Research Data Management & Open Data: a UK Perspective

Tuesday 16th March, 2013

Dr. James A J Wilson
James.wilson@it.ox.ac.uk
The Situation in the UK

- National funding councils each have research data management and sharing requirements
- Joint Information Systems Committee (JISC) funding Higher Education institutions to develop infrastructure for research data management (Managing Research Data Programme (since 2010))
- Digital Curation Centre (DCC) developing tools and information resources to assist Higher Education institutions
RCUK Common Principles on Data Policy

• Agreed and published in 2011

“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”

“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 use of public funds.”
Research Councils UK

- 7 government-funded research councils in the UK
- ALL now have policies regarding the preservation and sharing of data produced during research they fund
- Most put responsibility on researchers to follow policy
- EPSRC puts responsibility on research institution to ensure researchers follow policy – May 2015 compliance expected
- Unclear yet precisely what each council will pay for
- Enforcement of policies is patchy
Awareness of Funders’ RDM Policies

- Awareness of the research data management policies of funders amongst Oxford researchers is modest

"Are you aware whether the major research funder(s) in your field expect you to provide information in your funding proposals about how you will manage, preserve, and/or share the research data that you create during the course of your research?"

- Research institutions need to ensure that awareness is improved – otherwise researchers leave RDM planning until the last moment when applying for funding
Benefits of good Research Data Management

• Reduces time researchers spend seeking information they need to defend results
• Documenting data makes it easier to go back to and understand data produced months or years earlier (even when it’s your own data)
• Research groups don’t lose the ability to interpret data when postgraduates leave or staff change jobs
• Improves data security and reduces the risk of data loss
• Easier to share data between groups, or make it available to the general public
• Can enable new kinds of research, as datasets can be combined and analysed in quite different ways than their creators originally envisaged
• improved research data management ensures integrity and can improve the quality of research produced.
• Helps institutions keep track of their data and monitor impact
“I do believe that our research could be enhanced by having better ways of storing information, because the way I store my thoughts makes a difference to how I use them when progressing in my thinking. I can see that improving the way I store them might help the actual thinking – apart from saving time, it might be a bit more substantial than that: having a clear view of what I’ve already done, or how my different projects interconnect, might just be heuristic in a sense.”
The Digital Curation Centre (DCC)

- Funded by JISC
- Provides tools and resources to help institutions meet their RDM commitments
  - CARDIO, DAF, and DRAMBORA benchmarking guides
  - DMP Online (Data Management Planning tool for researchers)
  - Best practice information
- Some content useful beyond UK
- [http://www.dcc.ac.uk/](http://www.dcc.ac.uk/)
The JISC Managing Research Data Programme

http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx

• 1\textsuperscript{st} phase: 29 projects ; 2\textsuperscript{nd} phase: 27 projects

• Funded projects have:
  – Surveyed and interviewed researchers
  – Written institutional RDM policies
  – Developed software tools
  – Developed training materials for researchers & support staff
  – Created online guides and information sites for researchers
  – Prepared business cases for training and services
  – Customised data management planning tools
  – Planned services and workflows
  – Considered metadata requirements for research data

• Outputs are intended for re-use – please re-use them!
RDM at the University of Oxford

- Work on scoping university RDM requirements began in 2008
- Several JISC-funded projects
- Loose programme of infrastructure development
  
  Opportunistic  ➔  Coordinated

- Principles:
  - researchers need to be at the core of development
  - intra-institutional collaboration amongst service providers

- University RDM Policy now in place
- Working group formed to create and present business case for integrated suite of services
Oxford RDM Policy

• Indicates value of research data
• Sets minimum of 3 years retention period for data after publication or public release of the work of the research (but must meet funder requirements if longer)
• Researchers responsible for documenting procedures for collection, storage, use, re-use, access, retention and destruction of research data and records
• University responsible for:
  – Providing access to services and facilities for the storage, backup, deposit and retention of research data and records that allow researchers to meet their requirements under this policy and those of the funders of their research
  – Providing researchers with access to training, support and advice in research data and records management
  – Providing the necessary resources to those operational units charged with the provision of these services, facilities and training
Awareness of Oxford’s RDM Policy

- Awareness of Oxford RDM Policy is currently low
  - This is partly deliberate
  - Policy is to some degree ‘aspirational’ (like many in the UK)
  - Infrastructure to support policy not yet fully developed

Before reading this questionnaire, were you aware of the University of Oxford Research Data Management Policy?

- Yes
- No
- Not sure
Who is responsible for what?

Data Management Roll-out (DaMaR0)
Who is responsible for what?

Guidance, training, & planning services (Research Services)

Preservation, curation, & publication services (Library Services)

Data gathering, manipulation, & storage services (active data) (IT Services)

Deposit & discovery services (Library Services)

Central coordination
Who is responsible for what?

Experimental research (Oxford eResearch Centre)

Data gathering, manipulation, & storage services (active data) (IT Services)

Big data processing (Supercomputer Centre)

Deposit & discovery services (Library Services)

Preservation, curation, & publication services (Library Services)

Long-term data stewardship (Academic Departments)

Guidance, training, & planning services (Research Services)

WORK IN PROGRESS!

JISC Data Management Roll-out at Oxford (DaMaR0)
Understanding Researchers

- Previous projects have looked at specific research disciplines, but now need to generalize across institution.
- Large-scale benchmarking survey undertaken in November / December 2012 – 314 responses from researchers generating data.

![Pie chart and bar chart showing distribution of research disciplines and data management practices.]

- Humanities
- Mathematical, Physical and Life Sciences
- Medical Sciences
- Social Sciences
- Other

![Bar chart showing distribution of research disciplines and data management practices:]

- As part of a team, with our research data managed by the team.
- As part of a team, but each member of the team looks after their own data.
- As an individual.
- Some of my research is undertaken as part of a team, but I also conduct some research independently.

Data Management Roll-out at Oxford (DaMaRO)
Research Data at Oxford

Types of data

- Textual
- Numerical
- Statistical
- Geospatial
- Images
- Audio
- Multimedia
- Bibliographic
- Other (please specify)

How data is stored

- In word processor files
- In tables or spreadsheets
- In relational databases
- In document databases (or other 'unstructured' forms of database)
- With XML mark-up
- As RDF triples
- Other (please specify)
Researcher Attitudes toward RDM

**Importance of RDM**
- **Essential** -- My research would suffer significantly if my data were not properly managed
- **Important** -- My research benefits from the time spent managing data
- **Helpful up to a point** -- Time spent managing research data can make life easier further down the line, but it’s not a very significant aspect of research
- **Not important** -- Devoting time to managing research data would be a distraction from the real work of research

**Ever been inspired to undertake new research after looking at past shared data?**
- **Yes**
- **No**
- **Seeing existing research data may have played a part in shaping new research ideas, but has never been a particularly significant factor**
- **Don’t know / can’t recall**

But fewer than a quarter had received any information about RDM from the University.

Data Management Roll-out at Oxford (DaMaRo)
Researchers who think their data would be of interest to others (from a smaller 2011 survey)

- All or most
- A substantial portion
- Little or none

Good intentions? (2011 Survey results)

- I have made my own research data publicly available in the past
- I intend to make all or most of the data from my current research project publicly available in the future
- I have made my own research data publicly available in the past
- My funding body requires me to make my data publicly available
- I have shared data privately with colleagues (e.g. by emailing a file) in the past
- I would like to make my data publicly available, but don’t currently have a straightforward means of doing so
- I do not currently plan to make any data from my current research project publicly available
- I would be happy to share data privately with colleagues (e.g. by emailing a file) if asked
- I am in principle happy to make my data publicly available once I have completed the work I intend to do with it and published the results
- I generally prefer not to share my research data
- I do not intend to publish the complete dataset from my current project, but have already made or expect to make limited subsets of it publicly available —
Have you (or has your research team) ever deposited your research data in a dedicated repository or data store?

- **Yes**
- **No**
- **Can't recall**

If ‘no’, why not?

- Not aware of an appropriate data store
- Do not want to share it publicly
- First project generating data
- Nothing worth preserving previously
- Not had time to get around to it
- Other

‘Other’ includes confidentiality agreements, security concerns, copyright, and ‘will do so shortly’
Researcher Training Requirements

- Conducted survey of scientists (i.e. excluding researchers from the Humanities and Social Sciences) in December 2012 – 193 responses

Common RDM tasks ranked by mean confidence level: 1 = lowest, 5 = highest

<table>
<thead>
<tr>
<th>Rank</th>
<th>Task</th>
<th>Confidence Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dealing with copyright, licensing, or other IP (intellectual property) issues relating to datasets</td>
<td>2.1</td>
</tr>
<tr>
<td>2</td>
<td>Preparing datasets for long-term preservation</td>
<td>2.5</td>
</tr>
<tr>
<td>3</td>
<td>Preparing datasets for sharing with researchers outside your research group</td>
<td>2.9</td>
</tr>
<tr>
<td>4</td>
<td>Version control</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>Determining whether research datasets ought to be preserved after the end of a particular project</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>Data documentation</td>
<td>3.3</td>
</tr>
<tr>
<td>7</td>
<td>Data management planning</td>
<td>3.4</td>
</tr>
<tr>
<td>8</td>
<td>Storing data securely and backing up</td>
<td>3.5</td>
</tr>
<tr>
<td>9</td>
<td>Organizing and structuring data within files (e.g. for analysis)</td>
<td>3.7</td>
</tr>
<tr>
<td>10</td>
<td>Managing bibliographic data</td>
<td>3.7</td>
</tr>
<tr>
<td>11</td>
<td>Organizing, structuring, and naming files and folders</td>
<td>4</td>
</tr>
</tbody>
</table>
## Training Desired

Common RDM tasks ranked by mean level of desire for training:
5 = most desired, 1 = least desired

<table>
<thead>
<tr>
<th>Rank</th>
<th>Task</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Dealing with copyright, licensing, or other IP (intellectual property) issues relating to datasets</td>
<td>3.55</td>
</tr>
<tr>
<td>2</td>
<td>Preparing datasets for long-term preservation</td>
<td>3.42</td>
</tr>
<tr>
<td>3</td>
<td>Data documentation</td>
<td>3.34</td>
</tr>
<tr>
<td>4</td>
<td>Preparing datasets for sharing with researchers outside your research group</td>
<td>3.27</td>
</tr>
<tr>
<td>5</td>
<td>Storing data securely and backing up</td>
<td>3.13</td>
</tr>
<tr>
<td>6</td>
<td>Data management planning</td>
<td>3.13</td>
</tr>
<tr>
<td>7</td>
<td>Determining whether research datasets ought to be preserved after the end of a particular project</td>
<td>3.11</td>
</tr>
<tr>
<td>8</td>
<td>Organizing and structuring data within files (e.g. for analysis)</td>
<td>3.02</td>
</tr>
<tr>
<td>9</td>
<td>Version control</td>
<td>2.97</td>
</tr>
<tr>
<td>10</td>
<td>Managing bibliographic data</td>
<td>2.73</td>
</tr>
<tr>
<td>11</td>
<td>Organizing, structuring, and naming files and folders</td>
<td>2.66</td>
</tr>
</tbody>
</table>
“I'm pleased to see that this area has been recognised as one that requires formal training”

“Data management seems secondary to "proper science". This is why it is not taken seriously by those who plan research and decide on workload and funding. It is very important to educate "THE BOSSSES" about the importance of data management!”

“I think I've managed OK but I was probably reinventing the wheel…”

“Much data management is specific to the group you're in, so it would be great to have general training, but also group-specific task training, if possible”

“Teaching how to manage electronic data to large groups seems like it would be difficult because so much is specific to the type of data that is being used/generated”
Damaro Project Training Materials

• Challenge is to produce materials that can be used without relying on the Damaro Project staff!

Available via ‘Jorum’

Data Management Roll-out at Oxford (DaMaRO)
Research data lifecycle

Research idea → data gathering → documentation → curation → access and re-use

funding bid → analysis & research outputs → deposit → discovery

Tools & Services

DataStage

DMP Online
Oxford

ORDS (Online Research Database Service)

Departmental data stores

External data stores

colwiz (& other external research collaboration tools)

External data repositories

Institutional RDM Policy

Training & Support

Sustainability
Research data lifecycle

Research idea → data gathering → planning → lit./data review → file organisation & analysis & research organisation → funding bid

Tools & Services

- DataStage
- ORDS (Online Research Database Service)
- DMP Online Oxford

Departmental data stores

External data stores
- colwiz (& other external research collaboration tools)

External data repositories

Institutional RDM Policy

Training & Support

Sustainability
Research data lifecycle

- Research idea
- Data gathering
- Documentation & curation
- Repository storage
- Data outputs
- Access and re-use

Databank

- Welcome to Databank
- Browse data packages

Oxford tools & services

- Underpinning infrastructure
- External tools & services

Training & Support

Sustainability

External data repositories
Research data lifecycle

Tools & Services

Institutional RDM Policy
Training & Support
Sustainability
Business model(s)

• Still very much work in progress!
• Value of whole infrastructure is greater than the sum of parts
• Full service costing – e.g. ORDS
  – Staff, hosting, capital renewal costs all recouped (in theory) from up-front charge to use service
• Lightweight open source software – e.g. DataStage
  – Installed and set up by projects that wish to use it (e.g. on NAS); no charge, minimal support
• Embedding in existing funded channels – e.g. RDM Training
  – Modifying and adding content for delivery via ITLP & Divisions
• Central University funding – e.g. DataFinder
  – Free at the point of use (so as to not act as a disincentive)
• ‘Mixed model’ – e.g. DataBank?
  – No charge below a certain data volume
Final Thoughts

• Funders increasingly driving agenda, but researchers need to shape the solutions, else they won’t be used
• Central services require assiduous promotion to researchers
  – And this should be factored into cost models
• Try to assess potential uptake of services
  – Typically high fixed costs, low variable costs
• Different business models for different RDM services
  – Adopting a modular infrastructure enables components to be added (or removed) in future
• Researchers need better recognition for data outputs

Data Management Roll-out at Oxford (DaMaR0)
http://damaro.oucs.ox.ac.uk/
damaro@oucs.ox.ac.uk
Please email us to receive progress updates