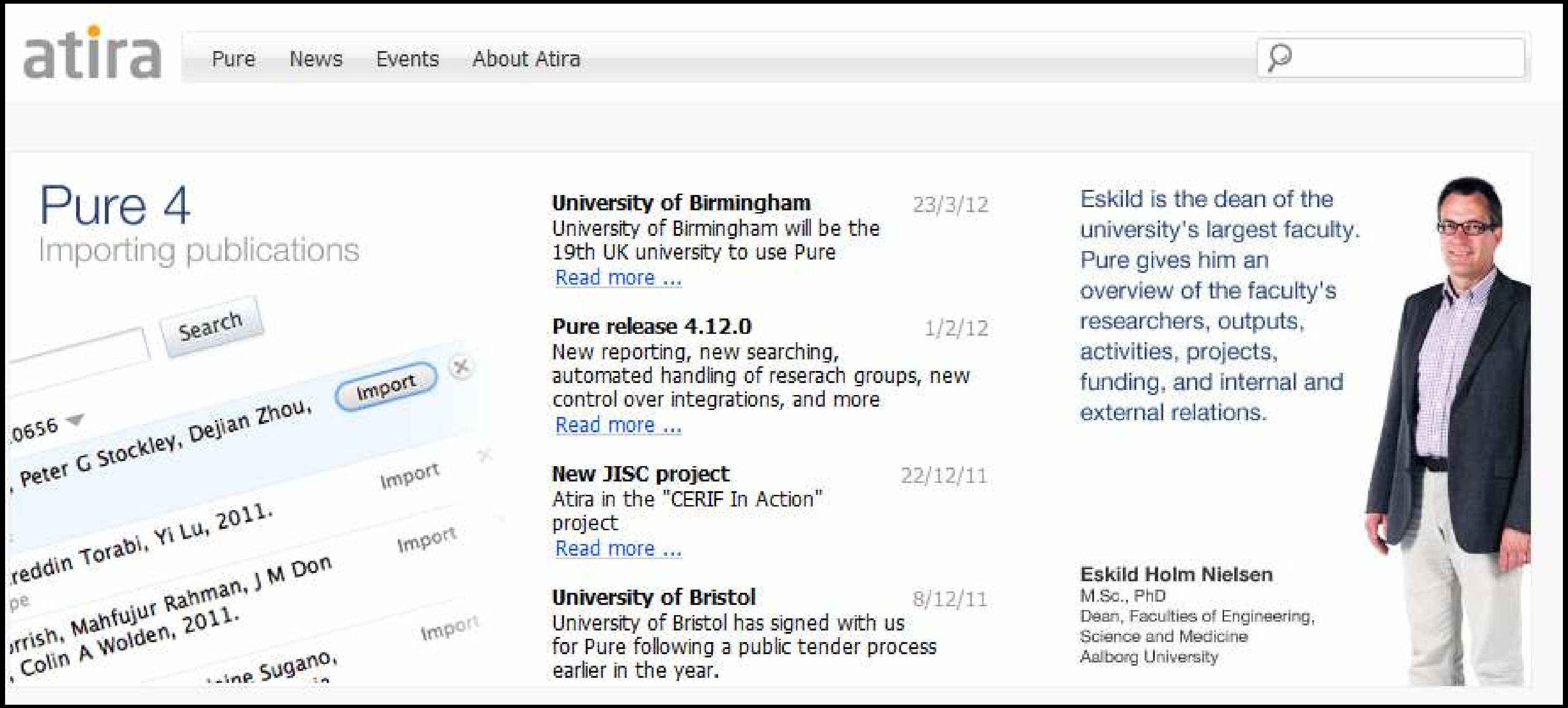
my collection

… or fetch a quick collection based on your Mendeley contacts and public groups »

http://total-impact.org/

- **Storage:** file-store, cloud, data centres, funder policy
- **Access:** embargoes, FoI





Pure 4

Importing publications

University of Birmingham 23/3/12
University of Birmingham will be the 19th UK university to use Pure
Read more ...

Pure release 4.12.0 1/2/12
New reporting, new searching, automated handling of reserach groups, new control over integrations, and more
Read more ...

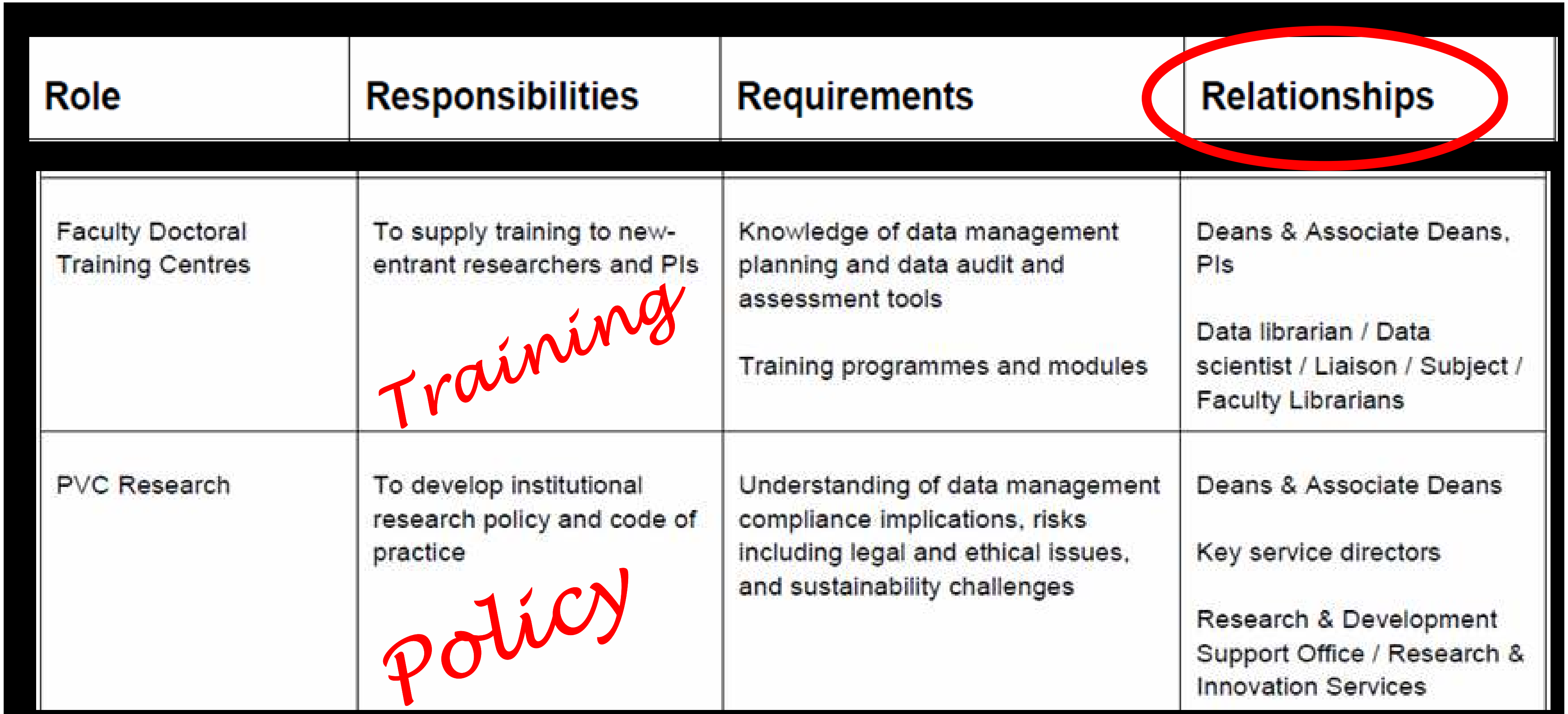
New JISC project 22/12/11
Atira in the "CERIF In Action" project
Read more ...

University of Bristol 8/12/11
University of Bristol has signed with us for Pure following a public tender process earlier in the year.

Eskild is the dean of the university's largest faculty. Pure gives him an overview of the faculty's researchers, outputs, activities, projects, funding, and internal and external relations.

Eskild Holm Nielsen
M.Sc., PhD
Dean, Faculties of Engineering, Science and Medicine
Aalborg University

CRIS integration, CERIF and data



Role	Responsibilities	Requirements	Relationships
Faculty Doctoral Training Centres	To supply training to new-entrant researchers and PIs	Knowledge of data management planning and data audit and assessment tools Training programmes and modules	Deans & Associate Deans, PIs Data librarian / Data scientist / Liaison / Subject / Faculty Librarians
PVC Research	To develop institutional research policy and code of practice	Understanding of data management compliance implications, risks including legal and ethical issues, and sustainability challenges	Deans & Associate Deans Key service directors Research & Development Support Office / Research & Innovation Services
Public Engagement Unit	To facilitate citizen participation in the research process	Understanding of open science methodologies and infrastructure	PVC Research Director, Communications Deans & Associate Deans, PIs The Media

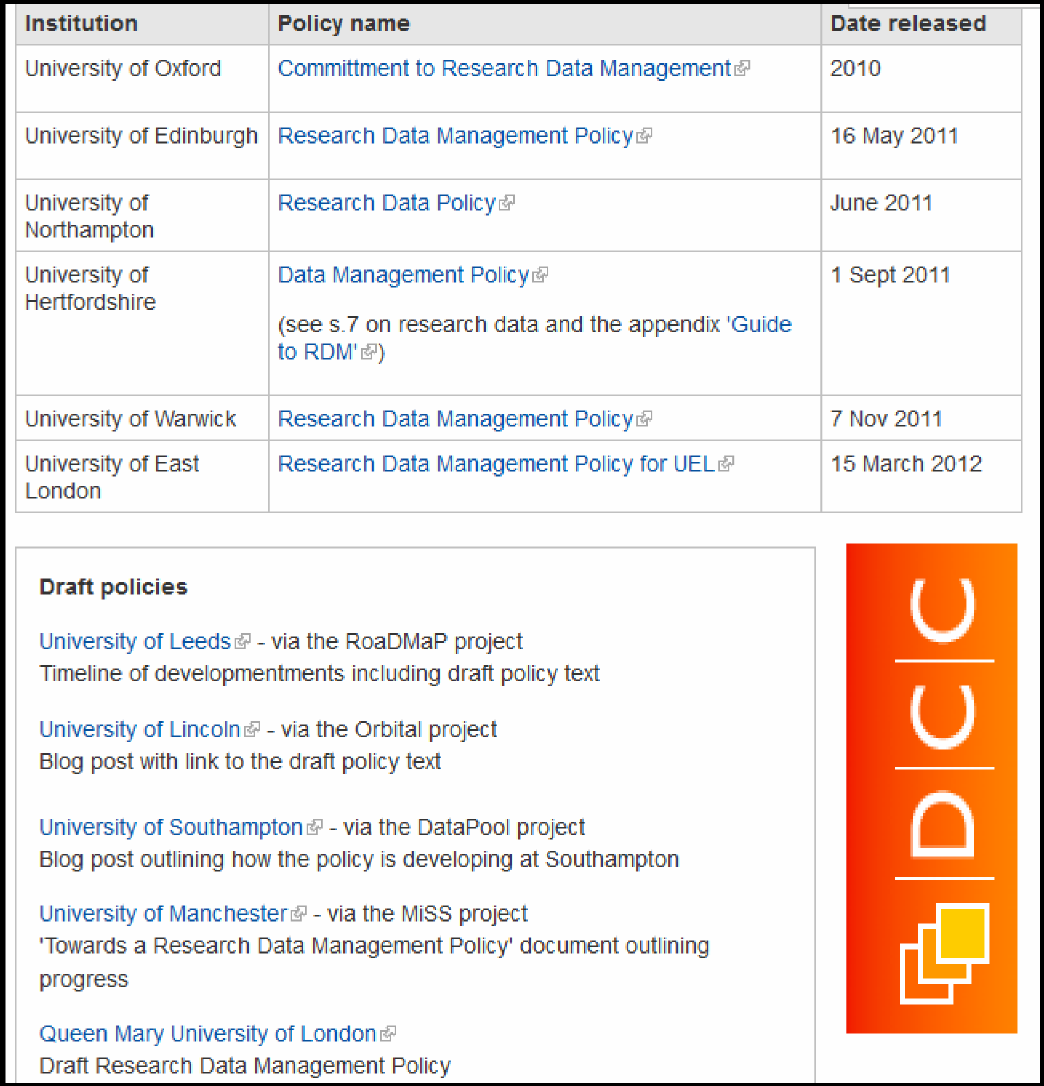
Training

Policy

Participation

Full mapping : Informatics Transform, IJDC Current issue, 2012

Institutional data policy development



Institution	Policy name	Date released
University of Oxford	Committment to Research Data Management	2010
University of Edinburgh	Research Data Management Policy	16 May 2011
University of Northampton	Research Data Policy	June 2011
University of Hertfordshire	Data Management Policy (see s.7 on research data and the appendix 'Guide to RDM')	1 Sept 2011
University of Warwick	Research Data Management Policy	7 Nov 2011
University of East London	Research Data Management Policy for UEL	15 March 2012

Draft policies

University of Leeds - via the RoaDMaP project
Timeline of developmentments including draft policy text

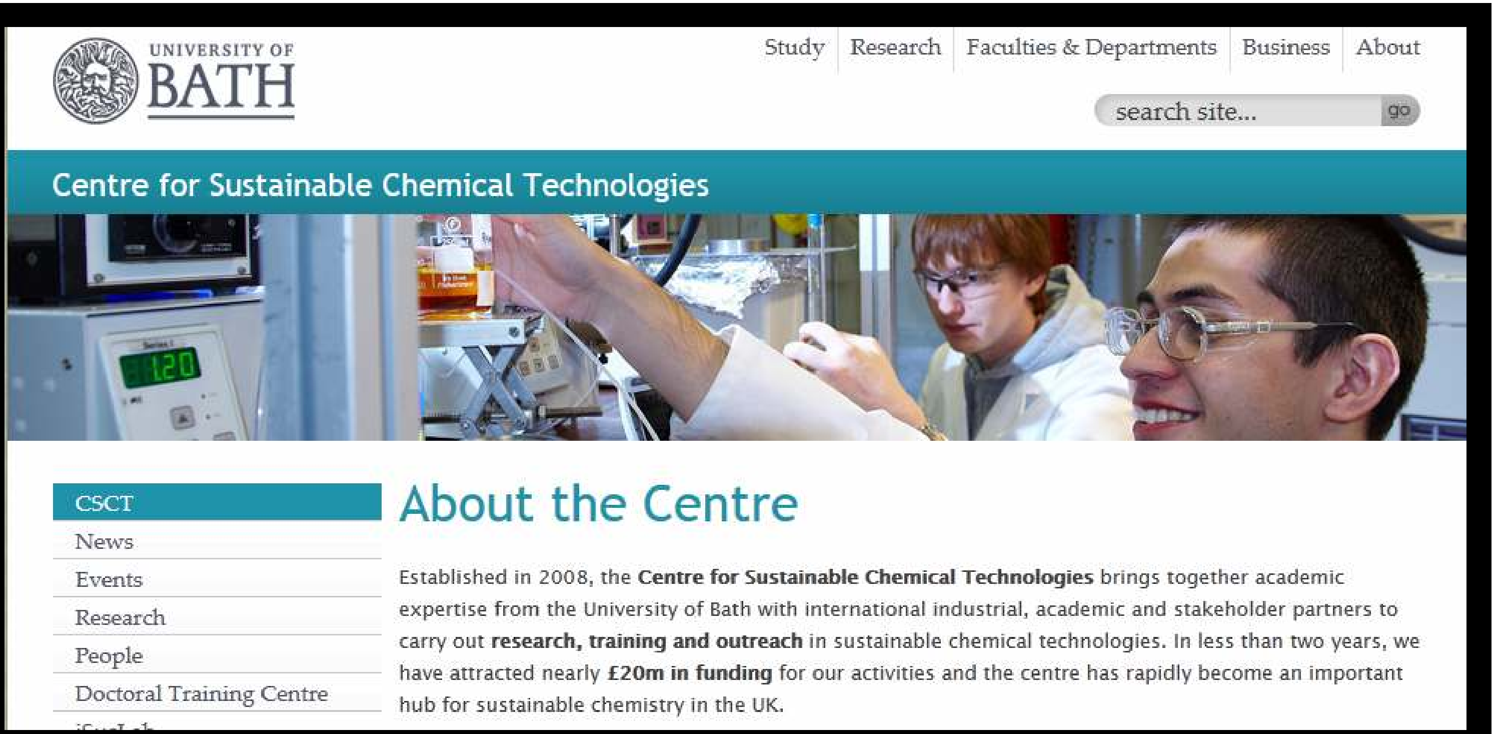
University of Lincoln - via the Orbital project
Blog post with link to the draft policy text

University of Southampton - via the DataPool project
Blog post outlining how the policy is developing at Southampton

University of Manchester - via the MiSS project
'Towards a Research Data Management Policy' document outlining progress

Queen Mary University of London
Draft Research Data Management Policy

- Aspirational?
- Pragmatic?
- Emergent?
- High-level?
- With teeth?

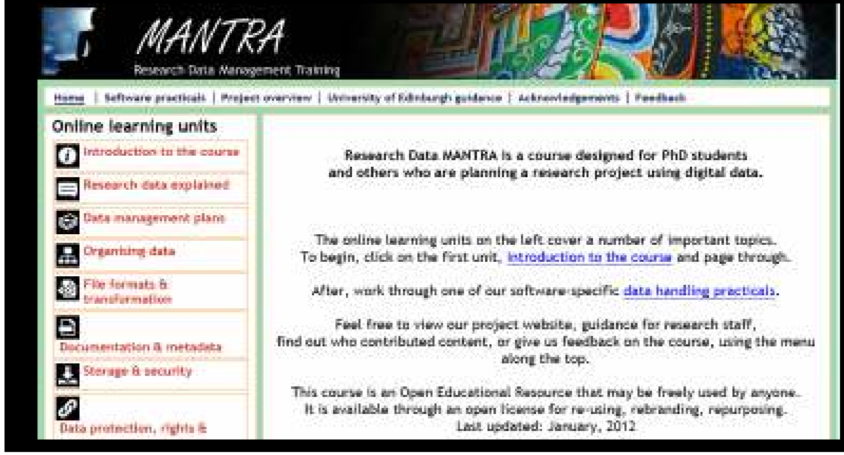


Centre for Sustainable Chemical Technologies

About the Centre

Established in 2008, the **Centre for Sustainable Chemical Technologies** brings together academic expertise from the University of Bath with international industrial, academic and stakeholder partners to carry out **research, training and outreach** in sustainable chemical technologies. In less than two years, we have attracted nearly **£20m in funding** for our activities and the centre has rapidly become an important hub for sustainable chemistry in the UK.

Doctoral Training Centres: Research360 Project @ Bath


MANTRA
Research Data Management Training
Home | Software practicals | Project overview | University of Edinburgh guidance | Acknowledgements | Feedback
Online learning units
Introduction to the course
Research data explained
Data management plans
Organising data
File formats & transformation
Documentation & metadata
Storage & security
Data protection, rights &

Research Data MANTRA is a course designed for PhD students and others who are planning a research project using digital data.

The online learning units on the left cover a number of important topics. To begin, click on the first unit, introduction to the course and page through.

After, work through one of our software-specific data handling practicals.

Feel free to view our project website, guidance for research staff, find out who contributed content, or give us feedback on the course, using the menu along the top.

This course is an Open Educational Resource that may be freely used by anyone. It is available through an open license for re-using, rebranding, repurposing.
Last updated: January, 2012

JISC projects

DCC resources

Digital Curation 101 training materials

If you are not able to attend DC 101 workshops and receive our training materials, why not buy a copy of training materials as a handy reference guide for your workplace?

- Leadership & co-ordination
- Strategy and planning
- Policy
- Legal and ethical (FoI, Data Protection)
- Advocacy (data informatics)
- Data repositories
- Data storage
- Data analysis
- Data visualisation
- Data mining
- Data modelling
- Data licensing
- Training….

Gaps? Opportunities??

Analyse LIS entry qualifications & increase STEM entrants

Target

- Biologists
- Chemists
- Mathematicians

Lyon, Informatics Transform, IJDC 2012

Gaps? Opportunities??

Define core components of data informatics and data science

- Metadata (discovery, preservation)
- Domain ontologies
- Visualisation e.g. VisTrails
- Workflow e.g. Taverna
- Analysis e.g. R

Lyon, Informatics Transform, IJDC 2012

http://www.flickr.com/photos/50542505@N08/5723947474/

Data scientist flavours?

- Analysis, mining, modelling
- Informatics, advocacy, training
- Repositories, preservation
- Visualisation, simulations

Infrastructure, Intelligence, Innovation: driving the Data Science agenda

8th International Digital Curation Conference, Amsterdam, 14-16 January 2013

Thank you!

Informatics Transform article

http://www.ijdc.net/index.php/ijdc/article/view/210

Slides

http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/presentations.html

DCC http://www.dcc.ac.uk