From data management to data science: current trends & future challenges

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

DST Seminar, Pretoria South Africa, November 2012

UKOLN

DCC

UKOLN is supported by:

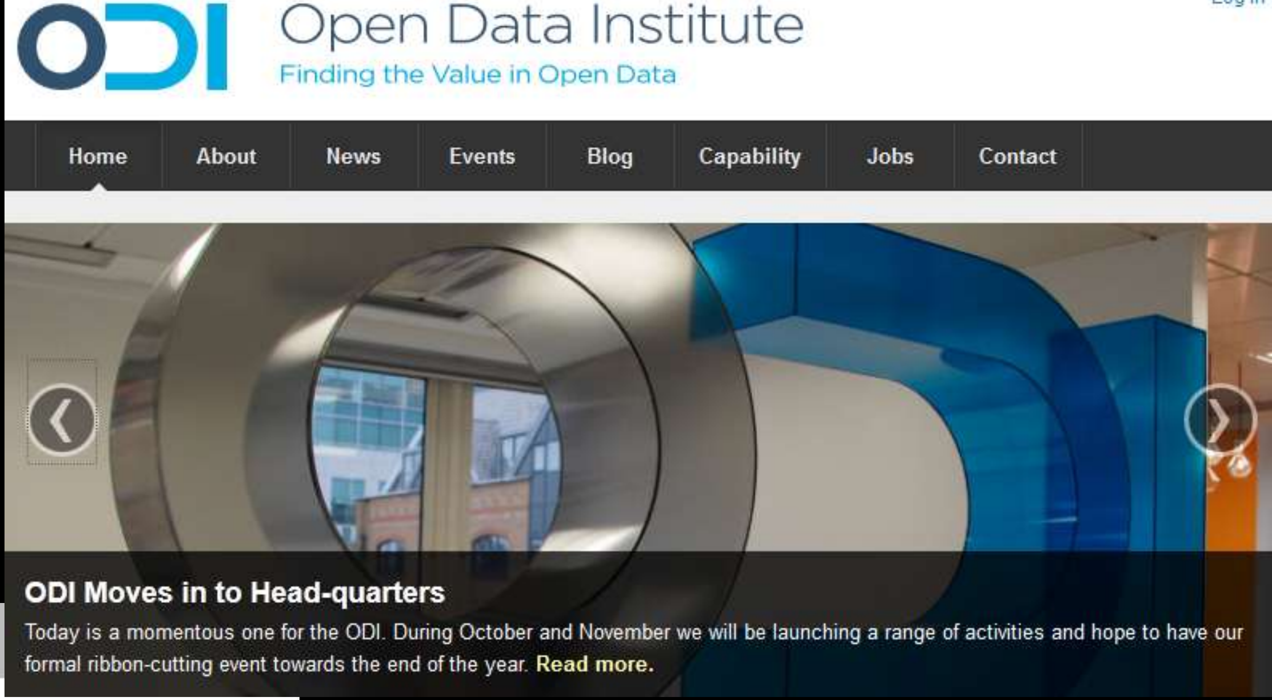
JISC Microsoft Research

UNIVERSITY OF BATH

THE SUNDAY TIMES UNIVERSITY OF THE YEAR 2011-12

Overview

- Trends, Launches, Readiness
- UK Social science data
- Institutions, data and the DCC
  University of Bath experience
- Data scientists: a new breed



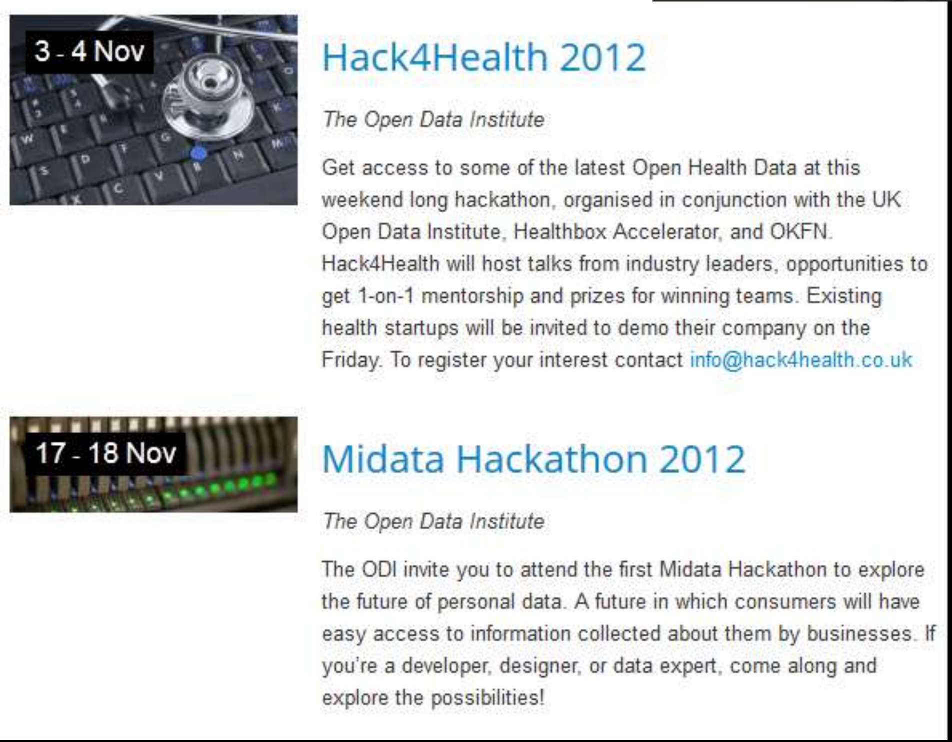
Open Data Institute

Finding the Value in Open Data

ODI Moves in to Head-quarters

Today is a momentous one for the ODI. During October and November we will be launching a range of activities and hope to have our formal ribbon-cutting event towards the end of the year. **Read more.**

Upcoming Events



3 - 4 Nov

Hack4Health 2012

The Open Data Institute

Get access to some of the latest Open Health Data at this weekend long hackathon, organised in conjunction with the UK Open Data Institute, Healthbox Accelerator, and OKFN. Hack4Health will host talks from industry leaders, opportunities to get 1-on-1 mentorship and prizes for winning teams. Existing health startups will be invited to demo their company on the Friday. To register your interest contact info@hack4health.co.uk

17 - 18 Nov

Midata Hackathon 2012

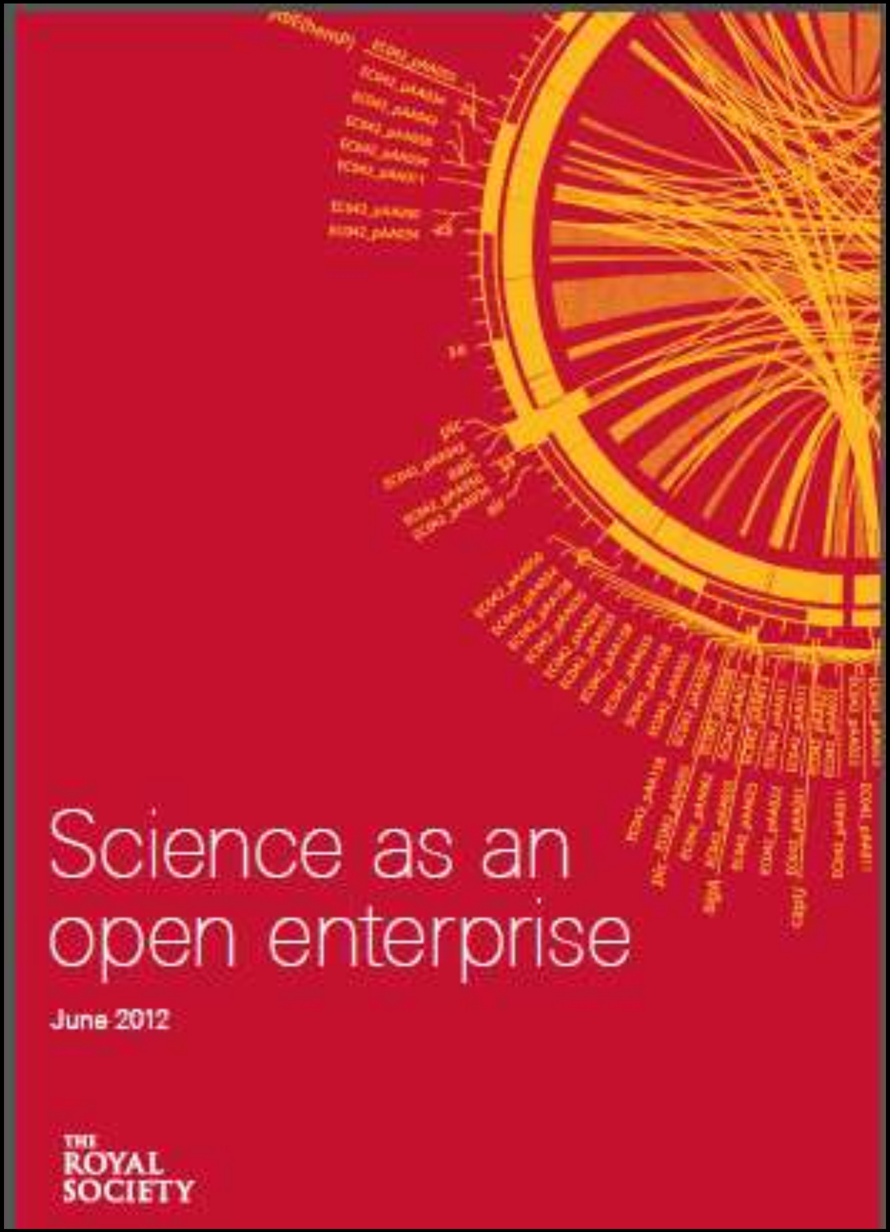
The Open Data Institute

The ODI invite you to attend the first Midata Hackathon to explore the future of personal data. A future in which consumers will have easy access to information collected about them by businesses. If you're a developer, designer, or data expert, come along and explore the possibilities!

Open Data Institute

Finding the Value in Open Data

UK government supports open data



Science as an open enterprise

June 2012

THE ROYAL SOCIETY

Royal Society Report

Science as an Open Enterprise

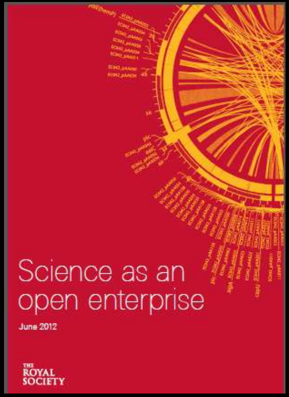
June 2012

10 Recommendations

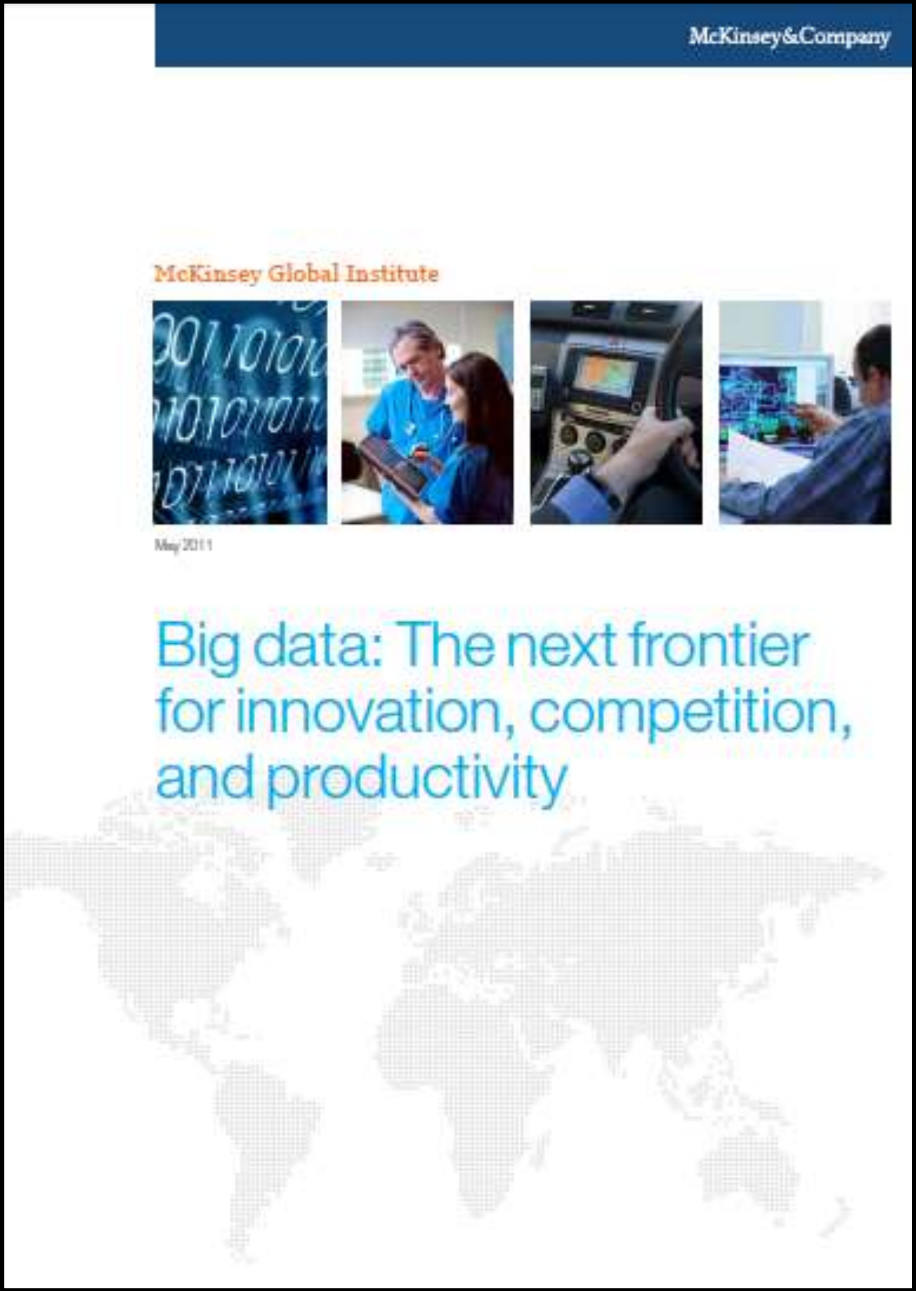
Cites the DCC

http://royalsociety.org/uploadedFiles/Royal_Society_Content/policy/projects/sape/2012-06-20-SAOE.pdf

Report sound-bytes



- “intelligently open data”
- “Scientists ….. are increasingly turning to their university libraries and institutional repositories for support for their data…..”
- “familiarity with … tools and principles of data management should be an integral part of the training of scientists in the future…. “
- “The skills of data scientists are crucial in supporting the data management needs of researchers and of institutions.”



McKinsey&Company

McKinsey Global Institute

May 2011

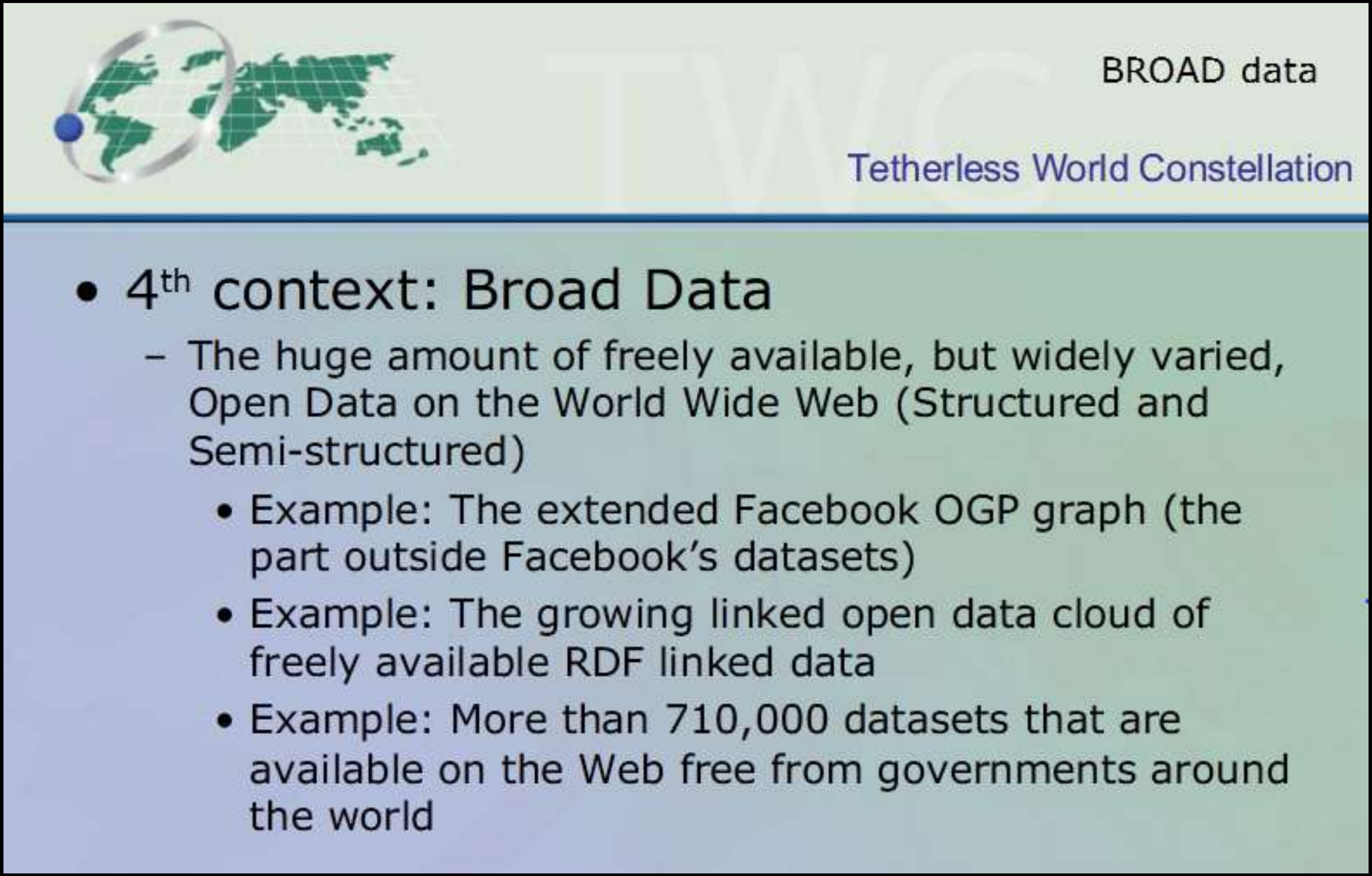
Big data: The next frontier for innovation, competition, and productivity

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

Predicts a shortage of 190,000 data scientists by 2019

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

Data : from Big to Broad (Jim Hendler)



BROAD data

Tetherless World Constellation

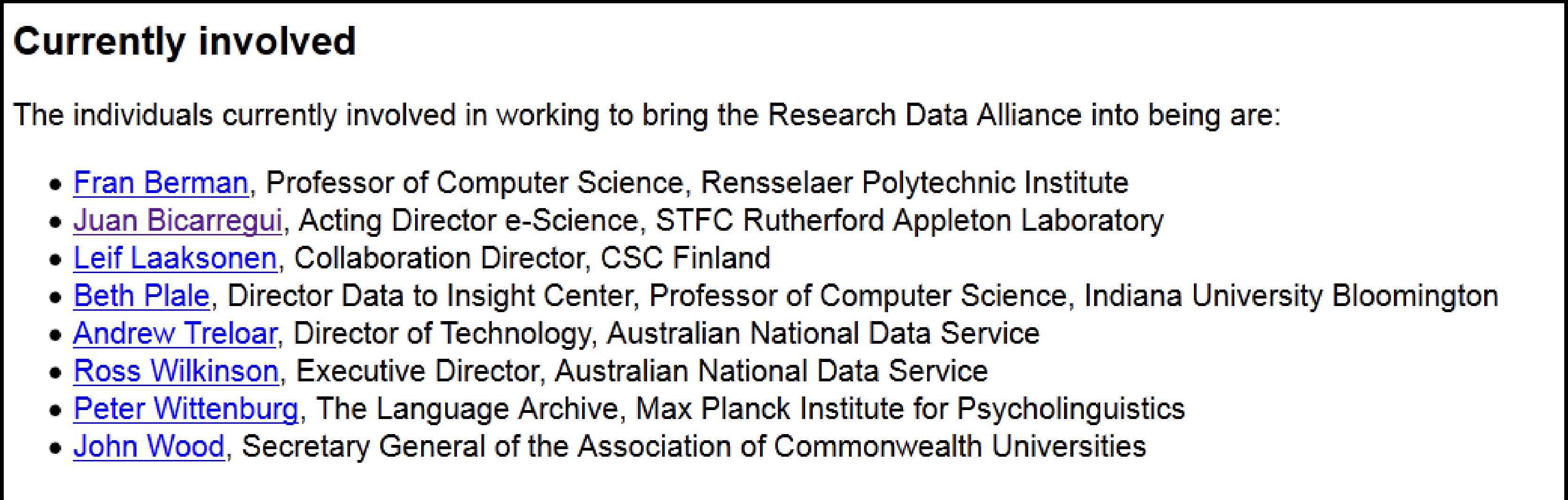
- 4th context: Broad Data
  - The huge amount of freely available, but widely varied, Open Data on the World Wide Web (Structured and Semi-structured)
    - Example: The extended Facebook OGP graph (the part outside Facebook's datasets)
    - Example: The growing linked open data cloud of freely available RDF linked data
    - Example: More than 710,000 datasets that are available on the Web free from governments around the world

Research Data Alliance

Vision

Researchers around the world sharing and using research data without barriers

http://rd-alliance.org//



Currently involved

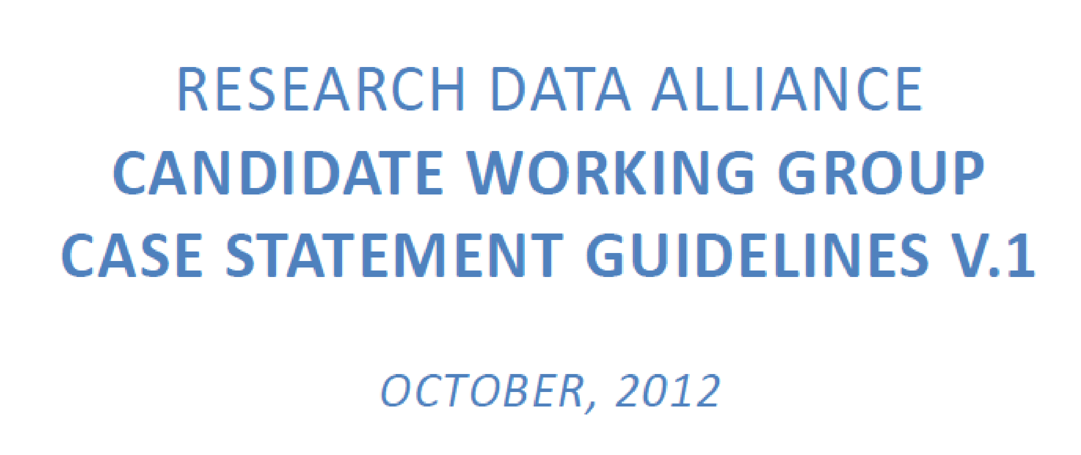
The individuals currently involved in working to bring the Research Data Alliance into being are:

- Fran Berman, Professor of Computer Science, Rensselaer Polytechnic Institute
- Juan Bicarregui, Acting Director e-Science, STFC Rutherford Appleton Laboratory
- Leif Laaksonen, Collaboration Director, CSC Finland
- Beth Plale, Director Data to Insight Center, Professor of Computer Science, Indiana University Bloomington
- Andrew Treloar, Director of Technology, Australian National Data Service
- Ross Wilkinson, Executive Director, Australian National Data Service
- Peter Wittenburg, The Language Archive, Max Planck Institute for Psycholinguistics
- John Wood, Secretary General of the Association of Commonwealth Universities

RD-Alliance

forum

RD-Alliance Launch March 2013, Gothenburg, Sweden

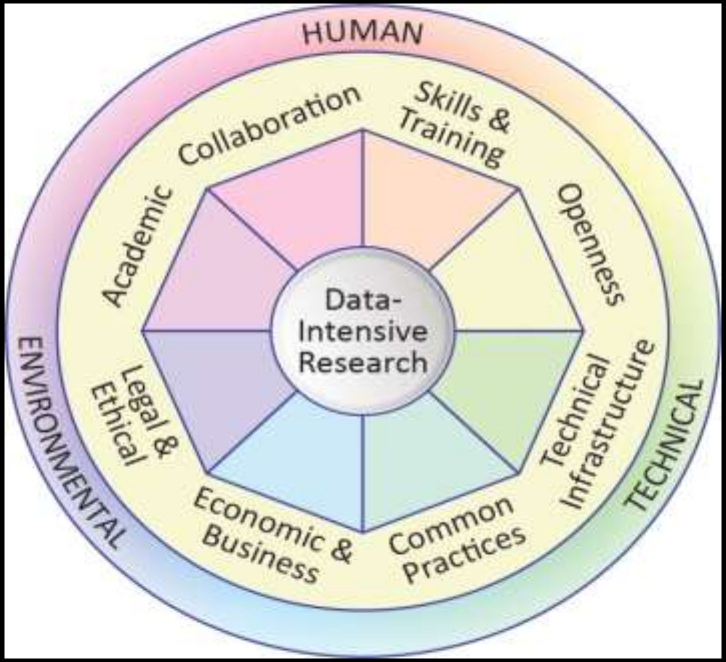


RESEARCH DATA ALLIANCE

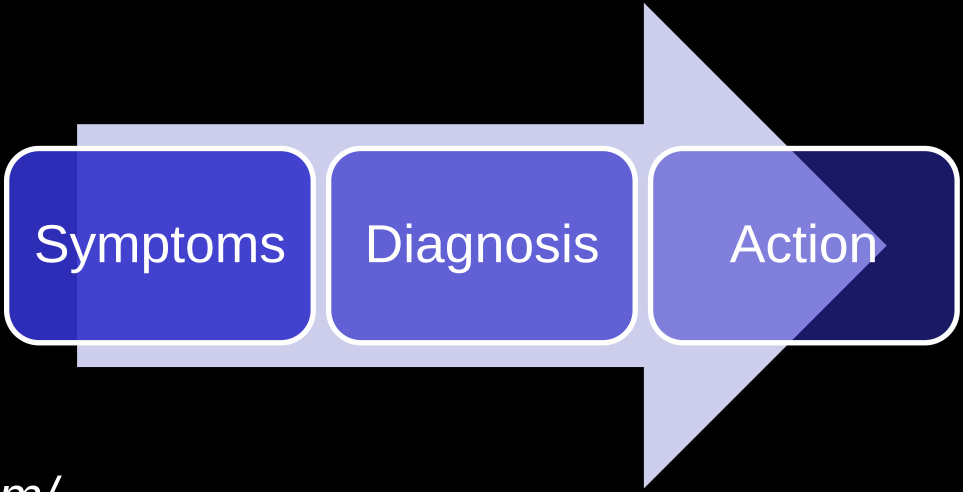
CANDIDATE WORKING GROUP CASE STATEMENT GUIDELINES V.1

OCTOBER, 2012

Community Capability Model Framework CCMF



- Understanding the data habitat :
- Funder, institution, PI



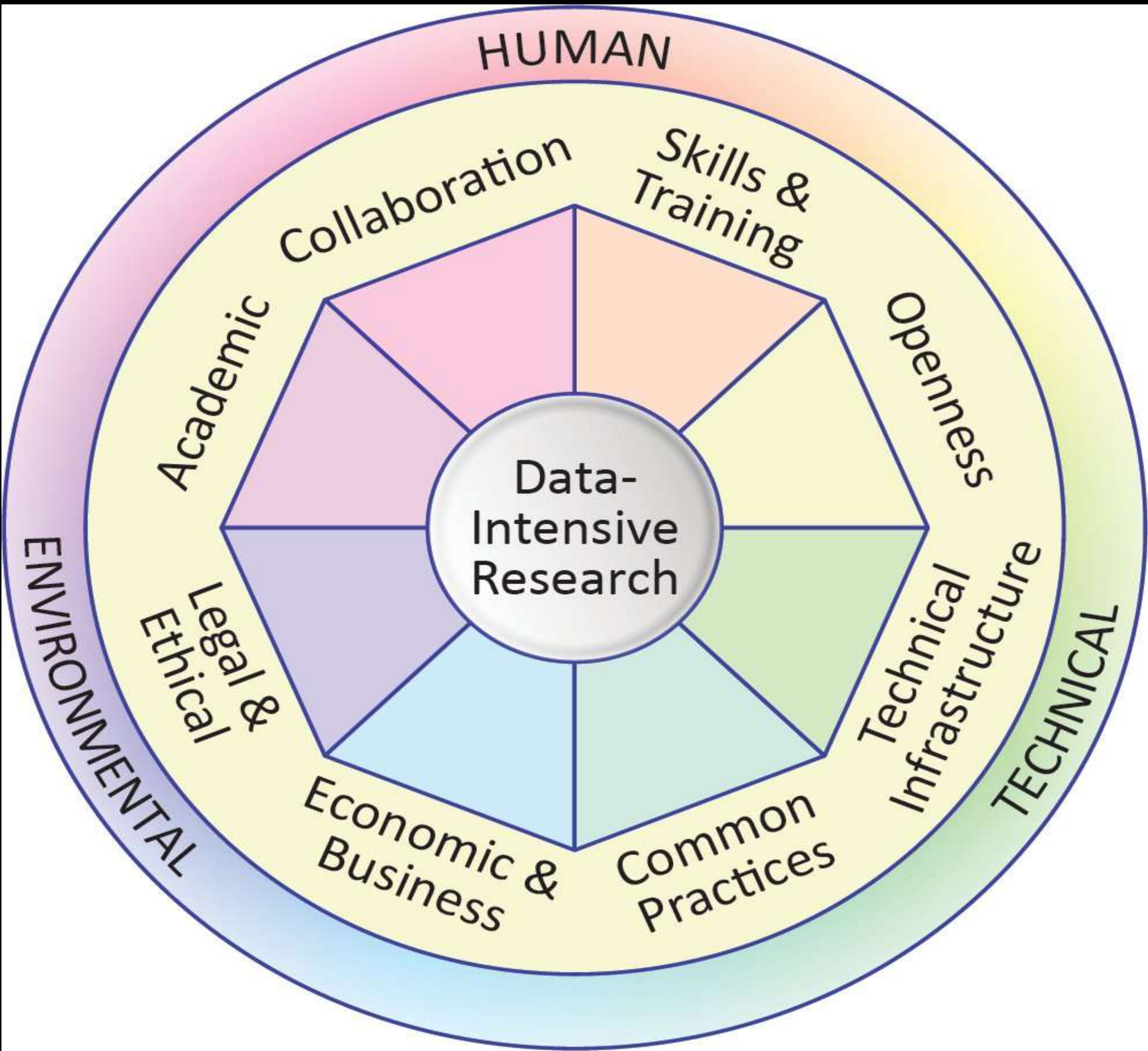
Symptoms

Diagnosis

Action

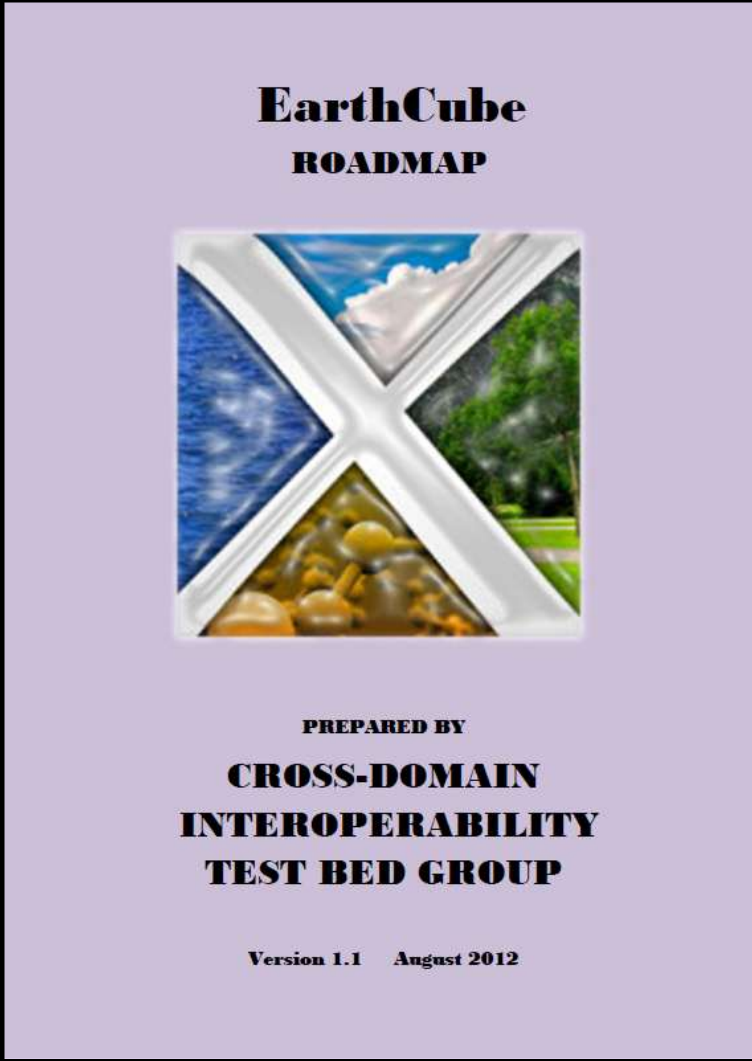
http://communitymodel.sharepoint.com/

CCMF 8 Capability Factors



Applying CCMF: geosciences

- Disciplinary readiness for Cyber-Infrastructure (CI)
- NSF geoscience initiative
- Cross-Domain Interoperability Roadmap
- Applied the CCMF model



EarthCube

ROADMAP

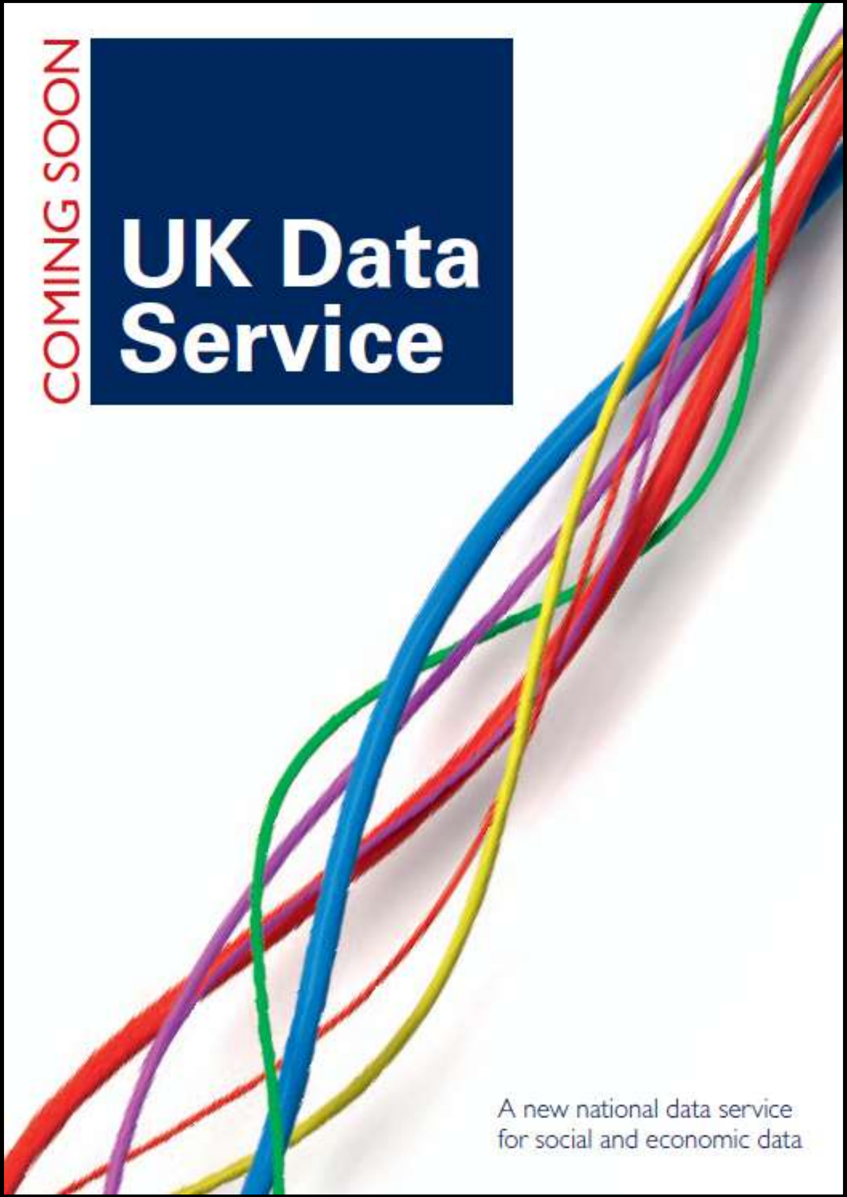
PREPARED BY

CROSS-DOMAIN INTEROPERABILITY TEST BED GROUP

Version 1.1 August 2012

https://www.dropbox.com/s/0oqk5ostahfokbg/interop_roadmap_master8_Aug16.pdf

UK social science: data ready



COMING SOON

UK Data Service

A new national data service for social and economic data

- UKDS Launch
- >40 yrs data archive
- Policy mandate to deposit datasets
- Admin data linkage

SECURE DATA SERVICE
enabling the research community

E·S·R·C ECONOMIC & SOCIAL RESEARCH COUNCIL

Economic and Social Research Council
Shaping Society

April 2011 - EPSRC Letter to VCs

EPSRC

Pioneering research and skills

Engineering and Physical Sciences Research Council

EPSRC expects all those institutions it funds

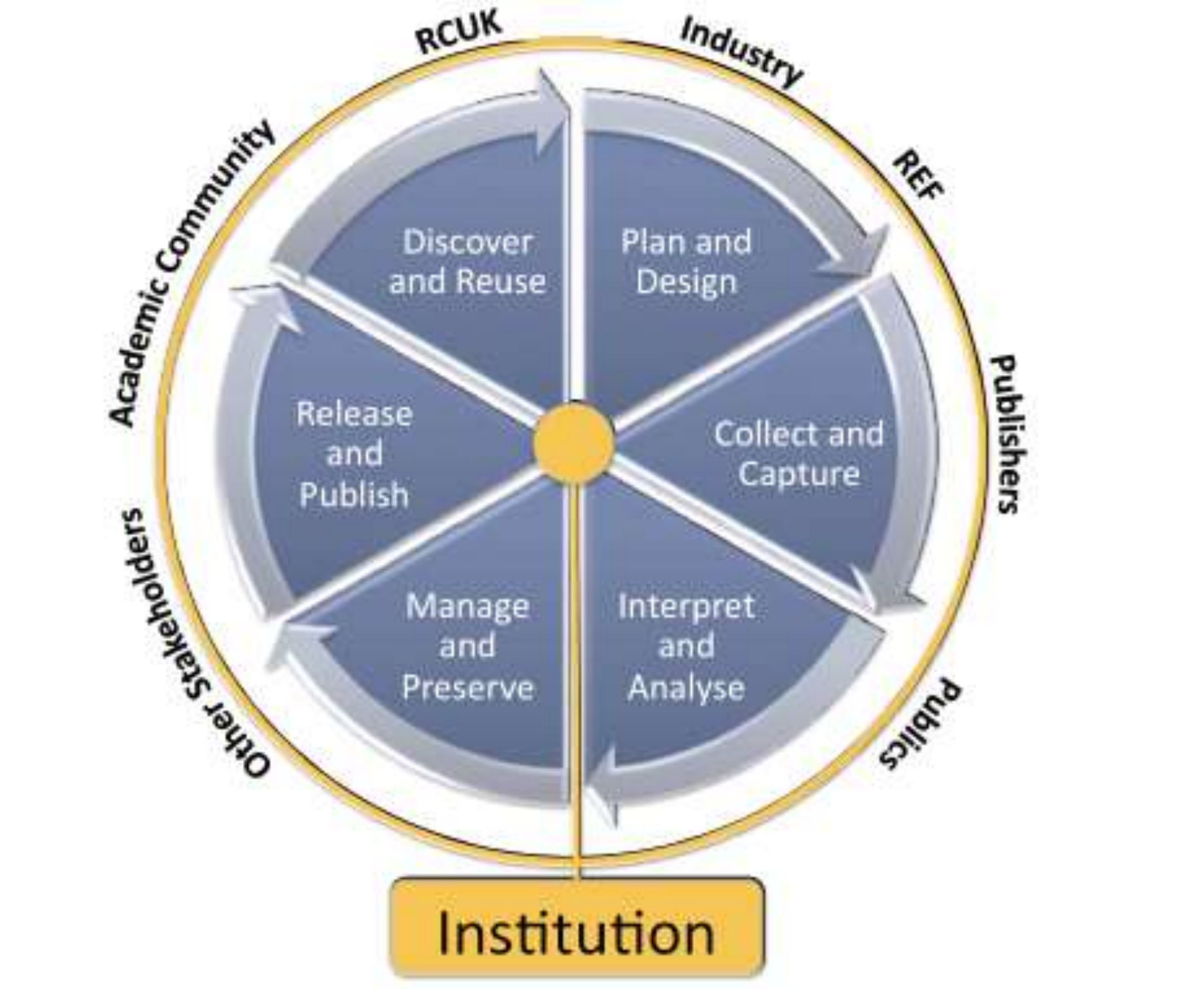
- to develop **a roadmap** that aligns their policies and processes with EPSRC's **expectations** by **1st May 2012**;
- to be fully compliant with these **expectations** by **1st May 2015**.

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

How ready are institutions?

Research360 Project



The Research360 Institutional Research Lifecycle Concept

JISC Managing Research Data Programme

Focus on academia-industry data

Focus on meeting EPSRC requirements

10 RDM challenges for institutions

1. ***Priorities, risks, benefits?***
2. What research data do we have?
3. Where is it stored?
4. What state is it in?
5. How long should we keep it?
6. Do we have a data strategy?
7. How much will it cost ?
8. Do we have a data policy?
9. Who has responsibility for RDM?
10. Are researchers and services staff “data-aware” and “data-savvy”?

1. Risk: where is your data?

Photo credits: Harvey Rutt http://www.ecs.soton.ac.uk/regenesis/pictures/

2. Reputation
3. Quality
4. Scale
5. Partnerships
6. Funding

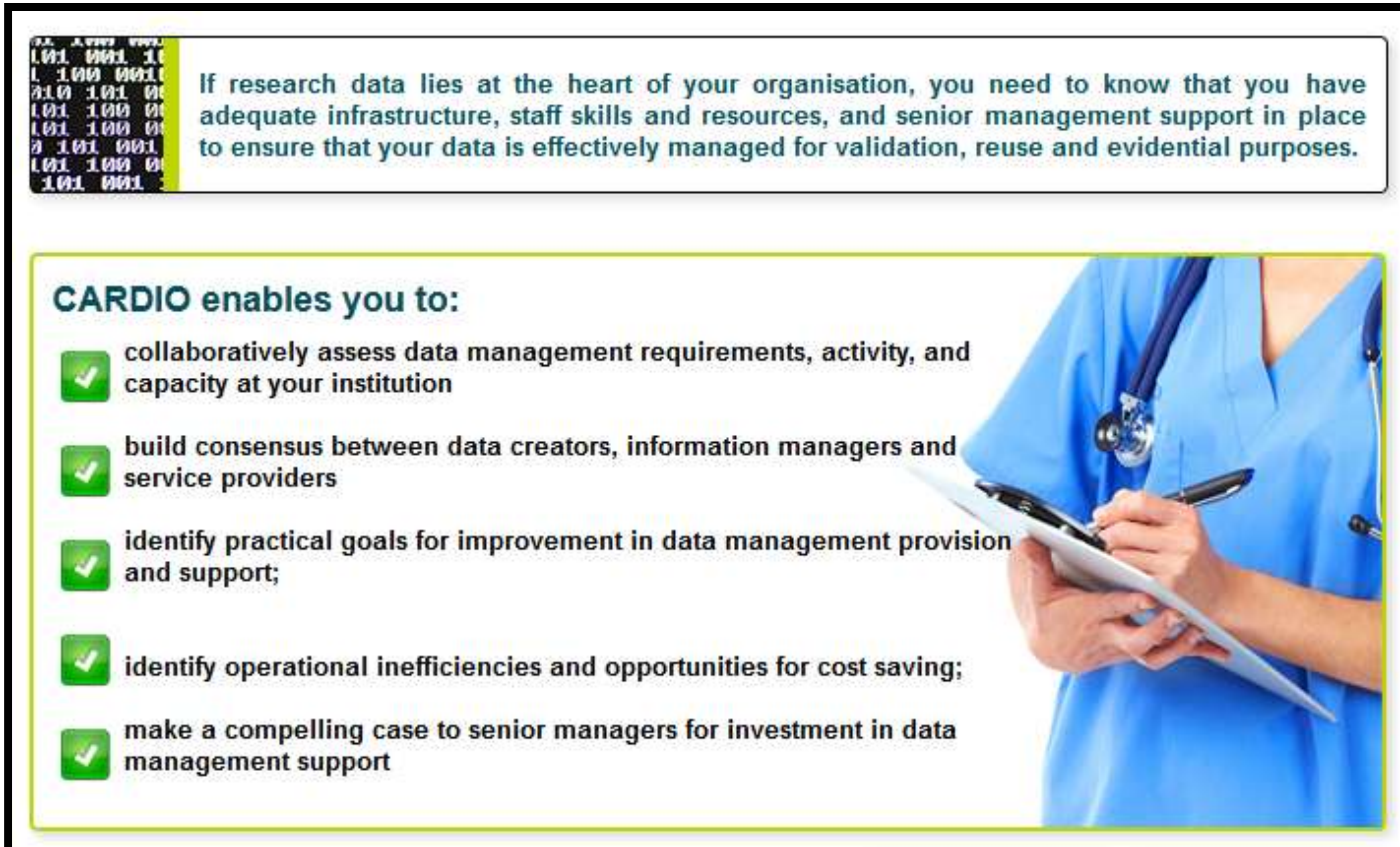
Getting attention….

Data Requirements at your Univ?

Data Asset Framework

BOS Bristol Online Surveys

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

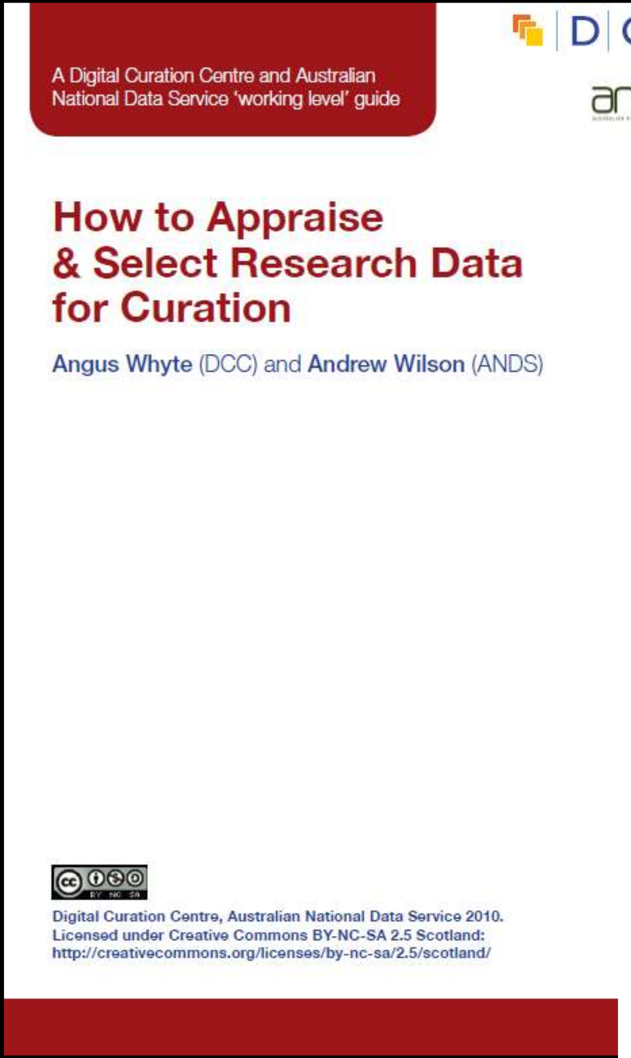
http://www.dcc.ac.uk/

BOS

Bristol Online Surveys

University of Bath Research Data Survey results

- 210 respondents (3.5% response rate) : PIs, ROs, postgrads
- ***Some preliminary findings:***
- Most have not had to produce a data management plan (81%)
- Much data is confidential, anonymised, under non-disclosure agreements, commercially sensitive, DPA, encrypted
- Data is also in non-digital form: lab notebooks, interviews
- Researchers store data on Univ Bath shared filestore ☺
- They also use Dropbox, USB sticks, home computers ☹
- Data loss: accidental deletion, hardware failure, obsolescence
- Open data is not the norm – often shared informally
- Lack of recognition for data sharing and reuse is an issue

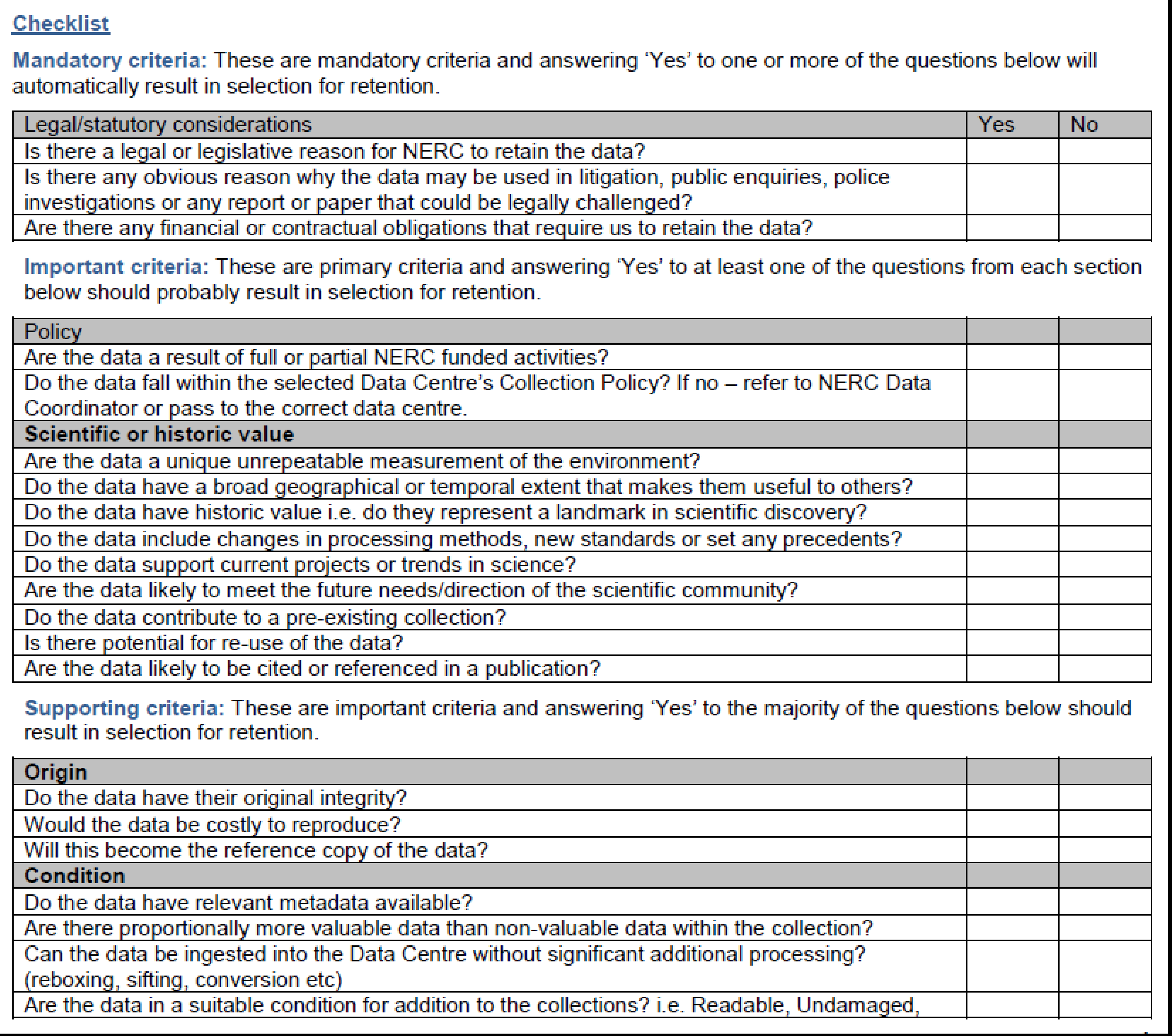


A Digital Curation Centre and Australian National Data Service 'working level' guide

How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and Andrew Wilson (ANDS)

NERC Data Value Checklist



Checklist

Mandatory criteria: These are mandatory criteria and answering 'Yes' to one or more of the questions below will automatically result in selection for retention.

Legal/statutory considerations	Yes	No
Is there a legal or legislative reason for NERC to retain the data?		
Is there any obvious reason why the data may be used in litigation, public enquiries, police investigations or any report or paper that could be legally challenged?		
Are there any financial or contractual obligations that require us to retain the data?		

Important criteria: These are primary criteria and answering 'Yes' to at least one of the questions from each section below should probably result in selection for retention.

Policy		
Are the data a result of full or partial NERC funded activities?		
Do the data fall within the selected Data Centre's Collection Policy? If no – refer to NERC Data Coordinator or pass to the correct data centre.		
Scientific or historic value		
Are the data a unique unrepeatable measurement of the environment?		
Do the data have a broad geographical or temporal extent that makes them useful to others?		
Do the data have historic value i.e. do they represent a landmark in scientific discovery?		
Do the data include changes in processing methods, new standards or set any precedents?		
Do the data support current projects or trends in science?		
Are the data likely to meet the future needs/direction of the scientific community?		
Do the data contribute to a pre-existing collection?		
Is there potential for re-use of the data?		
Are the data likely to be cited or referenced in a publication?		

Supporting criteria: These are important criteria and answering 'Yes' to the majority of the questions below should result in selection for retention.

Origin		
Do the data have their original integrity?		
Would the data be costly to reproduce?		
Will this become the reference copy of the data?		
Condition		
Do the data have relevant metadata available?		
Are there proportionally more valuable data than non-valuable data within the collection?		
Can the data be ingested into the Data Centre without significant additional processing? (reboxing, sifting, conversion etc)		
Are the data in a suitable condition for addition to the collections? i.e. Readable, Undamaged,		

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

- Research Council data requirements
- Institutional Roadmaps for EPSRC
- *http://www.bath.ac.uk/rdso/University-of-Bath-Roadmap-for-EPSRC.pdf*

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

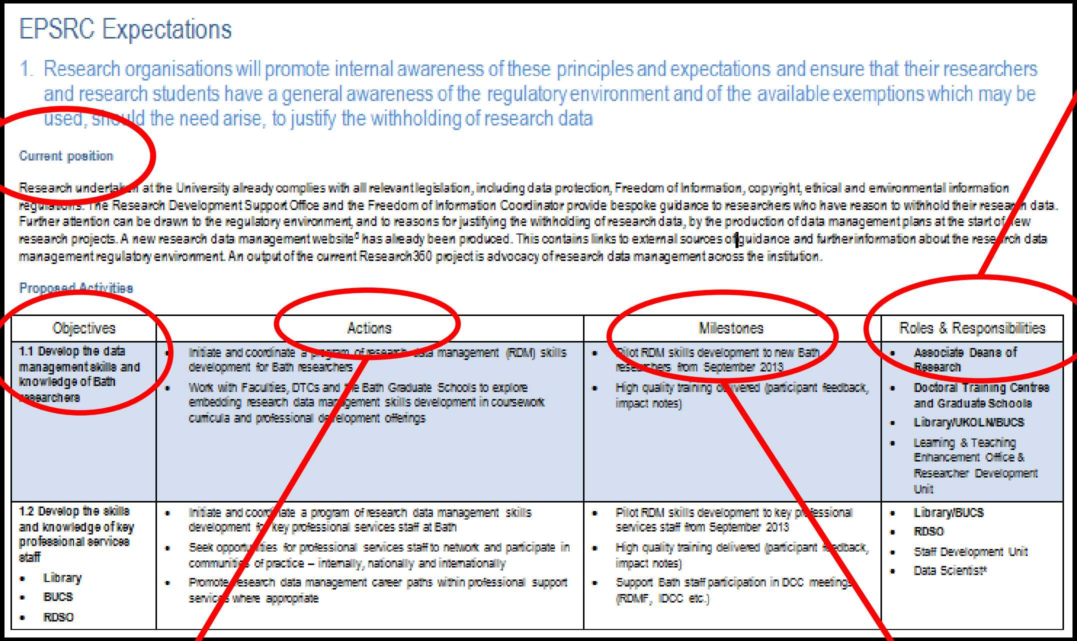
Acknowledgement

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.

Alignment with EPSRC Expectations

Roles and Responsibilities:
Who’s responsible

Where we are now



EPSRC Expectations

1. Research organisations will promote internal awareness of these principles and expectations and ensure that their researchers and research students have a general awareness of the regulatory environment and of the available exemptions which may be used, should the need arise, to justify the withholding of research data

Current position

Research undertaken at the University already complies with all relevant legislation, including data protection, Freedom of Information, copyright, ethical and environmental information regulations. The Research Development Support Office and the Freedom of Information Coordinator provide bespoke guidance to researchers who have reason to withhold their research data. Further attention can be drawn to the regulatory environment, and to reasons for justifying the withholding of research data, by the production of data management plans at the start of new research projects. A new research data management website[6] has already been produced. This contains links to external sources of guidance and further information about the research data management regulatory environment. An output of the current Research360 project is advocacy of research data management across the institution.

Proposed Activities

Objectives	Actions	Milestones	Roles & Responsibilities
1.1 Develop the data management skills and knowledge of Bath researchers	• Initiate and coordinate a program of research data management (RDM) skills development for Bath researchers • Work with Faculties, DTCs and the Bath Graduate Schools to explore embedding research data management skills development in coursework curricula and professional development offerings	• Pilot RDM skills development to new Bath researchers from September 2013 • High quality training delivered (participant feedback, impact notes)	• **Associate Deans of Research** • **Doctoral Training Centres and Graduate Schools** • **Library/UKOLN/BUCS** • Learning & Teaching Enhancement Office & Researcher Development Unit
1.2 Develop the skills and knowledge of key professional services staff • Library • BUCS • RDSO	• Initiate and coordinate a program of research data management skills development for key professional services staff at Bath • Seek opportunities for professional services staff to network and participate in communities of practice – internally, nationally and internationally • Promote research data management career paths within professional support services where appropriate	• Pilot RDM skills development to key professional services staff from September 2013 • High quality training delivered (participant feedback, impact notes) • Support Bath staff participation in DCC meetings (RDMF, IDCC etc.)	• **Library/BUCS** • **RDSO** • Staff Development Unit • Data Scientist*

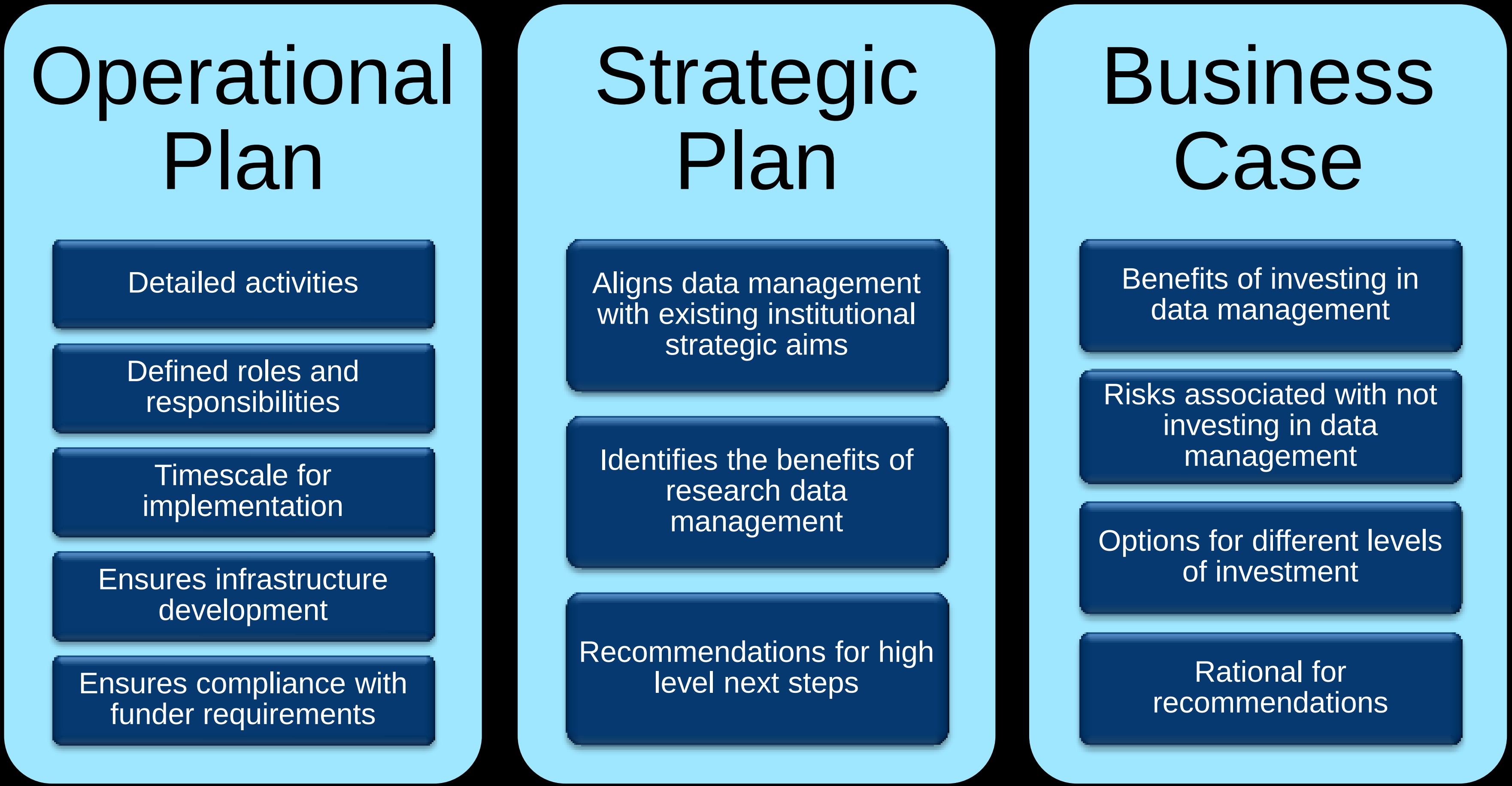
Objectives:
Where we need to be

How we’re going to get there

Milestones:
When it will be done by

- Finding the balance between ‘gold standard’ and ‘good enough’
- Ensuring that all stakeholders were satisfied

Roadmap → Strategic planning ↓



Operational Plan

- Detailed activities
- Defined roles and responsibilities
- Timescale for implementation
- Ensures infrastructure development
- Ensures compliance with funder requirements

Strategic Plan

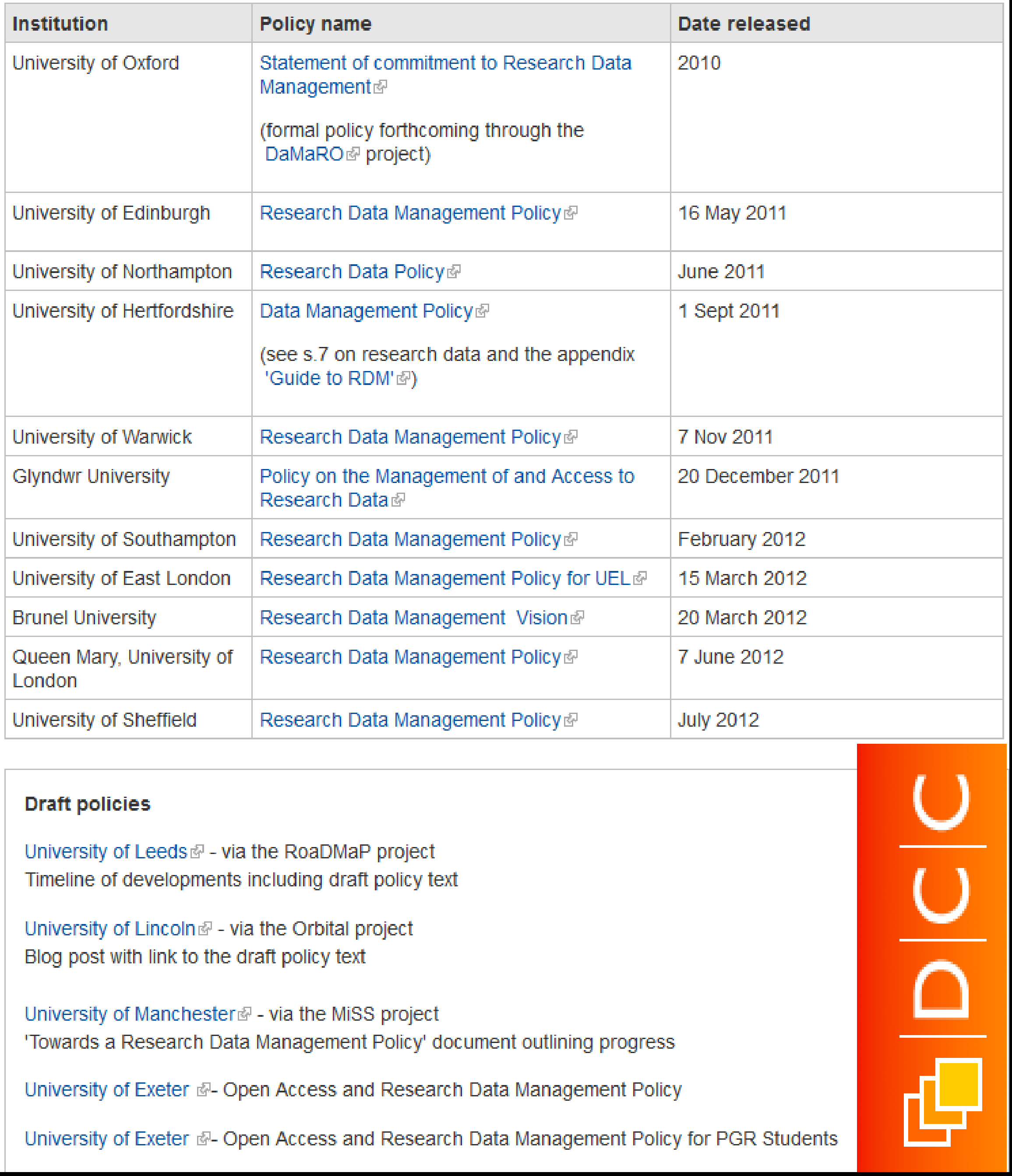
- Aligns data management with existing institutional strategic aims
- Identifies the benefits of research data management
- Recommendations for high level next steps

Business Case

- Benefits of investing in data management
- Risks associated with not investing in data management
- Options for different levels of investment
- Rational for recommendations

Transition from project-based to integrated infrastructure

Institutional data policy development



Institution	Policy name	Date released
University of Oxford	Statement of commitment to Research Data Management (formal policy forthcoming through the DaMaRO project)	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
Glyndwr University	Policy on the Management of and Access to Research Data	20 December 2011
University of Southampton	Research Data Management Policy	February 2012
University of East London	Research Data Management Policy for UEL	15 March 2012
Brunel University	Research Data Management Vision	20 March 2012
Queen Mary, University of London	Research Data Management Policy	7 June 2012
University of Sheffield	Research Data Management Policy	July 2012

Draft policies

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

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

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

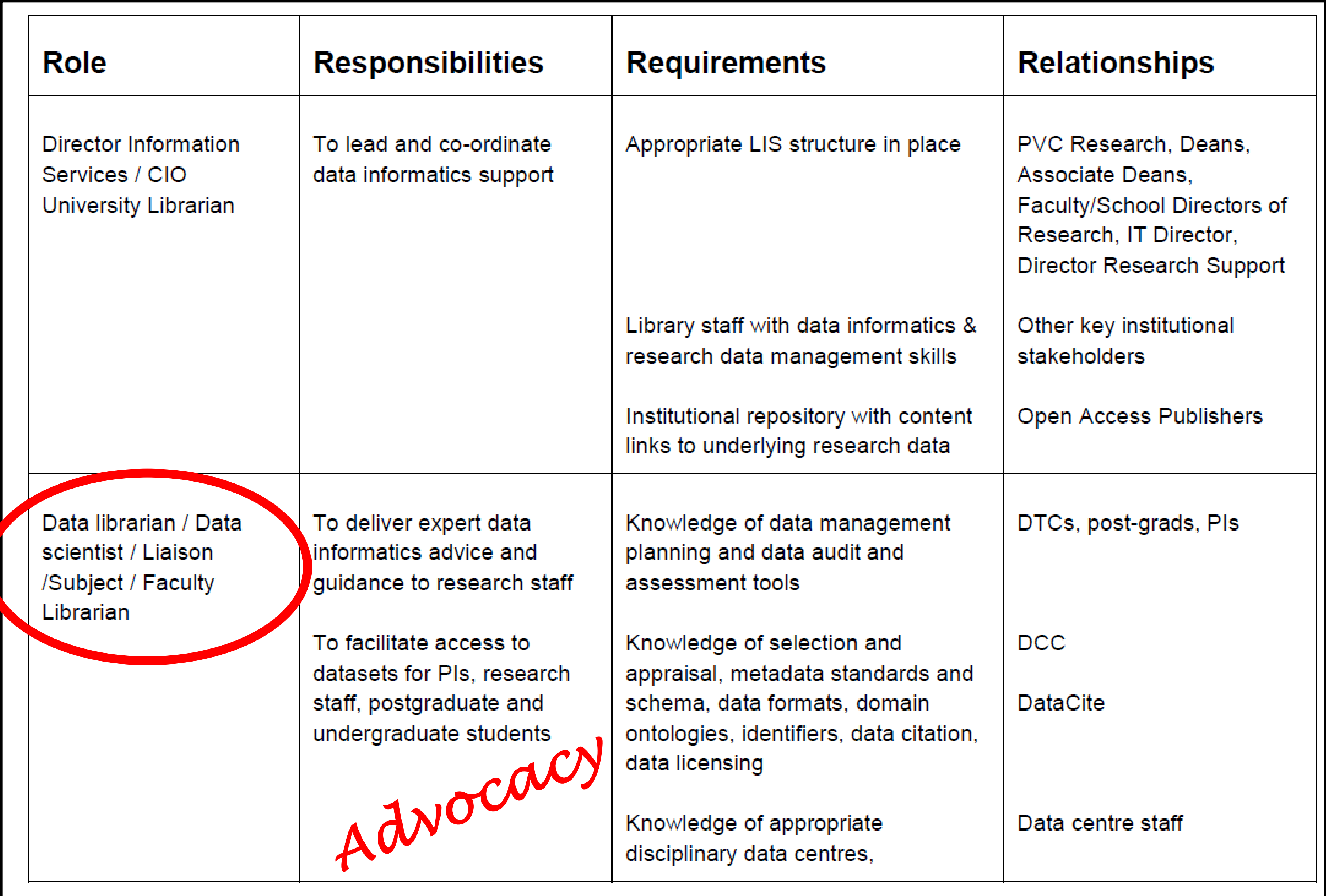
University of Exeter - Open Access and Research Data Management Policy

University of Exeter - Open Access and Research Data Management Policy for PGR Students

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

Developing a research data policy: reflections from Bath

- Keep it as succinct as possible
- It does need to have teeth
- Consult widely – many stakeholders
- Essential to discuss with the legal office
- Definitions of terms are helpful
- Provide detail in supporting procedures
- Research Data Steering Group approval
- PVC Research lead
- Roles and Responsibilities (again)



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

Advocacy

Full mapping : Informatics Transform, IJDC Current issue, 2012

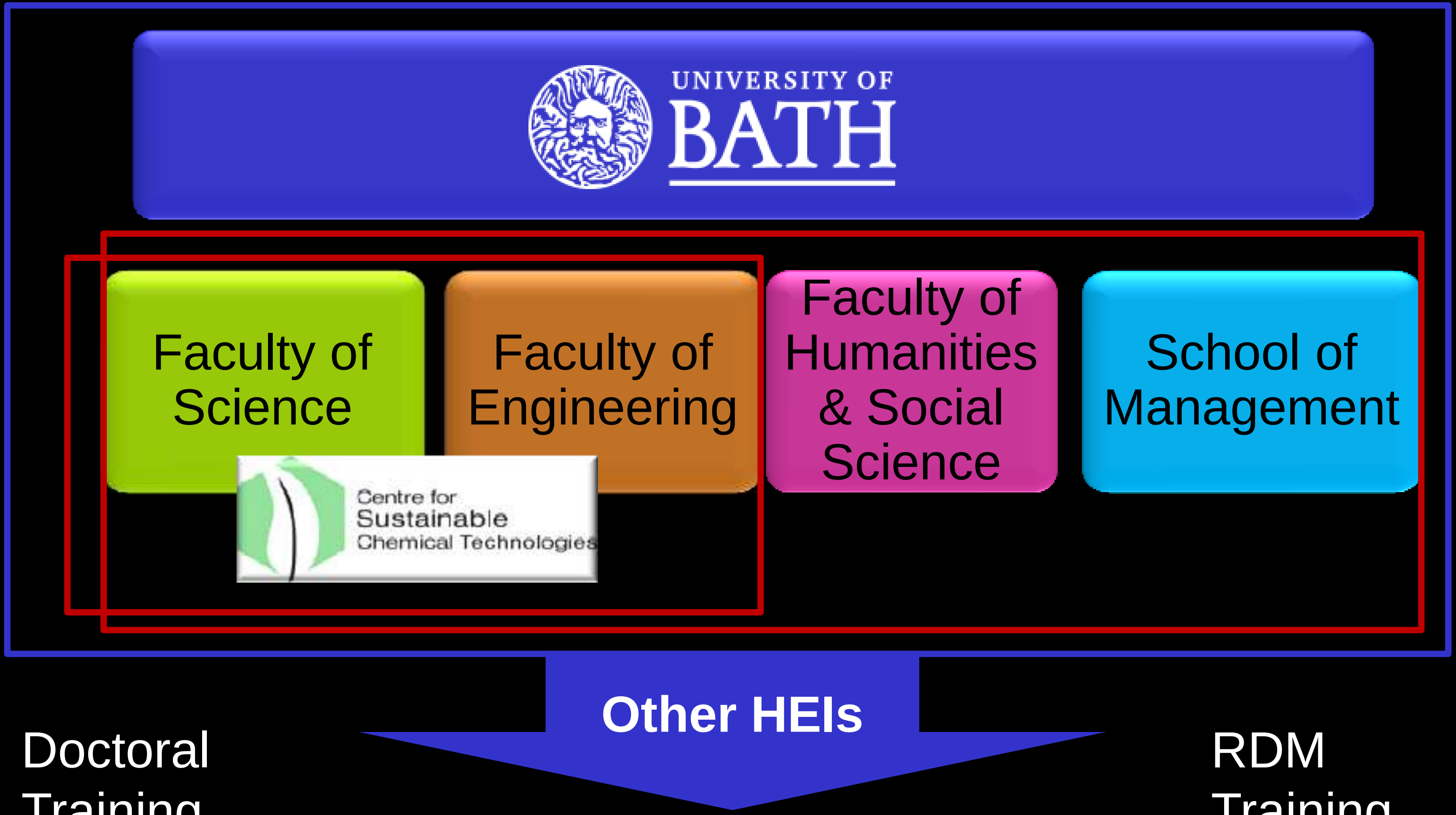
RDM Challenges for institutions

10. Are researchers and professional services staff “data-aware” & “data-savvy”?

Training cascade

Research360

Managing data across the institutional research lifecycle



UNIVERSITY OF BATH

Faculty of Science

Faculty of Engineering

Faculty of Humanities & Social Science

School of Management

Centre for Sustainable Chemical Technologies

Other HEIs

Doctoral Training Centres

RDM Training Module

Research360
Managing data across the institutional research lifecycle

Supporting resources

Research Development & Support Office

Home | For business | For staff | For graduates | About us | Contacts

Research Data

Research Data Management

What types of data are you creating? Who owns the data? What volume of data do you currently have? Where are the data stored? How frequently are your data backed up? Who can access your data? How long do you need to maintain your data?

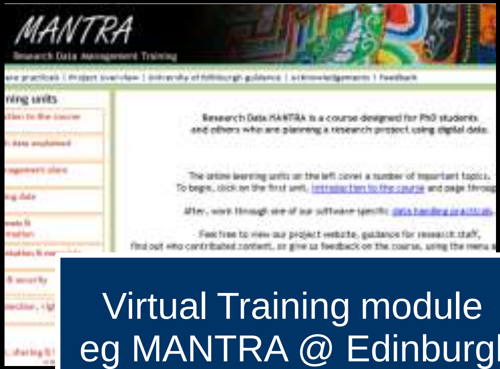
Research Data Management is organising and preserving research data so that it can be available for future use. Properly managed research data can be better shared among the wider research community, enhancing the long-term value of the work. This webpage has been designed as a central point of information about Research Data Management for researchers at the University of Bath. It utilises resources from other institutions, many of which are outputs from JISC funded projects.

For more information and guidance, contact research-data@bath.ac.uk

Research Data Management Website

Communication and dissemination (D2)
Cognitive abilities (A2)
Engagement and impact (D3)
Creativity (A3)
Engagement, influence and impact
The knowledge and skills to work with others and ensure the wider impact of research.
Knowledge and intellectual abilities
The knowledge, intellectual abilities and techniques to do research.
Domain D
Domain A
Domain C
Domain B
Professional conduct (C1)
Research governance and organisation
Personal effectiveness
Personal qualities (B1)

Researcher Development Framework
1-2 hour training sessions



MANTRA
Research Data Management Training

... practicals | project overview | University of Edinburgh guidance | acknowledgements | feedback

Learning units

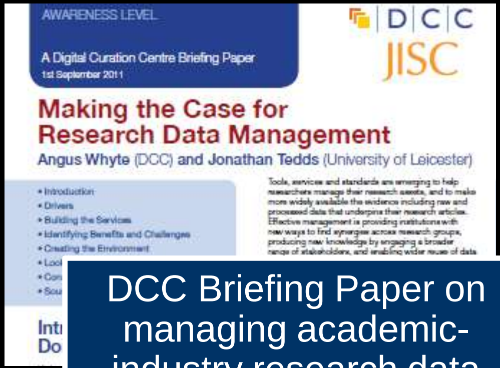
Research Data MANTRA is a course designed for PhD students and others who are planning a research project using digital data.

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 a...

Virtual Training module
eg MANTRA @ Edinburgh



AWARENESS LEVEL

A Digital Curation Centre Briefing Paper
1st September 2011

DCC
JISC

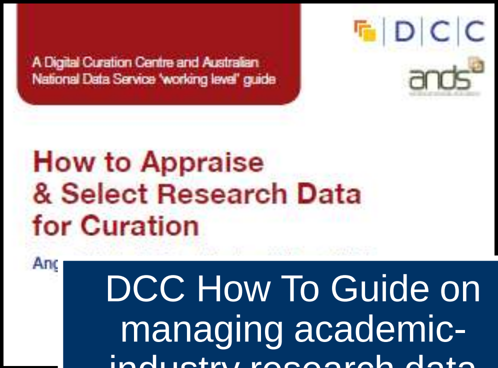
Making the Case for Research Data Management

Angus Whyte (DCC) and Jonathan Tedds (University of Leicester)

- Introduction
- Drivers
- Building the Services
- Identifying Benefits and Challenges
- Creating the Environment

Tools, services and standards are emerging to help researchers manage their research assets, and to make more widely available the evidence including raw and processed data that underpins their research articles. Effective management is providing institutions with new ways to find synergies across research groups, producing new knowledge by engaging a broader range of stakeholders, and enabling wider reuse of data

DCC Briefing Paper on managing academic-industry research data

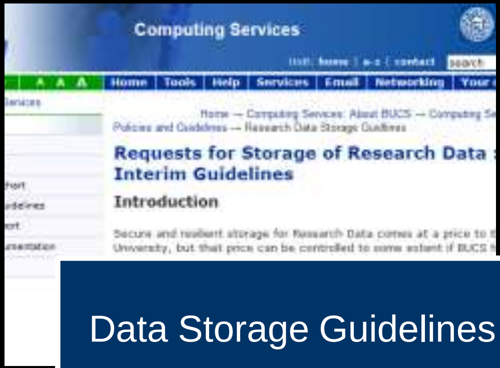


A Digital Curation Centre and Australian National Data Service 'working level' guide

DCC
ands

How to Appraise & Select Research Data for Curation

DCC How To Guide on managing academic-industry research data



Computing Services

Home | Tools | Help | Services | Email | Networking | Your...

Home → Computing Services: About BUCS → Computing Se... Policies and Guidelines → Research Data Storage Guidelines

Requests for Storage of Research Data: Interim Guidelines

Introduction

Secure and resilient storage for Research Data comes at a price to the University, but that price can be controlled to some extent if BUCS...

Data Storage Guidelines

Data Management Planning

DMP*online*

The DCC Data Management Planning Tool

DCC Checklist for a Data Management Plan

DCC

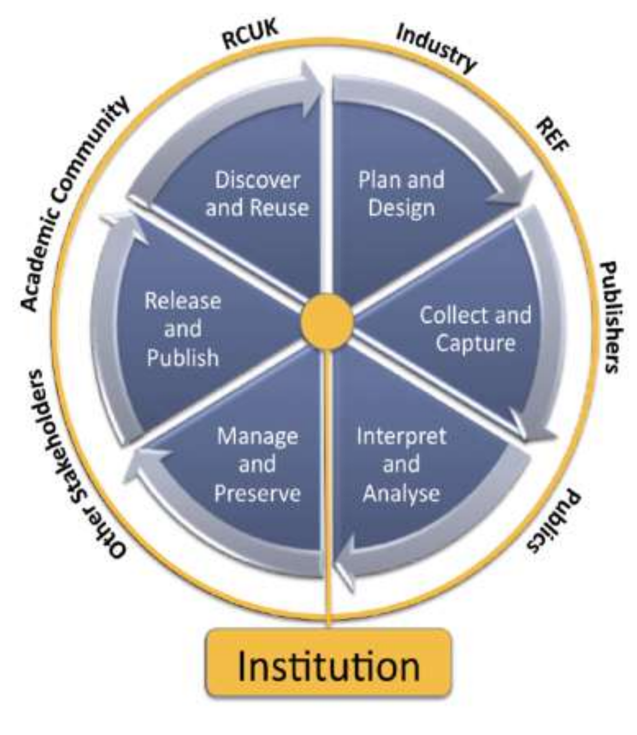
Institutional DMP template

Research360 is developing:

- Institutional template for **DMP***online*
- Guidelines for postgraduate students
- Support for researchers during grant application process:
  - Internal peer review college

Institutional data scientist role

- **Co-ordination and Collaboration**
  - Liaison / subject librarians
  - Repository manager
  - IT/Computing Services
  - Research Support & Development Office
  - Doctoral Training Centres
  - Researchers
- **Advocacy**
- **Training**



Research360
Managing data across the institutional research lifecycle

Liz Lyon, Informatics Transform, IJDC Current Issue, 2012

Family of data scientist roles

- ***data engineer -*** focus on software development, coding, programming, tools
- ***data analyst*** – focus on business/scientific analytics and statistics e.g. R, SAS, Excel to support researchers and modellers, business
- ***data librarian*** – focus on advocacy, research data management / informatics in a university / institute
- ***data steward –*** focus on long term digital preservation, repositories, archives, data centres
- ***data journalist*** – focus on telling stories and news

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

Posted by TEDxVancouver Team on October 17th, 2011 · No Comments »

Jer Thorp: Hope / Crisis, NYT Word Frequency

Infrastructure, Intelligence, Innovation: driving the Data Science agenda

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

Thank you.

DCC Resources can be downloaded from

http://www.dcc.ac.uk

Slides (including What is a data scientist?) at

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

Informatics Transform paper at

http://www.ijdc.net/index.php/ijdc/article/view/210/279