Research Data Management Infrastructure - A Case for Investment

Kathryn Dally, Sian Dodd, Wolfram Horstmann, Paul Jeffreys, Sebastian Rahtz, Andrew Richards, Dave de Roure, Sally Rumsey, David Shotton, Glenn Swafford, James Wilson

RIMSC
28 May 2013
Introduction
An initial reflection... Galathea Expeditions

- Three Danish ship-based scientific research expeditions

- Circumnavigated world from west to east and followed similar routes:
  - 1845 – 1847
  - 1950 – 1952
  - 2006 – 2007

The first two expeditions gathered large collections of research data - on paper
The third expedition collected a great deal of research data - in digital form

“Now, however, only a few data collections from the third expedition can be preserved. The remaining research data are dispersed over local hard disks with little or no chance of long-term preservation.”
Overview of Research Data Management Developments at Oxford

The Oxford RDM Working Group has been established to address the challenge of managing research data. The group aims to develop a strategy for managing research data in a way that is sustainable and adheres to best practices. The group has identified several key areas for improvement, including:

- Development of a centralized data management system
- Implementation of data management policies
- Training for researchers and staff

Approaches to RDM Challenges

The group has identified several approaches to addressing RDM challenges, including:

- Development of a data management policy
- Implementation of data management training
- Use of data management software

Further work

The group plans to continue its work on data management, with a focus on:

- Development of a centralized data management system
- Implementation of data management policies
- Training for researchers and staff

The group will also continue to engage with stakeholders to ensure that its work is aligned with the needs of the University.
Universities and Funders
Harvard – Dataverse Network

A Web Application for Sharing, Citing, Analyzing and Preserving Research Data

A repository for research data that takes care of long term preservation and good archival practices, while researchers can share, keep control of and get recognition for their data.

Supports the sharing of open data and enables reproducible research.
Information Services

Overview

Why manage research data?
Benefits of managing your data.

Data management planning
Ensuring all aspects of data management are fully perceived at the start of a project.

Documenting data
Guidance for describing your data for future use.

Data storage and backup
Storing and backing up research data to prevent data loss.

Data security, protection & confidentiality
Keeping data safe and controlling access to sensitive data.

Benefits of sharing data
Reasons to share or not to share data.

How to share your data
Making data accessible by depositing them in an open repository.

Support for data management
Major university services supporting research data.

University Homepage > Schools & departments > Information Services > Services > Research-support services

Research data management guidance
UCL Research Data

Introduction:

We are currently re-building these web pages. Please contact us directly for specific inquiries.

UCL Research Data will provide a service to facilitate better management of digital datasets by researchers.

- Key Objectives
- Further Information
- Access Guide for Pilot Users

The key objectives of the UCL Research Data service are to provide:

- Capability to store increasingly large volumes of electronic research data and data products
- Capability to support coordinated end-to-end research ‘workflows’ encompassing the use of both data storage and computational resources, by UCL researchers.
- The protection and preservation of UCL’s research data assets.
- Increased potential for re-use and possible future exploitation of UCL’s research data assets.
- Opportunities to consolidate UCL’s data storage infrastructure. Unified pan-UCL services (single system for all researchers)
- Improved compliance with increasingly stringent requirements for the management of project (data) outputs by the Research Councils.
Research Data Management

An Introduction to Research Data Management
Guidance for researchers new to the topic of research data management.
learn more

Help with Data Management Planning
Advice on creating a data management plan with links to subject-specific information.
learn more

Research Data Management Training
Information on research data management training options open to UoB staff.
learn more

Research Data Storage and Security
Data storage, security and ethical issues relating to research data.
learn more

Sharing and Publishing Research Data
Guidance on how to open up your research data to your collaborators or to the world.
learn more

Support at the University of Bristol
Institutional support for researchers engaged in research data management.
learn more
What is PURR?

The Purdue University Research Repository (PURR) provides an online, collaborative working space and data-sharing platform to support the data management needs of Purdue researchers and their collaborators.

Start Your Research Project

- Create a Data Management Plan
  Learn about the detailed requirements for your data management plan (DMP). Funding agency requirements are very specific and our DMP resources can help you to clear up any confusion. [Get Started]

- Upload Research Data to Your Project
  Create a project to upload and share your data with collaborators using our step-by-step form to guide you through the process. Invite collaborators from other institutions to join your project. [Create a Project]

- Publish your Dataset
  Package, describe, and publish your dataset with a Datacite DOI. Publishing will ensure your dataset is citable, reusable, and archived for the long-term. [See Published Datasets]

Featured Dataset

- Evaluation of Function Predictions for Moonlighting Proteins
  By Ishita K. Khan, Meghna Chitale, Catherine Rayon, Daisuke Khara
  Purdue University, Université de Picardie Jules Verne

Supplemental datasets used for evaluation of function predictions for moonlighting proteins.
Start Your Research Project

Create a Data Management Plan
Learn about the detailed requirements for your data management plan (DMP). Funding agency requirements are very specific and our DMP resources can help you to clear up any confusion. Get Started ›

Upload Research Data to Your Project
Create a project to upload and share your data with collaborators using our step-by-step form to guide you through the process. Invite collaborators from other institutions to join your project. Create a Project ›

Publish your Dataset
Package, describe, and publish your dataset with a Datacite DOI. Publishing will ensure your dataset is citable, reusable, and archived for the long-term. See Published Datasets ›
Relevance to the University of Oxford?

- Leading universities are developing suites of RDM services, following similar processes and investing – often collaboration between Library, IT and others
- Some of the universities are almost as devolved as Oxford with considerable RDM done locally based on central services (cf Big Data Centre)
- Our RDM Programme provides the essential components of an RDM infrastructure (already cut back to the minimum)
- If we do not act, we risk losing competitiveness in research capability and research income

“I think Oxford is paying coherent attention to the topic and that makes you competitive with the bulk of institutions”
Making research data available to users is a core part of the Research Councils’ remit and is undertaken in a variety of ways. We are committed to transparency and to a coherent approach across the research base. These RCUK common principles on data policy provide an overarching framework for individual Research Council policies on data policy.

**Principles**

- Publicly funded research data are a public good, produced in the public interest, which should be made openly available with as few restrictions as possible in a timely and responsible manner that does not harm intellectual property.

- Institutional and project specific data management policies and plans should be in accordance with relevant standards and community practice. Data with acknowledged long-term value should be preserved and remain accessible and usable for future research.

- To enable research data to be discoverable and effectively re-used by others, sufficient metadata should be recorded and made open to enable other researchers to understand the research and re-use potential of the data. Published results should always include information on how to access the supporting data.

- RCUK recognises that there are legal, ethical and commercial constraints on release of research data. To ensure that the research process is not damaged by inappropriate release of data, research organisation policies and practices should ensure that these are considered at stages in the research process.

- To ensure that research teams get appropriate recognition for the effort involved in collecting and analysing data, those who undertake Research Council funded work may be entitled to a limited period of privileged use of the data they have collected to enable them to produce the results of their research. The length of this period varies by research discipline and, where appropriate, is discussed further in the published policies of individual Research Councils.

- In order to recognise the intellectual contributions of researchers who generate, preserve and share key research datasets, all users of research data should acknowledge the sources of their data and abide by the terms and conditions under which they are accessed.

- It is appropriate to use public funds to support the management and sharing of publicly-funded research data. To maximise the research benefit which can be gained from limited budgets, the mechanisms for these activities should be both efficient and cost-effective in the public funds.
### Overview of funders' data policies (thanks to DCC)

n.b. the research funder links in the table take you to further details of each funder's policy

<table>
<thead>
<tr>
<th>Research Funders</th>
<th>Policy Coverage</th>
<th>Policy Stipulations</th>
<th>Support Provided</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Published outputs</td>
<td>Data</td>
<td>Time limits</td>
</tr>
<tr>
<td>AHRC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>BBSRC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>CRUK</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>EPSRC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>ESRC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>MRC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>NERC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>STFC</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
<tr>
<td>Wellcome Trust</td>
<td>○</td>
<td>○</td>
<td>○</td>
</tr>
</tbody>
</table>

- ○ Full Coverage
- ○ Partial Coverage
- ○ No Coverage
EPSRC - Engineering and Physical Sciences

The EPSRC mandated open-access publication in 2009.

A policy framework on research data was released in May 2011.

POLICY STIPULATIONS

Time Limits

The EPSRC expects research organisations to publish appropriately structured metadata online describing the research data they hold, normally within 12 months of the data being generated.

In line with the RCUK Common Principles, it is expected that data will be made available in a timely and responsible manner.

The EPSRC expects data to be maintained securely for 10+ years.

Data Plan

The EPSRC does not require researchers to submit data management or sharing plans in grant applications.

However, research organisations are expected to have a roadmap in place by May 2012 to ensure their policies and processes are in line with the its policy framework on research data by May 2015.

Access/Data Sharing

Researchers are expected to make publications freely available. The EPSRC also requires that published research papers include a short statement.
RDM required to facilitate best research

By managing your data you will:

- Ensure research integrity and replication
- Ensure research data and records are accurate, complete, authentic and reliable
- Increase your research efficiency
- Save time and resources in the long run
- Enhance data security and minimise the risk of data loss
- Prevent duplication of effort by enabling others to use your data
- Comply with practices conducted in industry and commerce

Views supported by DaMaRO survey of researchers (footnotes to Table 2)
Progress since last term
Progress made by RDM Group since HT 2013

- Split into four quadrants, developed priorities; and service business cases
- Prioritized and consolidated components within RDM infrastructure
  - Identified what we consider to be essential services only
    - Set priorities, but all essential ...
  - From Business Cases, chose least expensive option
  - Recognised that central services designed to complement local services
- Proposed 2 year RDM Programme ending at end of AY 2014/15
  - Capital injection over 2 year period to progress services
  - Not asking for long term commitment
  - After break-point, decision point for University – recurrent services?
- Very difficult to estimate capital injection for individual services at this stage
  - Business Cases circulated to RIMSC
  - Overall request envelope realistic, funds may need to be vired
## Identified researcher needs and services

<table>
<thead>
<tr>
<th>Tools to allow the researcher to:</th>
<th>Service name and provider</th>
<th>Funding model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Create data management plans (DMPs) with Oxford-specific metadata and automated deposit of DMPs into a dedicated repository</td>
<td>Oxford DMP Online, hosted at Edinburgh’s Digital Curation Centre</td>
<td>Free storage in DataBank instance at Bodleian Libraries</td>
</tr>
<tr>
<td>Work with a high-quality networking and backup infrastructure</td>
<td>Network, HFS and Advanced Research Computing (IT Services)</td>
<td>Centrally funded</td>
</tr>
<tr>
<td>Manage data whilst research is active</td>
<td>DataStage as user interface, linking to underpinning active storage</td>
<td>Software managed by research groups and departments; storage provided centrally, charged to research grants above a base level</td>
</tr>
<tr>
<td>Manage and publish relational databases online</td>
<td>Online Relational Database Service (IT Services)</td>
<td>Chargeable service, based on storage provided centrally</td>
</tr>
<tr>
<td>Store and preserve data after completion of research</td>
<td>DataBank (Bodleian Libraries)</td>
<td>Free at point of use, but funded research projects will be charged an ingest and cataloguing fee</td>
</tr>
<tr>
<td>Describe and catalogue datasets for finding by other researchers</td>
<td>DataFinder (Bodleian Libraries)</td>
<td>Centrally funded</td>
</tr>
<tr>
<td>Access information, advice, services and training</td>
<td>RDM web site (RSO); help and training from: RSO, Bodleian Libraries and IT Services</td>
<td>Centrally funded</td>
</tr>
</tbody>
</table>
RDM Services

- Backup (HFS)
- Network
- Data plan (Oxford DMPonline)
- Live Data management (DataStage)
- Leadership, facilitation and support
- Archiving (DataBank)
- Storage (Staas)
- Metadata (DataFinder)
- Online Database (ORDS)
## RDM Services – Funds Required

<table>
<thead>
<tr>
<th>Service</th>
<th>Priority (High²=5)</th>
<th>How funded</th>
<th>Cost over 2 years</th>
<th>Anticipated contribution from funders or charged</th>
<th>Oxford underwriting / investment</th>
<th>Likely annual recurrent costs from 2015/16</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataFinder</td>
<td>5</td>
<td>University</td>
<td>380</td>
<td>0</td>
<td>380</td>
<td>110</td>
</tr>
<tr>
<td>DataBank</td>
<td>5</td>
<td>University, with some charges to research grant</td>
<td>940</td>
<td>100</td>
<td>840</td>
<td>140</td>
</tr>
<tr>
<td>Program Coordination and Communication, RDM web-site, Oxford DMPonline, Engagement, Active data management</td>
<td>5</td>
<td>University</td>
<td>300</td>
<td>0</td>
<td>300</td>
<td>70</td>
</tr>
<tr>
<td>Training</td>
<td>4</td>
<td>University</td>
<td>120</td>
<td>0</td>
<td>120</td>
<td>60</td>
</tr>
<tr>
<td>Storage as a Service (St-a-a-S)</td>
<td>3.5</td>
<td>University</td>
<td>100</td>
<td>0</td>
<td>100</td>
<td>50</td>
</tr>
<tr>
<td>Online research database service (ORDS)</td>
<td>3</td>
<td>Researchers pay for service; storage University</td>
<td>278</td>
<td>25</td>
<td>253</td>
<td>70</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td></td>
<td><strong>2118</strong></td>
<td><strong>125</strong></td>
<td><strong>1993</strong></td>
<td><strong>500</strong></td>
</tr>
</tbody>
</table>
Bottom line and request to RIMSC

Total capital required: £2.12m
Total funding recovered from funders and through charging £0.13m
Total amount University must contribute to programme £1.99m

RIMSC is asked to decide whether it supports a case for this capital investment over a 2 year period in RDM infrastructure?
Preparation before submission to R.C.

- Modify in response to feedback
- Refine business cases
- Gather detailed plans from other universities (where possible)
- In consultation with the CIO, seek support from, and views of, the IT Committee
- Meet all members of Research Committee for briefing (as far as possible)
Thank you,
Questions
Funding – taken from Business Cases

- **DataFinder**
  
  £191k per year for operations cost plus embedding project cost  
  Recurrent operations cost: £110k per year (assuming that there will not be more than 5000 new records per year to curate)  
  Cost over 2013/14 and 2014/15 = 191 + 191 = 380  
  Recurrent cost = 110

- **DataBank**
  
  £218k per year for operations cost plus embedding project cost and £500k for 100 free DataBank silos  
  Projected income during embedding proj.: £100k, i.e. 20 DataBank silos charged out through grants  
  Recurrent operations cost: £136k per year; after the embedding project, cost for DataBank silos should be exclusively charged out.  
  Cost over 2013/14 and 2014/15 = 218 + 218 + 500 = 940  
  Recouped payments = 100  
  Cost to University = 840  
  Recurrent cost = 140
Funding – taken from Business Cases

RDM Coordination, DMPonline, Live data man

- Full-time service manager: 68 p.a. – recurrent
- *One project manager for 18 months (OeRC): 56*
- *One RA for 18 months (ARC): 45*
- Travel, consultancy, workshops, publicity: 10

Oxford DMPonline:

- Consultancy and training and dissemination = 10
- Recurrent = 2
- Live data man
- DataStage: 15 + 25
- Cost over 2013/14 + 2014/15 = 0
- Recurrent cost = 0
- Cost over 2013/14 and 2014/15 = 68 + 68 + 56 + 45 + 10 + 10 + 40 = 300
- Recurrent cost = 68 + 5 = 73
Funding – taken from Business Cases

Training and Support

- Cost / year = 59

StaaS

- 100 over a two year period

ORDS

- Option 1, cost-recovery basis
- We have: £136k in the first year, and (£136k + .22k*40) = £144k in the second
- Cost over 2013/14 and 2014/15 = 136 + 144 = 280 (278); recurrent cost = 150
- The business case estimates that charging will be 0.75/year. If 0 projects in first year, and 30 in the second year making 30*0.75 = £25k in the second year.
- In the third year it will be (60+30) projects = 90*.75 = £80k from charging etc., so recurrent additional costs is 70k
Visual representation of stages in the Research Lifecycle – data issues and support available (where there is no external research grant funding)

Pre-proposal
- Scoping data support services
- Collaboration/collective working
- Research Facilitators
- DataFinder – the catalogue for research data in Oxford
- Domain-specific IT support
- Librarians

Project plan (or internal proposal)
- Online RDM prep tools
- Tailored face-to-face advice
- Online advice
- DMP Online
- RDM Website
- Research Facilitators
- SRFs etc

Award/Offer
Research data issues may at this stage be influenced by:
- Collaborative agreements
- Sub-contracts
- MTAs
- Legal Services
- IP Rights
- Isis Innovation/OUC

Conducting the Research
RDM issues:
- Collection
- Management
- Storage/Retrieval
- Back-up
- Collaborative environments
- Emails/lab notebooks
- ORDS/DAAS
- STAAS
- Local dept’l
- DataStage
- Ext’l – UKDA or Colwiz
- HPC and VM hosting

Data Curation & Accessibility
- Non-digital data – making it shareable if requested
- Ensuring long-term access to data
- DataFinder
- DataBank
- ORA
- Data librarians

Publication
Publisher may require details re data or deposition of data:
- Data deposit
- Metadata deposit
- Metadata defines t&c for access
- Link /integrate with existing policy requirements (ORA)

DPhil thesis
- Link /integrate with existing policy requirements (ORA)

AND DO YOU HAVE DATA?...

- SYMPECTIC: ‘safety net’ for data capture
- DAMASC: data deposition request at article submission (JISC/OUP/Ox)
- DATAFINDER: starting point

- DataFinder - harvesting metadata and offering a catalogue
- Symplectic
- DataBank & ORA
- Librarians skilled in RDM

“..AND DO YOU HAVE DATA?”...