JISC Final Report

Project Information			
Project Identifier	*To be completed by JISC*		
Project Title	Leeds RoaDMaP (Leeds Research Data Management Pilot)		
Project Hashtag	#leedsrdm		
Start Date	01/01/2012	**End Date**	30/06/2013
Lead Institution	University of Leeds		
Project Director	Brian Clifford, Deputy University Librarian and Head of Learning and Research Support, b.e.clifford@leeds.ac.uk		
Project Manager	Rachel Proudfoot		
Contact email	roadmap@leeds.ac.uk		
Partner Institutions	Digital Curation Centre, F5, National Instruments		
Project Web URL	http://library.leeds.ac.uk/roadmap-project		
Programme Name	JISC Managing Research Data Programme 2011-13		
Programme Manager	Simon Hodson		

Document Information			
Author(s)	Rachel Proudfoot, Brenda Phillips, Tim Banks, Graham Blyth		
Project Role(s)	Project Manager		
Date		**Filename**	
URL	*If this report is on your project web site*		
Access	This report is for general dissemination		

Document History		
Version	**Date**	**Comments**
0.1	23/06/2013	Draft circulation to team for comment and additional content
0.2	5/07/2013	Revised with input from Project Team meeting
0.3	12/07/2013	Corrections from RDWG
1.0	17/07/2013	Sign off by RDSG

Table of Contents

NB : This table of contents 'auto-populates' - to update the table of contents – place cursor in the table of contents, right-click your mouse, click 'update field', select appropriate option

Acknowledgements

The RoaDMaP Project was funded by Jisc under the *Managing Research Data Programme 2011-3013* and supported by the University of Leeds. We are grateful to our academic case study leads Professor Richard Hall, Professor Bren Neale and Dr Ian Sapiro who provided support and expertise for the project. We received valuable support from the members of our working groups and the institution's Research Data Working Group and Research Data Steering Group. The Digital Curation Centre was one of our project partners and provided valuable input, particularly to our training and data management planning work packages; we worked most closely with Martin Donnelly, Sarah Jones, Joy Davidson and Monika Duke and also thank Alex Ball and Kerry Miller. We have benefitted from the work of many other projects in the JISCMRD Programme and are particularly grateful to the KAPTUR project for their technical analysis work and Essex for their work developing an EPrints data plug-in. Like many others we thank Edinburgh for their institutional data policy which proved an excellent starting point for the Leeds policy. Thanks are also due to our Programme Manager Simon Hodson and evidence gatherer Jonathan Tedds.

1 Project Summary

Leeds RoaDMaP (Leeds Research Data Management Pilot) which ran from Jan 2012 to June 2013 was one of several infrastructure projects funded under the JISC Managing Research Data Programme 2011-13. Our project investigated core elements of a research data management infrastructure including policies, data management plans and guidelines, processes, systems, support and training. Our findings are likely to be of interest to HEIs at an early stage of creating a research data management service and those with a particular interest in our three case study disciplines: biomedical engineering, music (specifically film music) and sociology (particularly qualitative, longitudinal data).

During the project, a new Research Data Management Policy was introduced and a research data web site created to support it. We piloted the Digital Curation Centre's data management planning tool, DMPOnline, feeding back its pros and cons. Stakeholder groups for research data management were identified and training delivered to early career researchers and research support staff; materials are available online for re-use. An online survey of research data at the institution illustrated some Faculty variations in data management practice and provided a profile of the size and location of research data at the institutions (the survey and results are available online). Interviews with case study researchers revealed varying attitudes towards data sharing and some concerns about the level of openness expected by research funders. This led us to consider managed access to research data, building this into our functional requirements for a research data repository. Various repository candidates were considered; towards the end of the project most effort was directed towards the EPrints open source platform, which will be taken forward post-project as our pilot data repository. Archival storage models were reviewed with particular attention to archiving as a service. We analysed how very large data sets or data with access restrictions may be served to requesters and started proof of concept work in this area.

The project acted as a catalyst for institutional change, raising awareness of research data management issues and highlighting areas where Leeds could improve support and gain efficiencies; it provided a springboard for further institutional funding and helped define what immediate actions we should take. Project staff also contributed to national and regional research data management discussions.

2 Main Body of Report

2.1 Project Outputs and Outcomes

Table 1: Project outputs and outcomes

Output / Outcome Type *(e.g. report, publication, software, knowledge built)*	**Brief Description and URLs (where applicable)**
Policy and Guidance	
Institutional RDM Policy	Research data policy agreed July 2012: http://researchdata.leeds.ac.uk/management-policy
RDM web site	Managing research data web site, geared towards supporting the RDM policy and good RDM practice: http://researchdata.leeds.ac.uk/
Policy roadmap	Brief chronology of policy and lessons learnt http://library.leeds.ac.uk/roadmap-project-outputs
User requirements	
Research data survey report	Results of the 2012 Research Data Survey are online: http://library.leeds.ac.uk/roadmap-project-outputs
Three RDM case study reports	• Data profiles and lessons learnt from Sociology (the Timescapes Archive); Engineering (SpineFx Project); Music (music of Trevor Jones): http://library.leeds.ac.uk/roadmap-project-outputs • Blog post analysing lessons learnt from the case studies from an award lifecycle perspective – pre-award, live award, post award • Case study interview questions: http://library.leeds.ac.uk/roadmap-project-outputs
Data Management Planning	
Data Management Planning Report & blog posts	• DMP Report - http://blog.library.leeds.ac.uk/blog/roadmap/post/169 (blog post) • DMPOnline plan formatting - http://library.leeds.ac.uk/blog/roadmap/post/134 (blog post) • DMPOnline developments - http://library.leeds.ac.uk/blog/roadmap/post/115 (blog post)
Repository	
Pilot repository	Investigation of DataFlow, CKAN and EPrints. Test EPrints repository available and will continue to be tested post-RoaDMaP.
Functional Requirements	List of functional requirements for a research data repository http://library.leeds.ac.uk/roadmap-project-outputs
Metadata	Joint meeting with White Rose colleagues to discuss metadata requirements. We are piloting the schema developed by JISCMRD colleagues at Essex and have deployed the ReCollect plug-in in our EPrints test repository.

Storage	
Virtualised Storage Assessment Report	Test criteria and results against the ARX file virtualisation system from F5 (commercial partner). http://blog.library.leeds.ac.uk/blog/roadmap/post/136
Archiving as a services	Investigation of archiving as a service model in conjunction with Arkivum Assured Archiving[1]; scoping for initial proof of concept.
Training	
Training materials	Training materials available for re-use can be downloaded from http://library.leeds.ac.uk/roadmap-project-outputs or from Jorum at ….
Training delivered to researchers	A full list of training and awareness raising events, including attendance figures and audience make up is online at: http://library.leeds.ac.uk/info/377/roadmap/271/evidence_and_impact/3
Training delivered to support staff	
Intangible outcomes	
Project Working Groups	Involvement of more staff in the RoaDMaP project; building expertise and awareness. Two of the working groups will continue to meet post project as they have proved valuable to discuss issues beyond the scope of RoaDMaP.
Improved understanding of RDM roles and workflow	Formal and informal discussions around RDM have helped progress understanding of who the key players in RDM are at the institution and potential differences in roles and workflows in academic Faculties. This work has helped identify further areas for investigation and clarification post-RoaDMaP.
Emerging referral network	The team to support RDM is emerging and various interested players have been brought together through RoaDMaP activity.
Greater understand of funding options	Various funding options have been explored to identify sustainable funding options for an institutionally based RDM service, including which RDM elements can/will be included in direct costs on grants; discussions with central finance at the University have been helpful.
RoaDMaP as catalyst for improved system interoperability	RoaDMaP brought together representatives of a number of institutional systems which could work more efficiently together to the mutual benefit of all e.g. Symplectic Research Information Management System, KRISTAL in-house grants management system, the Equipment Sharing and Management System, the EPrints scholarly publications and etheses repositories.
Project Management and Impact	
Project documentation, website & blog	http://library.leeds.ac.uk/roadmap-project Project Bid, Project Plan, Work Packages, Consent Form, Final Report: http://library.leeds.ac.uk/roadmap-project-management
Benefits Report	Qualitative and quantitative evidence of project impact, primarily in the areas of data management planning, training and repository requirements. The benefits reports and additional accompanying evidence is online at: http://library.leeds.ac.uk/info/377/roadmap/271/evidence_and_impact
Building the service	
RDM Service	A short account of our RDM policy evolution and lessons learnt is online

[1] Arkivum http://www.arkivum.com/

Roadmap	from our project outputs page: http://library.leeds.ac.uk/roadmap-project-outputs
Interim Funding Bid	The Research Data Steering Group in collaboration with RoaDMaP has secured interim funding from the University to continue to develop the RDM service. Further details are online at: http://library.leeds.ac.uk/roadmap-project-outputs

2.2 How did you go about achieving your outputs / outcomes?

2.2.1 Aims and Objectives

The RoaDMaP project built on work started at Leeds as part of the UKRDS (UK Research Data Service) feasibility study (2008) and the Leeds Building Capacity Project (2010-11)[2]. These initiatives brought together interested parties from across campus and led to an increased understanding for all involved of the wide-ranging issues surrounding research data management. RoaDMaP's **main aim** was to investigate requirements for, and pilot the implementation of, an institutional research data management infrastructure.

Several detailed **objectives** were included in the RoaDMaP Project Plan[3], broadly these were to:

- Develop a University of Leeds Research Data Management Policy and accompanying guidelines.
- Work closely with case study projects in order to understand their data management requirements.
- Pilot the DMPOnline data management tool from the DCC, defining and sharing enhancements.
- Define and implement a training strategy.
- Pilot a Research Data Management System (RDMS); use the pilot to inform planning for a RDMS to serve the institution.
- Pilot virtualised storage to stitch together multiple storage silos.
- Provide evidence to inform the business case for the IT infrastructure and personnel needed to meet the data management requirements of research funders, including EPSRC, by 2015.
- Share our project outcomes with the JISCMRD02 Programme

There were no significant changes to the objectives. We expanded our data requirements analysis to undertake an institution-wide research data survey, not included in the original Project Plan. Implementation of a pilot research data repository platform took longer than anticipated; we judged there was insufficient time towards the end of the project to ingest metadata from lab based data capture equipment – though this work has been started. Work on ingest by import was carried out and will continue as we develop the pilot repository platform. More time was spent with our Engineering researchers, defining their practices and requirements and applying RDM principles to their data so it will be better prepared for ingestion into a data repository.

2.2.2 Project governance

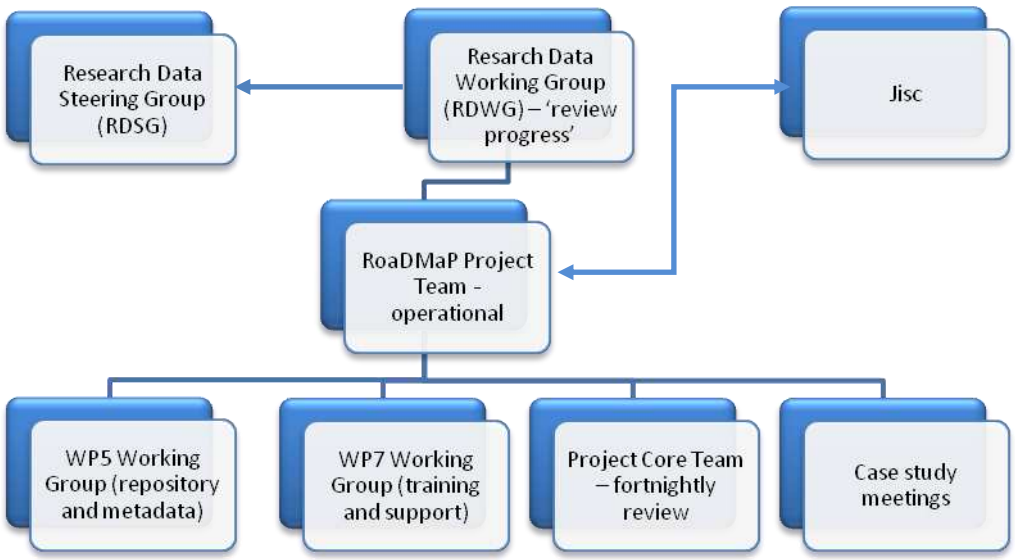
Two key groups were in place at the University of Leeds before the start of the RoaDMaP project, partly as a response to the 2011 EPSRC data requirements[4]: the Research Data Steering Group

[2] RoaDMaP Policy Timeline
http://library.leeds.ac.uk/info/377/roadmap/162/research_data_management_policy_evolution/2
[3] RoaDMaP Project Plan http://library.leeds.ac.uk/roadmap-project-management

(Chaired by the PVC for Research and Innovation) and the Research Data Working Group (Chaired by the Pro-Dean for Research in the Performance, Visual Arts and Communications Faculty). Group membership included researchers and support services staff[5]; RDWG provided oversight and advice for the project and RDSG provided strategic guidance. The Project Team - all those directly involved in RoaDMaP work packages delivery – met bi-monthly and a small, core team met fortnightly for on-going review and priority setting. In addition, specialist working groups were formed around Work Package 5 (Repositories and Metadata) and Work Package 7 (Training). We also had ad hoc meetings with our three case study leads. The project governance and reporting structure is summarised in Figure 1 below.

Broadly, the structure worked well; the small scale fortnightly meetings were particularly useful to keep up project momentum and discuss any issues arising.

Figure 1: RoaDMaP Project Structure



2.2.3 Project Work Packages

The main areas of activity under RoaDMaP's eight project work packages are outlined in *Table 1* below. Work with our case studies from Sociology, Music and Engineering cut across several of the work packages, particularly 2, 4, 5 and 7.

Table 2: Work Packages and Methodology

WP1: Project management: project coordination and review, liaison with Jisc, documentation, dissemination
• The project was supported by and reported to the groups in Section 3.2.2. • Participation in Jisc Programme activities was undertaken where possible. • The project made regular use of SharePoint to record progress, collaborate on reports, gather evidence etc.

[4] EPSRC Policy Framework on Research Data
http://www.epsrc.ac.uk/about/standards/researchdata/Pages/policyframework.aspx
[5] RDWG and RDSG membership can be found online
http://library.leeds.ac.uk/info/377/roadmap/122/roadmap_project_outputs/2

WP 2: Requirements analysis: analyse RDM requirements for three case studies

- Case studies were drawn from projects at different stages of maturity (pre, live and post award) and from different disciplines. Case study leads (researchers from the case study projects) became members of the RoaDMaP Project Team.
- Researchers and support staff from the case studies were interviewed: sometimes in groups or on the telephone, but mainly one to one and in person.
- Three detailed case study reports were produced and analysed for commonalities and differences to generate recommendations.
- Requirements from other stakeholder groups were gathered across formal and informal meetings and at training events.
- An online research data survey was conducted using Bristol Online Surveys, primarily to inform capacity planning. The survey was targeted at 'data owners'.

WP 3: Institutional research data management policies: agree and promote an institutional RDM policy supported by implementation guidance notes

- A high level policy, based partly on the University of Edinburgh model[6], was drafted with input from RDWG and RDSG. Feedback was sought from Faculty and University research committees and the Policy was endorsed by Senate in May 2012 with final wording agreed in July 2012.
- A supporting web site was put in place and is being gradually developed to better support the RDM Policy.
- The RDM Policy was promoted at training events and in publicity for the online research data survey in WP2.

WP 4: Data management planning: pilot DMPOnline; feedback suggestions to the DCC

- DMPOnline was used to create data management plans for 15 grant applications; it was also tested with our three case study leads
- Suggestions for adaptations and improvements were made to the DCC
- The process of data management planning was mapped against current institutional processes and systems.

WP 5: Software systems and metadata: create a working research data management platform with implementation guidelines and metadata templates

- Staff time from central IT was negotiated to help install and test DataFlow
- A list of functional requirements for a data repository was drawn up to assist with platform testing and choice
- A pilot repository was set up using EPrints software and contacts made with other projects testing EPrints for data.
- The project investigated long term archiving as a service, primarily with Arkivum Assured Archiving.

WP 6: Virtualised storage: working single virtualised storage area created from file systems on central University SAN, Faculty managed storage and 3rd party cloud storage

- Commercial partner F5 loaned two ARX-2500 devices (March – August 2012); staff time from central IT was agreed to help with the installation and testing.
- Test objectives were agreed by the project team, most tests performed and summary report

[6] University of Edinburgh Research Data Management Policy http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy

written.
- (The final phase of user experience testing was dropped when F5 changed the focus of its product away from rule-based data tiering)

WP 7: Training / people: develop RDM training materials making use of existing best practice; embed training opportunities for RDM stakeholders

- A WP 7 Working Group was formed.
- Training stakeholders were identified and current training provision at University of Leeds was reviewed to identify gaps.
- External RDM training resources were reviewed.
- Several training sessions were run for researchers and for support staff in consultation with the DCC.
- Training materials and feedback were made available online.
- Options for embedding training outlined to RDWG.

WP 8: Dissemination, evaluation, and exit /sustainability strategy: ensure lessons learnt and RDM practices embedded and sustained

- Regular reports on project progress were presented to RDWG and RDSG.
- The work of the project was promoted through training events, at committees and meeting and by directly involving a range of colleagues in project activity.
- Training events were evaluated by participants and the results used to inform planning.
- A spin out group from RDWG known as the 'Timeline Group' was formed to identify which activities, with approximate timings, would be needed to build a RDM service at the University, including those areas out of scope for RoaDMaP. The group also considered potential funding models.
- A case for interim funding to continue to build the RDM service and scope options for longer term resourcing was put to the Vice Chancellor's Executive Group in June 2013.
- Options for areas of shared service were outlined and discussed with colleagues from the White Rose University Consortium and N8 Research Partnership.

Out of scope

- Detailed research data management requirements analysis beyond the identified case studies.
- Recommendations for institution-wide storage solutions.
- A comprehensive research data audit across Faculties.

2.3 What did you learn?

2.3.1 Institutional RDM Policy

The high level policy progressed rapidly through committees and met with little resistance; some rewording was required to emphasise the policy was consistent with legal and ethical requirements:

> *"Data management plans should take account of and ensure compliance with relevant legislative frameworks which may limit public access to the data (for example, in the areas of data protection, intellectual property and human rights)."*

Specific questions and issues have tended to arise as the policy is translated into practice: for example, which researchers must complete a data management plan and the cost implications of data deposit.

Having a policy has proved useful for RDM advocacy with both researchers and support staff and provided a framework for considering what RDM actions and system developments are needed to fully implement and support the policy. Rather than a compliance 'stick' the policy has acted more as a 'heads up' for stakeholders, indicating that change is imminent and that RDM is being taken seriously.

2.3.2 Researcher requirements

Securing input from case study researchers at an early stage has been very helpful, providing us with ready access to academic perspectives throughout the project.

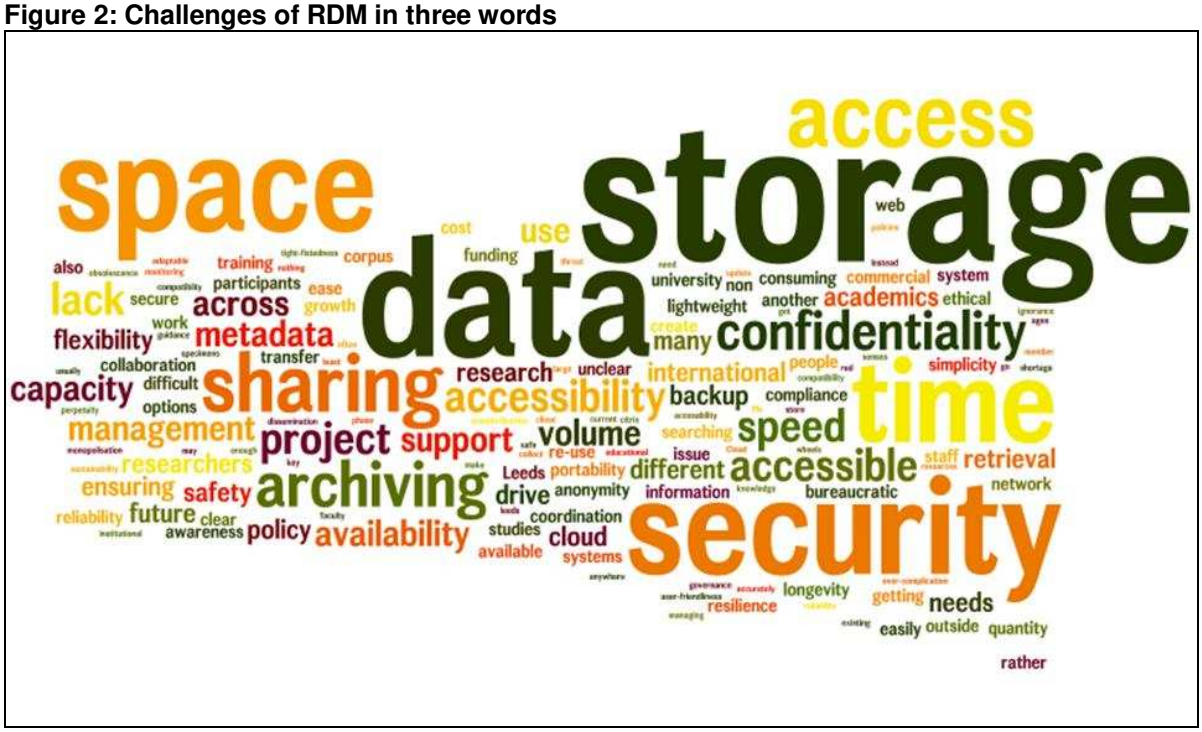
We reviewed various sets of interview questions but found that a relatively short set of semi-structured questions was sufficient for most purposes when interviewing researchers, who are more than ready to discuss their work and the challenges they face. In some cases, the interviewees preferred to have the questions in advance. Participants were often willing to give permission to be directly quoted but sometimes this was on the proviso they could see the quotation in context first.

We did not anticipate need formal ethical clearance from the University for our RoaDMaP work but our Research Data Working Group decided it would be best practice to do so and it was informative to go through the process. We were required to create a more formal consent form for completion by interview and pilot training participants and reword the online research data survey pre-amble.

The online research data survey we ran aimed to (i) gain a better understanding of the scope and scale of research data on University networks to help with capacity planning and (ii) poll research data management practice. Our response size (N=242) represents about 10% of our target population; however, because we asked 'data owners' to fill in the survey to avoid double counting, the true number of researchers represented is likely to be considerably higher. The response rate in some Faculties was very low; thus the findings should be interpreted with caution. It could have been useful to have agreed an 'acceptable' completion rate across the institution and in Faculties. Nonetheless, we now have a fuller picture of our research data assets and the respondents had the opportunity to suggest examples of good practice and raise any research data management concerns. Survey participants were invited to sum up the challenges of research data management in three words: the Wordle (word cloud) in *Figure* 2 below is generated from the results, showing a variety of responses but also demonstrating researchers' focus on storage and security.

Our three case studies (Sociology, Music and Engineering) proved a rich source of information on researcher requirements – and those of support staff. As well as variations across the disciplines- for example, the size and nature of data being generated – the case studies highlighted commonalities across projects, particularly the impact of data management training and support on research practice. (See Impact Section 3.4)

Figure 2: Challenges of RDM in three words



2.3.3 Data management planning

Several enhancements and new areas for development were fed back to DCC including plan sharing features, formatting improvements, removal of redundant questions, AHRC funder template problems, pre-creation of user accounts and possibilities for pre-populating project fields using query strings. A single, recommended tool to create data management plans is potentially attractive, particularly if it can accommodate customised templates for specific purposes such as a very lightweight template for small scale or unfunded research projects. We are working towards pre-populating plans with information from our grants management system; having to re-key information will be a significant barrier to engaging researchers with the process.

Examples of data management plans used in our training events received positive feedback from participants; it can be useful for trainees to see more than one plan, including from outside their own discipline area as this can prompt new avenues for consideration or strategies for data management.

Research support staff were keen to have examples of 'boilerplate' text for inclusion in plans; there is a danger that too much boilerplate can lead to less engagement with the planning process, but there are opportunities to save time by suggesting content where a standard response may be appropriate: for instance, regarding institutional storage and backup arrangements.

Introducing data management planning into research practice and to institutional administrative and IT systems poses challenges. Although the Leeds RDM policy requires a data management plan for every project proposal, the reality is a phased-in change to practice. The pre-existing institutional data risk assessment process is gradually being replaced by data management planning and work is underway to better integrate the DMP process with the grant application workflow.

Creating a data management plan for the first time can require significant levels of support, which our feedback – for example in the Timescapes case study - showed was highly valued. However, intensive support will be difficult to scale across the institution, particularly in Faculties with high grant application rates. We recognise that the 'ownership' and support models for DMPs may look slightly different from Faculty to Faculty and will continue to explore these models after the project. Our training session with pre-award staff suggested a costing checklist, such as that produced by UK Data Service [7] is helpful but will be most effective when informed by real costing examples, ideally from successful bids. We are also looking at ways to build capacity across multi service teams in a 'train the trainer' approach; this work has been started under RoaDMaP but there is more scope to involve colleagues in the data management planning process, with the agreement of the PI, as a way of up-skilling more staff in this area.

2.3.4 Repository platform and data catalogue

It became clear, particularly through our work with the Timescapes and Music case studies, that some data will need a level of restricted access, including an access permission process. It is not yet clear to what extent these are 'edge cases'; it's likely that the majority of data from some subject areas can be made openly available – indeed this is already the culture in, say, bioinformatics[8] - however, for other disciplines in the arts and social sciences, the access picture may well be more complex. Responses from the repository community (summarised here http://bit.ly/1adI6jT) suggested widespread interest in exploring access control more fully; we have scoped access scenarios which we are sharing with the community and University of Southampton with a view to enhancing EPrints' functionality in this area. This work dovetails with our proof of concept work around archival storage with Arkivum (see 3.3.5).

Examples of research data in institutional repositories may consist of a single metadata record with multiple files attached – essentially a [Collection][Subcollection + files] structure. Edinburgh's DataShare service, which runs on DSpace, is an example we looked at with our Music case study[9].

[7] UK Data Service – *Data management costing tool and checklist*
http://data-archive.ac.uk/media/247429/costingtool.pdf
[8] European Bioinformatics Institute http://www.ebi.ac.uk/
[9] For example see A Collection of Dinka Songs, http://datashare.is.ed.ac.uk/handle/10283/155

This type of structure may well be suitable for smaller, less complex data sets. We have found the data generated by Engineering requires a deep hierarchical structure: we need to ensure a multi-level folder structure can be reflected within the data repository, with appropriate metadata at each level of hierarchy. The qualitative, longitudinal data from the Timescapes archive poses other challenges, requiring different views into the data by 'case' and by 'wave'[10]. There is still much work to do to ensure institutional data platforms can create appropriate relationships both within and between 'data sets' and display these in effective ways to enable search, browse and download/request.

A question for Leeds, which will be replicated at other institutions, was whether to build on our existing EPrints repository infrastructure, which is used to manage our scholarly outputs and Digital Library. Our Research Data Steering Group suggested that simply using EPrints for data because it was expedient would not be an appropriate approach. Our original project bid and work plan identified DataFlow as the candidate system we intended to pilot as it appeared to offer the best fit when considered against our project needs. We hit some technical issues with DataFlow – in particular, implementing the link between the DataStage file management environment and the DataBank repository. To mitigate risk, we considered other repository options; it became clear that there are no obvious market leaders in the data repository space and few platforms at a sufficient level of maturity for immediate deployment. We drew up a list of repository functional requirements[11] from different RDM stakeholders' perspectives, aiming to be platform neutral. We then looked at our three main candidates: DataFlow, CKAN and EPrints, concluding that EPrints was the best bet for a pilot service given the relatively short timescale required for compliance with EPSRC's data management requirements (May 2015). Having a demonstrable repository platform is a sine qua non for full engagement with researchers and support staff; it is perhaps better to have something basic which can be improved on and from which data can be readily extracted for migration than implement a complex system in the first instance.

Ingest by Import
At the end of a project the research team typically publish the results and conclusions of their work. The associated data and metadata will have been added to a repository, a DOI generated, and the DOI referred to in the publications. In many research projects data is collected, stored unchanged, and used for subsequent analysis possibly over several years. Rather than waiting until publication for the collection of metadata that will be used during repository ingest the Engineering case study in particular suggested that the metadata was best collected at the time the data was collected. The intention being to store the metadata in a form that permits easy ingest. Trials with EPrints and its import function show promise and will be pursued. This will also be important in studies that generate large volumes of data for a sometimes large number of samples which is again the case with the Engineering case study. Each scan operation will lead to data that could be of future use in its own right making it desirable to capture the associated metadata and ensure that this gets into the repository. With the likelihood of a large number of such operations the option of ingest by import becomes even more attractive to the researcher as compared to completing a metadata form for each dataset when using manual ingest.

2.3.5 Management of active and archived data

Researchers in Engineering – from our SpineFx case study and beyond – have provided a useful testing ground to explore metadata collection as part of routine research practice rather than a later 'add on' when data is ready to archive. Retrospective creation of metadata can be onerous and require much staff time; better to capture metadata as the data is created.

The Engineering case study data highlights a category of data – in this case, spinal scans - which are 'live' but potentially appropriate to archive as soon as they are generated as they are static and must not change.

[10] Case could be a specific individual or organisation. Waves are successive collections of data through time.
[11] Research data repository requirements http://blog.library.leeds.ac.uk/blog/roadmap/post/163

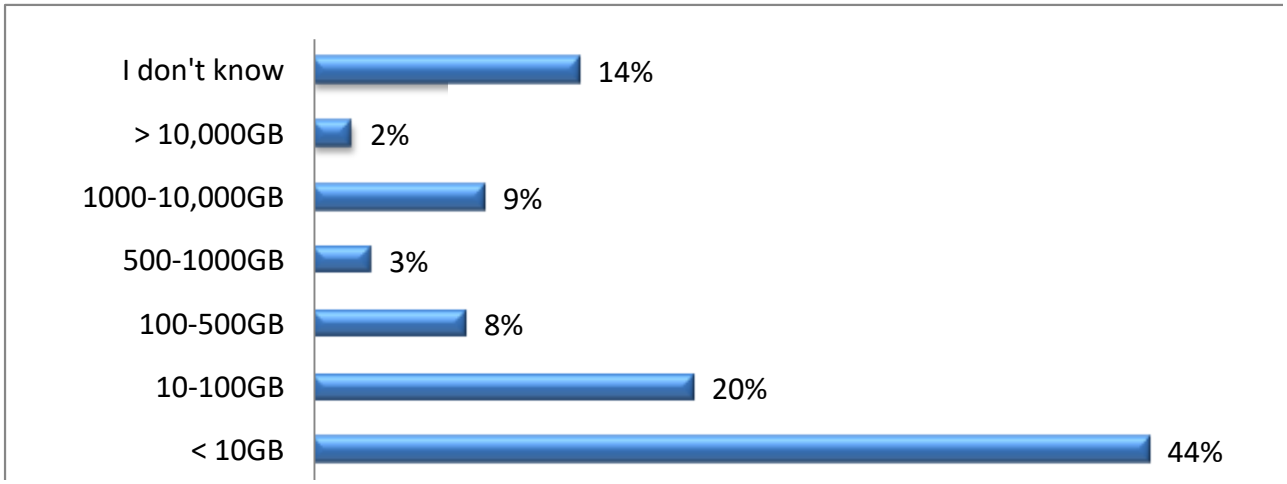
We are interested in archival storage for data sets and how this type of service might support access to data via different workflows. We have started to test these processes with Arkivum Assured Archiving[12] and have secured funding to continue this proof of concept work with Arkivum or a similar service. Workflows include:

(i) large data sets via request-restore-ready-deliver (e.g. engineering data generated by SpineFx, our engineering case study)
(ii) sensitive data sets via request-decide-retrieve(or decline)-deliver (e.g. restricted access data sets generated by the Timescapes programme, our sociology case study)

The research data survey run by RoaDMaP in 2012 provided a picture of research data at the institution, including the most commonly used formats, short term storage locations, the volume of data generated and how much data researchers anticipate keeping longer term. Full details are available from our survey report[13] but some headline findings are illustrated below, showing the majority of researchers estimate they generate less than 100 gigabytes of data a year (*Figure 3*) and there is a polarisation when it comes to estimates of how much data will need to be kept (*Figure 5*): for example, several maths and physical science researchers anticipate keeping a relatively low percentage of their research in the longer term in contrast researchers in medicine and health where the majority of researchers anticipate keeping over 75% of their data.

To advise researchers and inform institutional decision making about what data to keep longer term, we need a better understanding of how to appraise data and a clear selection policy for the institutional data repository; both will be essential and are identified as key early deliverables in our interim service development period (July 2013-July 2014).

Figure 3: How much research data would you typically generate in a year? (in Gigabytes)



[12] http://www.arkivum.com/
[13] http://library.leeds.ac.uk/roadmap-project-outputs

Figure 4: How much research data would you typically generate in a year? By Faculty

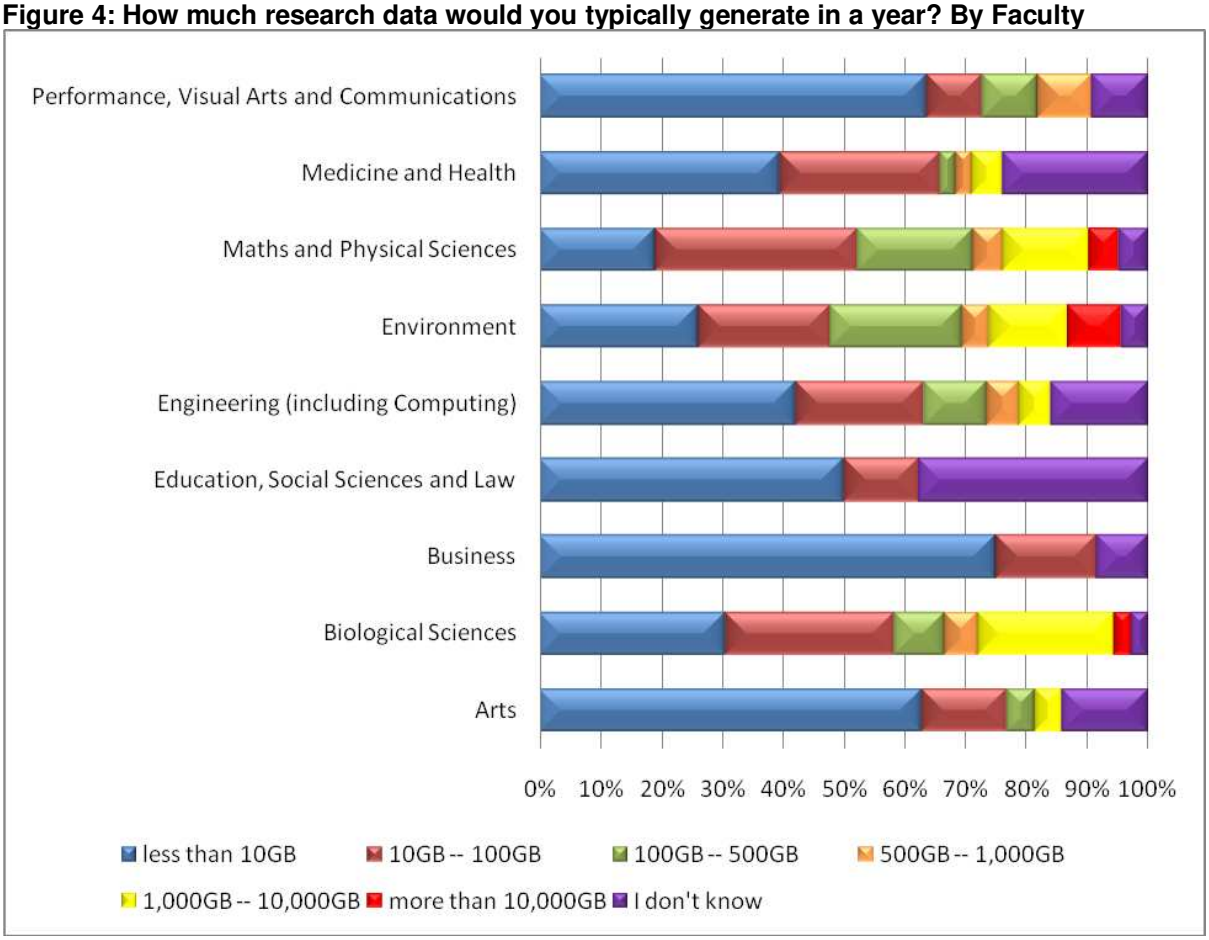


Figure 5: What % of research data generated would you need to keep for others to validate your research findings?

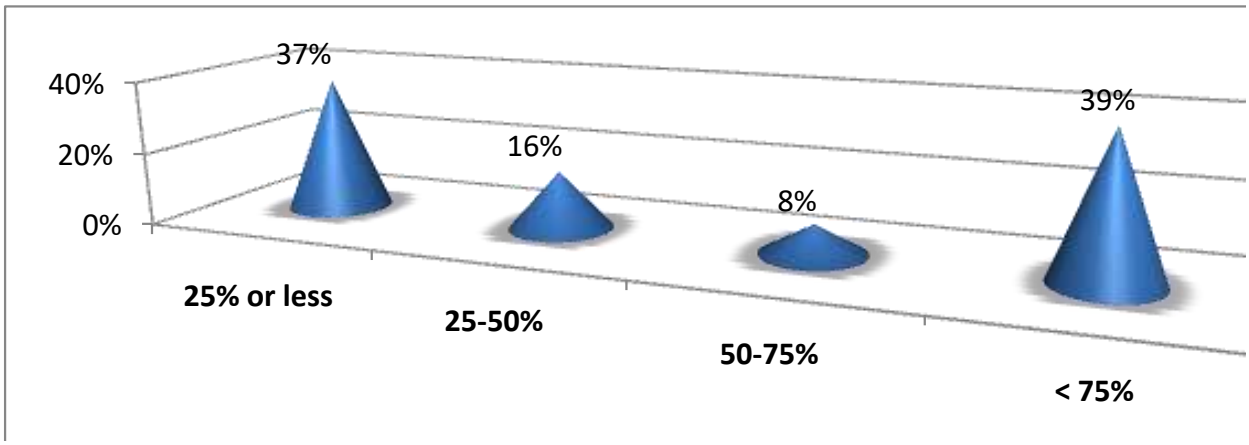
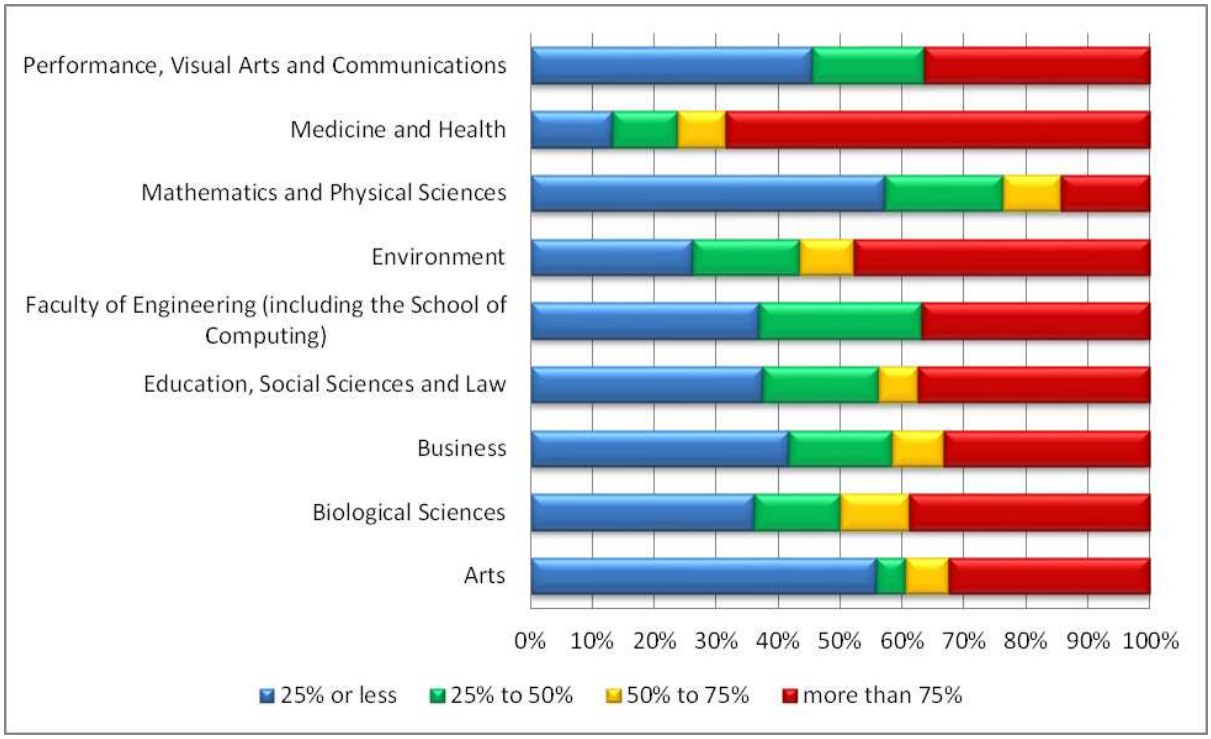


Figure 6: What % of research data generated would you need to keep for others to validate your research findings? By Faculty.



Fewer than 10% of respondents store their research data in an external data repository – the most frequently mentioned being British Atmospheric Data Centre and Protein Data Bank. Perhaps those whose data is already taken care of were less likely to complete the survey; but this is speculation. We plan to investigate data holdings in more detail at the Faculty level to flesh out our knowledge of what data assets are on our networks and their likely destination, however, taken as a whole, the survey results suggest a significant volume of data, varying in format, scale and complexity, will be in scope for a locally provided research data repository.

Table 3: % respondents using specific data formats

Data format	%
Documents (e.g. text, Microsoft Word, PDF), spreadsheets:	68%
Statistical data sets (e.g. SPSS, Stata, SAS):	34%
Books, Manuscripts (including musical scores):	31%
Laboratory notebooks, field notebooks, diaries:	30%
Questionnaires:	28%
Photographs / other images:	28%
Interviews (including transcripts):	28%
Laboratory instrument data (e.g. from microscopes, chemical analysers, monitors etc.):	24%
Computer software (e.g. modeling / simulation, schemas):	24%
Models, algorithms, scripts:	20%
Qualitative data sets (e.g. NVivo, ATLAS.ti, NUD*IST):	18%

Archival data:	17%
Audio:	15%
Existing databases (e.g. images, audio, video, text documents):	14%
Video:	13%
Microscope slides, artefacts, specimens, samples:	12%
Other:	12%
Observational data (e.g. Astronomical data):	9%
Standard operating procedures and protocols (health research):	7%
Methodologies and workflows:	7%
Test responses:	5%
Codebooks:	1%

2.3.6 Training and guidance

Working with colleagues from the University's training department (SDDU) and from other professional service areas was beneficial to the design, promotion and embedding of training. From academic year 2013-14, ULTRA (University of Leeds Teaching and Research Award) will include a half day session on research data management based on the pilot materials developed during RoaDMaP.

RDM training is most effective when structured around meaningful, practical activity: e.g. creating a data management plan for your own data; identifying costs in a data management plan. Theory should be linked to practice at the earliest opportunity. This can be difficult where roles and responsibilities are still emerging.

As soon as training is delivered, expectations are raised – participants want to know 'what next' to explore RDM further and embed in their professional practice. In this way, training activity acts as a driver for the development of other elements of a research data service; it is important that training activity does not run too far ahead of infrastructure development and vice versa.

Our discussions with researchers highlighted the importance of research support staff who are seen as a trusted source of advice and information, for example, offering checklists of activity researchers should consider as part of their bid preparation.

Training can be delivered to professional groups – so participants are broadly similar and may want to address similar issues – or to diverse groups so each can benefit from a range of perspectives on the topic (this was the model for our White Rose event *Perspectives on Research Data Management*[14]). Both approaches have merit; it may be that the mixed group approach is most appropriate when the basic RDM service has more maturity so that the end-to-end RDM process can be discussed in a more coherent way.

Some generic elements of research data management training (the RDM context and drivers, good practice in file management and naming, funder requirements, basic repository options) can be delivered by a range of staff with reasonable knowledge of the RDM landscape. We are also looking at online delivery options for these elements. For credibility, however, delivery by a professional peer of the training recipients is ideal. We envisage training will become more specialised once the full impacts of institutional and funder RDM requirements kick in. In this case, more input from specialists in a particular field will be essential and this may be an area ripe for collaboration across institutions – for example, the N8 Research Partnership.

[14] Perspectives on Research Data Management - 24th May 2012, Ron Cooke Hub, University of York http://library.leeds.ac.uk/info/377/roadmap/123/roadmap_events/5

We found the delivery of training acted as a catalyst for heightened interested in RDM and further participation in RoaDMaP. This was particularly true in engineering where the impact of training has been long lasting and has extended beyond those who attended the initial face to face session.

2.3.7 Business plans and sustainability

Anticipating how much a research data management service will cost, both in the short term and as it develops over time, has proved to be difficult. There are few mature examples of institutional RDM services to benchmark against. High levels of up-front investment may be too risky when RDM solutions are only just emerging. Securing the backing of the PVC Research and Innovation was vital to make the case for funding at the institutional level and working with central Finance has also proved to be an important consideration in building a business case.

To improve embedding of project outcomes, it is usually better to utilise existing groups and structures, where these exist, rather than creating new ones. However, we found that securing participation in working groups from professional colleagues outside the direct project team has proved a good way to build areas of shared interest and expertise. At least two of the RoaDMaP groups will continue and already have a programme of continuing business.

2.3.8 Building the Service

To scope the work required to take our developing RDM service forward, it was helpful to identify headline areas under which activity was needed. The Jisc/DCC *Components of an RDM Service* diagram[15] provided a good starting point. The activity headings proved unwieldy for some purposes – for example, reports to senior groups – but it is useful to retain some granularity to ensure areas of activity are not missed or 'assumed to happen' within broader areas.

Table 4: RDM Service development areas

RDM Service areas	Condensed list
• Policies and roadmaps • Guidance • Training programme • Data management planning • Management of active data / metadata • What to keep / appraisal • Repository platform • Data catalogue, metadata and identifiers • Storage of live and archived data • Business plans and sustainability • **Interoperability*** • **Regional / shared service options*** • **Stakeholder consultation*** * additions to the DCC list	• High level policy • Detailed RDM guidance to support policy • Data management planning • Storage and management of active data • Data repository and data catalogue • Training • Business plans and sustainability • Regional / Collaborative shared service

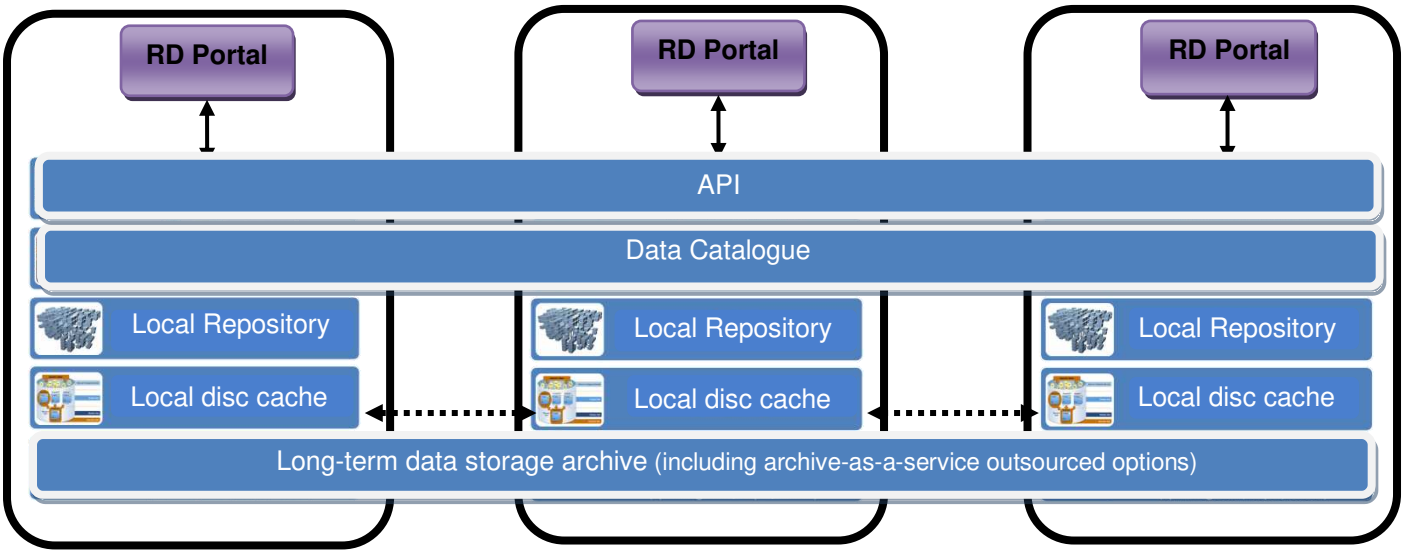
2.3.9 Regional / shared service options

During the project we discussed possibilities for sharing components of an RDM service with consortial partners: for example, the White Rose University Consortium or the N8 Research Partnership. We found it useful to split the RDM infrastructure into layers to illustrate there were options that involved sharing some, but not all service components. Figure 7 illustrates one possible configuration.

[15] How to Develop RDM Service: a guide for HEIs http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services

We concluded that there would be significant challenges to a shared 'repository' layer because of the need to integrate with other local system. Potential candidates for a shared service would be the data and catalogue / API layers and long-term storage for data. Shared training and advice services are another strong candidate being actively investigated: this could take the form of a collaboratively developed online training module for research data management building on existing examples such as MANTRA from the University of Edinburgh[16] - RoaDMaP interviewees, training participants and our training working group all demonstrated support for an online resource.

Figure 7: Potential shared service layers within an RDM infrastructure: illustrative example



2.3.10 Cultural change

We learned a great deal about researcher attitudes, awareness and culture from our case studies – including variation across subject disciplines. For some, altruism can be a driver for data management and sharing: for example, some of the researchers generating the spinal scan data in Engineering recognised it was unlikely the local research team could fully exploit the huge volume of data being generated and saw the value of managing data for subsequent cohorts of PhD students at the institution but also for sharing with a wider community, particularly as the data may be used for research to benefit human health.

On the whole, though, case study interviewees and other researchers who input to the project felt there is still a long way to go to promote a culture of sharing and re-use. Some of the key agents for change suggested were:

- methods training for early career researchers
- promoting RDM in the context of good research practice rather than compliance
- being supported to create data management plans – particularly where this is a new activity; the thought process is the most impactful element
- improving awareness and supplying tools to pre-award research data administrators
- providing a significantly improved, low barrier infrastructure around the research data lifecycle (live, storage, repository, metadata tools, archival storage)

Parallels were drawn with impact statements which were seen as having changed researchers' behaviour: specifically, they may cause researchers to think about their research questions in a different way, considering the community(ies) which may be interested in and benefit from their

[16] MANTRA Research Data Management Training http://datalib.edina.ac.uk/mantra/

research. The Timescapes case study in particular shows an approach where research data sharing and re-use contributes to the process of formulating research questions.

2.4 Immediate Impact

- **A catalyst for cross-team collaboration:** though led by the Library, all aspects of the project were collaborative, bringing together researchers and professional support staff through our governance structures, training delivery and consultation.
- **A focus to build a professional RDM network:** the project has brought together individuals with a shared interest in RDM and helped to form collaborative relationships which will extend well beyond the end of the project, helping the institution to expand its RDM provision. The project started to explore roles and responsibilities – for example, in the creation of data management plans.

Sample quotes from attendees at training events

"I basically knew nothing about data management other than simple things i.e. name file naming. I learned that data management should have structure, that it's not just something that happens."

"I will circulate the slides, DMP templates etc within my institute and will be incorporating this information into my guidance for researchers pack."

"I got to speak to people for whom RDM is their bread and butter - IT staff, faculty or school based data managers and others who are tangling with RDM issues on a daily basis."

- **Training and awareness raising:** the project has had a direct impact on colleagues who attended training event. Sample quotes are included in the box below. There are also training materials available for reuse by the community. Our training materials have been re-used by the DCC and are referenced in DCC's publication *How to Develop RDM Services – a guide for HEIs.*[17]

- **Dispelling fears and myths:** the open conversations we have had with various stakeholder quelled at least some of the fear and trepidation around RDM - for example, simply seeing examples of data management plans can be informative and reassuring. Compliance driven 'open access' to data is a particular fear; researchers are not always aware that there is a recognition of "legal, ethical and commercial constraints on release of research data"[18].

- **Support for data management planning:** several projects received direct support to create data management plans. We also raised awareness of this process and the availability of the DMPOnline tool to a range of stakeholders, including research support staff. RoaDMaP has prompted the re-evaluation of a current business process at the institution – that of data risk assessment – and secured support from our Working and Steering Groups to replace this with a more comprehensive data management planning process.

- **A use case for enterprise architecture**: the institution is looking at standardising its IT systems and processes through the 'OneIT' programme; RoaDMaP has highlighted research data management as an activity which requires systems which can readily share core information about projects, people and publications and processes which minimise duplication of effort and present a low barrier for engagement.

[17] http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services
[18] RCUK Common Principles on Data Policy http://www.rcuk.ac.uk/research/Pages/DataPolicy.aspx

- **A framework to explore the potential role of the Library in RDM**: the extent of the Library's role in RDM is still under discussion – for example, how much direct support will be provided by subject librarians or similar, whether a dedicated team should be formed for research support etc. RoaDMaP has helped progress thinking in this area. During the interim funding period (July 2013-July214) the Library will coordinate the RDM advice and support service, whether centralised or distributed and lead on repository development work.

2.4.1 Impact on RoaDMaP case studies

Input from RoaDMaP played a part in recent research bids for each of our case study areas:

- **Music**: won a £570,000 AHRC grant for *The Professional Career and Output of Trevor Jones*, starting October 2013. Feedback on the Technical Appendix was very positive and our case study lead felt interaction with the project had raised awareness of and changed the approach to research data management and that, specifically, input to the Technical Appendix may well have helped secure the bid. Our RoaDMaP case study lead, Dr Ian Sapiro, is Co-I on the project.

- **Engineering:** *Life Long Joints*[19] (2013-2018) is a Leeds-led project funded by the European Union under the 7th Framework Programme up to a sum of 13.3 million Euros. The project includes funds for a Leeds based 0.2FTE research data management post. Our RoaDMaP case study lead, Professor Richard Hall, is the project Coordinator.

- **Sociology:** *Changing Landscapes for the Third Sector: Enhancing Knowledge and Informing Practices* is a £150,000 ESRC award which is a follow on project from the Timescapes programme, makes use of the Timescapes Archive. It includes a 50% FTE Archive Officer post which will be based in the University Library. Our RoaDMaP case study lead, Professor Bren Neale, is project CO-I.
  RoaDMaP also provided a framework to review the feasibility of migrating the Timescapes Archive to a new repository platform.

2.5 Future Impact

- On-going RDM training and awareness raising, including the development of an online training module, should become embedded at the institution and move RDM towards becoming standard practice.
  Impacts on: researchers, professional support staff, trainers.
  Tracked by: monitoring uptake and feedback from training. We also plan to start a 'trainer's forum' to exchange experience and ideas.

- The introduction of a research data repository should make more research data available for re-use, increasing the University's international research profile and possibly resulting in new research collaborations.
  Impacts on: researchers, institution
  Tracked by: monitoring the growth of data repository content and usage statistics; encouraging formal citation by data re-users.

- More widespread uptake of data management planning coupled with provision of improved data storage and curation options should save researchers time and improve their visibility and impact. Raising the profile of data management planning in the context of the grants management system

[19] http://lifelongjoints.eu/

should ensure creating a DMP becomes more firmly and efficiently embedded in researcher workflows.
Impacts on: researchers, professional support staff
Tracked by: liaison with Research and Innovation Service to improve DMP monitoring

- The University RDM Policy should have greater impact in the future; once the supporting infrastructure (technology and staff) is more developed, the institution will be able to promote the policy and benefits of data sharing more actively, resulting in greater uptake of RDM services and more research data available for re-use. We anticipate developing a formal communications policy as the RDM service develops.
  Impacts on: institution, researchers, professional support staff
  Tracked by: monitoring awareness of the RDM policy at training events; monitoring web traffic to the policy text

- RoaDMaP identified the value of further data audit activity at the University to gain a fuller picture of data assets – particularly those at risk – in different Faculties. We are already working with Geography on a pilot which could act as a model across other Faculties. A fuller picture of what data is being generated should help with capacity planning and identifying any data sets at immediate risk.
  Impacts on: institution, researchers, professional support staff.

- Our case study contacts, colleagues who participated in training and those receiving direct support to create data management plans indicated their work with RoaDMaP had changed their approach to research data management. This impact should persist into the future and we hope improved awareness and practice will be cascaded to other colleagues.

3 Conclusions

Development of research data management infrastructure in UK HEIs is still at a relatively early stage: the JISCRDM programme has helped to scope what developments are required to move forward. It has been helpful to share issues and learning with the RDM community centred around the programme and it will be extremely valuable if this sharing continues.

It may be feasible to collaborate with partners to develop some aspects of a research data management service, for example, provision of archival storage and development and delivery of RDM training.

There is no obvious ‘market leader’ in the data repository space.

The project has provided a useful framework to draw together staff from different areas of the institutions – both researchers and professional support departments. As is often stated, research data management requires cross-team collaboration; the RoaDMaP project coupled with our institutional working and steering groups provided good mechanisms to enable this.

It will be challenging to tackle research data of varying size and complexity, including managing access to live and archived data.

More work is needed to understand how much University of Leeds research data can be made openly available (either immediately or after an embargo period) and how much requires ongoing, managed access.

Requirements for re-keying data into research data management systems (for example, a data management planning platform or data repository) will present a barrier to uptake by researcher and support staff; relevant systems need to talk to each other.

4 Recommendations

4.1 *Recommendations for the wider community*

It will be mutually beneficial for those involved in developing and support research data management services to share training materials: for example data management plans, examples of costing, examples where data sharing has been beneficial, examples of data catastrophes, examples arising from a whole range of different disciplinary areas generating different types of data. Perhaps DCC could have a coordinating role in gathering and categorising examples, ideally with direct updating from the community so up to date examples can be readily added.

Most institutions are still in the process of making a case for RDM support; it will be valuable to share any success stories showing a relationship between the data management planning process and bids that are successful and appropriately resourced; examples of where money has been recouped through appropriate direct costs in bids.

It is unlikely that a single department will be well placed to build a research data management service for the institution: RDM should be a collaborative, cross-team but researcher focussed service.

It is valuable to work closely with colleagues involved in methods training and ethics training to develop a consistent approach to RDM and make the best use of pre-existing training opportunities.

Raising awareness of RDM will increase demand from researchers to ingest research data, including legacy data sets and non-digital data; it is worth agreeing criteria for prioritising which data are initially included in a local data repository and considering how/whether to tackle legacy data sets and non-digital data.

As data repository services start to evolve, appraisal guidelines will be an immediate priority. Researchers are best placed to know the value of their data and what proportion of it should be retained but it may be helpful to suggest appraisal criteria to avoid wholesale deposit of data where this is not really necessary and, conversely, undervaluing research data where its reuse potential has not been fully considered. Universities could work together to develop appraisal criteria and consider to what extent they are generic or whether significant variation in approach is required by discipline.

RoaDMaP had more impact as its governance structure allowed for top down input whilst the involvement of researchers through the case studies and interviews allowed for bottom up input. Both approaches are necessary and it has proved valuable to have a range of perspective to inform thinking and planning.

4.2 Recommendations for Jisc

- **Explore different repository ecosystem models.** Further investigation of the repository ecosystem at the national / international level would be valuable: institutional data repositories; institution-based thematic or subject data repositories; national data centres. This could include (i) consideration of how limited resource can be most effectively directed (ii) the pros and cons of institutions or consortia developing specialist capacity to handle research data from particular subject areas and/or by characteristics of the data.
- **Standardise 'open access' terminology in the data arena.** The term 'open access' means different things to different people; in the data context, 'controlled' or 'managed access' may be a more appropriate description for how data is handled in practice.
- **Provide a clear guide – and recommended practice - for data licensing options**. Data licensing goes hand in hand with access. There may be a danger researchers will tend towards a conservative approach to data sharing and apply restrictions if these are offered, regardless of whether these are strictly necessary for the data or not. Spelling out the potential problems of multiple licences and re-use conditions would be valuable, as would illustrating the benefits accruing with greater levels of openness.

- **Collate examples of how RDM activity can be costed.** In our interviews and in the discussions between the Timescapes programme and the Library, resourcing and sustainability came up time and again. Institutions need more help in understanding the nature of data archiving roles, how to cost activities and infrastructure into bids and appropriate resourcing for a centrally managed data repository service.
- **Work with the community to collate more evidence of cost savings through data management planning –** for example, the costs of preparation of data for archiving where metadata is being created at the time of deposit rather than when the data is generated.
- **Sustain the JISCRDM community**. JISCRDM has been a true 'programme', enabling discussion of issues, sharing of solutions and promoting a more coherent approach to RDM across institutions through the email list but also the many events surrounding the programme. Some ongoing input from Jisc could help maintain the sense of community and facilitate continued sharing of lessons learnt, potential solutions etc.
- **Work with subject data repositories** to establish whether a report on data sets associated with specific HEIs can be supplied to help populate their data catalogue/data registry, ideally as an ongoing business process rather than a one off activity.

5 Implications for the future

The Research Data Steering Group and Research Data Working Group will continue to lead RDM development at the institution. We have bid successfully for funding to retain the core RoaDMaP staff and for some non-staff elements such as a modest amount of archival storage. Funding covers the period July 2013-July 2014. Many of the work packages started under RoaDMaP are being built upon during this next phase though the emphasis has shifted from scoping to service delivery. We plan to make a longer term business case to suggest the level of investment and resourcing necessary to enable the RDM service to scale up across the institution.

Training resources developed under RoaDMaP will be in use at the institution and so will be kept under regular review. We anticipate that any major changes to materials will be reflected in a new version deposited into JORUM.

EPrints is a well established platform with a flourishing user base and is increasingly used to manage a variety of digital content. We can see the benefit in strengthening the user community and ensuring there is a ready mechanism for sharing thoughts and resources; this may include a special interested group around research data and we would certainly participate actively in such a forum into the future.

Our Timescapes case study illustrates a thematic repository based at an institution but with close ties to the UK Data Archive. As suggested in Section 5.2, there is still much to do to map out the data repository landscape. We will continue to work with the Timescapes researchers, including on future bids, and will meet with the UKDA to discuss the on-going relationship of our two services.

6 References

GARRETT, L., SILVA, C. and GRAMSTADT, M-T. (2012) *KAPTUR: technical analysis report.* Project Report. VADS Visual Arts Data Service, a Research Centre of the University for the Creative Arts. Available online: http://www.vads.ac.uk/kaptur/outputs/

JONES, S., PRYOR, G. and WHYTE, A. (2013) *How to Develop Research Data Management Services - a guide for HEIs*. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

UNIVERSITY OF EDINBURGH. *Research data management policy.* http://www.ed.ac.uk/schools-departments/information-services/about/policies-and-regulations/research-data-policy