Activities, solutions and experiences within UK universities to meet the EPSRC research data policy
Monica Duke, DCC, Jonathan Rans, DCC
Verena Weigert, Jisc
University of Leeds

Interview with Rachel Proudfoot, Research data management advisor

20 January 2015

Interview led and recorded by Verena Weigert, Jisc and Monica Duke, DCC
Introduction to the University of Leeds

The University of Leeds is a large Russell Group institution with a strong research-led ethos. It is part of the N8 partnership, a collaborative group of northern universities. The institution has been active in the development of RDM services through funded projects and participation in community discussion, and it has a history of collaborative effort through the White Rose consortium. Rachel Proudfoot is an RDM advisor based in the library and has been involved in the RDM effort at Leeds in different capacities over the last few years.

Policy, strategy, governance and sustainability

A service is born

The University of Leeds has been working consistently on developing its RDM service, first through a Jisc-funded project\(^1\), followed by a period of interim funding from the university, and now with funding from the university allocated up to September 2016 after the presentation of a business case\(^2\).

Throughout, the service has been developed holistically, not specifically for EPSRC compliance. The university identified core service areas and developed a range of activities under each one. These activities were mapped against the EPSRC expectations to ensure there were no gaps and this practice continues into the current live service implementation. Activities and progress can be presented under the main service areas but also against the nine EPSRC expectations. The level of resource requested in the business case was linked to compliance against funders with research data management requirements, known in shorthand as RCUK+ (all the UK Research Councils, Cancer Research UK [CRUK], Wellcome Trust and the EU’s Horizon2020). Staffing resource was justified by breaking down the activities needed to set up and support the repository infrastructure and estimating the number of advice hours for data management plans (DMPs) based on the university’s research bids and awards.

---

1. Jisc RoadMAP Project (Jan 2012-Jun 2013): http://library.leeds.ac.uk/roadmap-project-outputs
2. The full business case is not available publicly yet, but a partial case can be viewed at http://blog.library.leeds.ac.uk/blog/roadmap/post/276
The first two phases of development were governed by a research data steering group (chaired by the PVC for research) and a research data working group (chaired by the pro-dean for research in the performing and visual arts faculty). The institutional policy for research data management was developed and passed through various committees, finally being signed off by senate in July 2012. The policy is owned by the research and innovation board and is due for review shortly.

One challenge for the university in developing the RDM policies has been aligning all other existing policies and the messages that they send. In a large university there is a risk that conflicting messages come from different policies; for example around what data can be retained, what can be made available and what actions are needed. A data protection policy might emphasise that data should not be kept longer than needed, which could be seen to be in conflict with a requirement to share data. It is important to make sure that researchers get a coherent, consistent message across policies. Liaison with service areas and relevant decision-making bodies across the institution is proving vital to achieve policy consistency. For example, issues of data security have been discussed with the university IT security co-ordinator with a view to agreeing a consistent approach to data classification based on - and consistent with - the university’s policy on safeguarding data. This policy classifies

3 Leeds Institutional RDM policy evolution: [http://library.leeds.ac.uk/info/377/roadmap/162/research_data_management_policy_evolution](http://library.leeds.ac.uk/info/377/roadmap/162/research_data_management_policy_evolution)
4 Leeds RDM policy: [http://library.leeds.ac.uk/research-data-management-policy](http://library.leeds.ac.uk/research-data-management-policy)
5 Leeds policy on safeguarding data: [http://iss.leeds.ac.uk/info/362/policies/782/policy_on_safeguarding_data](http://iss.leeds.ac.uk/info/362/policies/782/policy_on_safeguarding_data)
data in terms of how confidential it is, so it makes sense that advice from the RDM service links with this classification process which researchers are already being asked to do. Discussions about policy spill into the practicalities of the service so they are quite informative.

Moving to implementation

Now that the RDM service has moved into the delivery phase, governance has been taken up by the scholarly communication steering group (SCSG), which also has a remit to cover OA. The university librarian chairs the SCSG and membership includes representatives from key partner services (the university research and innovation service and IT) and several senior academics. The SCSG reports to the research and innovation board, which is chaired by the PVC research. The research data team is part of the broader research support team in the library.

The university now has four FTE in place, partly to implement the data repository and registry and partly to deliver an advice and guidance service. The new service will foster good data management practice and enable the university to comply with funder requirements.

Data management support and staff development

Determining requirements

The University of Leeds undertook a range of activities to inform the development of its services and training as part of the RoaDMaP project\(^6\). These included pilot case study work in engineering, sociology and music; repository functional requirements analysis; pilot face-to-face training sessions; reviewing the university’s approach to data management planning, including a pilot of DMPOnline and a service roadmap or timeline which informed the overall business case. It was also useful to see the outputs from other institutions and to participate in external meetings and discussions.

The university carried out a partial data audit during the RoaDMaP project but has not undertaken a comprehensive, institution-wide audit; the immediate priority is data generated as part of live grants and data underpinning publications. An open question remains around how far back to go in order to address what is already out there.

Based on the DCC/Jisc service areas diagram, the university has developed its service areas; these service areas are likely to evolve over time - See diagrams and list of service areas on page 11.

Providing support and training

The university has developed a communication strategy (which identifies key stakeholders and messages) and this is being fleshed out with more detailed communication plans. Different methods are employed to raise awareness of support and good practice: training activities, presentations at committees, one-to-one liaison (e.g. with faculty IT managers and faculty research managers). Promotion of the service with the arts faculty resulted in a steady stream of technical plans for comment.

---

\(^6\) The outputs from the project can be found at: [http://library.leeds.ac.uk/roadmap-project-outputs](http://library.leeds.ac.uk/roadmap-project-outputs)
The RDM service supports researchers who are writing DMPs, taking referrals from faculty research support colleagues but also taking direct approaches from researchers via phone and email. DMPonline is promoted as a useful tool (although it is not recommended as the only tool).

Pilot training sessions identified some of the different requirements of PhD students vs. PIs. PIs, for example, want a much more procedural session. Early career researchers with little research experience are required to undertake the ULTRA (University of Leeds Teaching and Research Award) course, which includes a half-day session on data management planning. Doctoral training centres have also been identified as an important partner; training will be delivered to one of the White Rose DTCs in conjunction with the data management services at University of York and University of Sheffield.

To raise awareness of EPSRC requirements, compliance is covered in the online information and in training sessions. The university is also planning to tie in targeted information for EPSRC grant holders more closely with the content of the RDM web site. Training and awareness has been delivered to some key support staff including research support and central and faculty IT staff.

An RDM website is available with mostly generic content. Work within the library web team is in progress to redesign the site to help people to navigate the content more easily. In particular, redevelopment will make links more meaningful to researchers’ processes.

Future plans

Although the service has successfully delivered sessions that provide an overview of the general principles of data management, applied, practical training has been identified as an area for further development: i.e. ‘this is what you need to do in practice and this is how it is plugged into university systems’.

In the next phase of training delivery, the university plans to come to the researchers, e.g. to piggy back onto existing training programmes. Some schools set aside time to prioritise bid-writing, with a series of events and presentations in the lead-up; incorporating RDM into this type of activity will help target the right people at the right time. The RDM service is also working with colleagues in the library scholarly communications team to share some training promotion and delivery. The service is fine tuning how best to match the message, audience and delivery route and also considering how often to target specific audiences: for example, whilst it was not too hard to reach professional staff on a one-off basis, there was little demand for sessions more than once a year. It is important to consider communications between training events in order to keep up the RDM momentum with particular audiences.

Infrastructure

Storage and infrastructure

The approach to live and archival storage at the University of Leeds is under active review, university-wide, led by the central IT service. Establishing the strategic ‘ownership’ of storage, particularly for archival storage underpinning the data repository, required careful discussion. Research data is only one digital asset managed at the institution and so storage provision needs to be viewed in the wider institutional context.

Leeds RDM webpages: http://researchdata.leeds.ac.uk
For long-term storage, work has been carried out with Arkivum; this has been helpful in determining storage requirements and potential solutions. At the moment, the approach to long-term storage varies by faculty; the institutional storage review will consider the costs and benefits of possible solutions, including the potential for collaboration across the N8 Consortium\(^8\). Discussion of RDM within the N8 Consortium gave rise to a reference architecture - coordinated by enterprise architects at the University of Leeds – highlighting areas of similarity and difference in RDM systems across the partner institutions. This work suggested potential areas for collaboration, with shared storage being the most promising candidate.

A team led by the library, which includes a member of the RDM team and a colleague from special collections, is reviewing the institutional approach to digital preservation.

**Repository**

The University of Leeds has set up an instance of the EPrints repository for data. Its publications are handled through the White Rose repository\(^9\) (also using EPrints software) – however, the data instance is separate from this. The decision was partly strategic – a common approach to a data repository had not been agreed across the White Rose group when Leeds secured specific project funding to address data – and partly practical, based on the known complexities of linking the publications repository with local systems across the three institutions.

The choice of repository was not completely straightforward. After a repository requirement analysis\(^10\) the university decided to go for EPrints as a pragmatic medium-term solution. An evaluation will be made in four to five years to review the choice of platform. For the data held by the university, tracking of access will be managed via EPrints, although this may need fine-tuning.

Pending decisions on an institutional approach to archival storage some storage from Arkivum is available; the workflows for determining how particular files move from researcher to storage are still being worked through.

**Data appraisal and access control**

Processes to help researchers determine which data are of long-term value and should be preserved or shared is an area identified by the university for active development. Current guidance consists of basic information on the website and referral to DCC resources. Realistically, the best people to understand the data are the researchers who generated it but RDM support staff could give them some criteria for how they could decide what to prioritise or whether they should be thinking of keeping it all. There is the potential to work with the university library special collections to identify overlaps in terms of how issues such as appraisal and value selection are addressed, thinking about both digital and non-digital data. Rachel thinks that, despite slightly different approaches, it could be very valuable to exchange lessons learned.

Consideration of access control requirements\(^11\) included discussion on the JiscMail research data management list (in May 2013), discussion with the EPrints user community and discussion with researchers. Leeds was also able to draw on several years’ experience managing the Timescapes Archive\(^12\). It became clear that some

---

8 N8 Research Partnership: [n8research.org.uk/](http://n8research.org.uk/)

9 The White Rose Group is the University of Leeds, the University of Sheffield and the University of York

10 Research data repository requirements: [http://blog.library.leeds.ac.uk/blog/roadmap/post/163](http://blog.library.leeds.ac.uk/blog/roadmap/post/163)

11 Research data repository requirements: [http://blog.library.leeds.ac.uk/blog/roadmap/post/163](http://blog.library.leeds.ac.uk/blog/roadmap/post/163)

12 Timescapes repository: [http://timescapes.researchdata.leeds.ac.uk](http://timescapes.researchdata.leeds.ac.uk)
researchers will need to keep control over who can access their data and for what purpose. The university encourages open availability of data where possible, however, in recognition of more complex data control requirements, the data management service directed significant resource to scoping and developing an access control layer for the EPrints repository platform (documented in a wiki\textsuperscript{13}). The results were presented to the EPrints UK user group in January 2015\textsuperscript{14}. Controlled access to data introduces some sustainability issues – for example, when PIs move institution or retire. The approach of a named ‘data steward’ has been suggested, which must always be updated for data with access restrictions. Discussions on this issue are being held with the faculty research ethics committees that could, for example, have a role in dealing with ‘orphaned’ data. Even more challenging is tracking whether data that sits outside the university in a variety of other repositories has been used or not.

The university will use CC-BY as a default reuse license but will offer depositors other licensing options.

\textsuperscript{13} http://wiki.eprints.org/w/Access_Control_Layer
\textsuperscript{14} Slides, EPrints user group 13 January 2013: http://slideshare.net/ULCCEvents/eprints-user-group-january-2015 (slides 115-136) and video youtube.com/watch?v=a2AtOseXCGU (2:05:08 – 2:32:32)
Metadata and identifiers

The N8 Consortium metadata group has been influential in informing the decision around metadata collection. The ReCollect plugin (for EPrints) offered a metadata specification, which largely met the requirements of the N8 group; decisions on mandatory fields and any additional fields were devolved to each individual institution with Leeds opting for a slightly different schema from the original ReCollect specification.

For the first phase of repository development, metadata will be collected via a structured spreadsheet which can then be ingested, rather than creating the metadata directly into EPrints. The rationale for the spreadsheet method is that it’s a way to collect metadata earlier in the research process, giving academics an idea of the metadata that is going to be needed for deposit. It will also act as a metadata quality control check. The metadata fields are available online.

Looking to the future for metadata, the N8 are keen to see the schema more widely adopted and are discussing with Jisc how to facilitate this via the Research at Risk initiative. Additional fields around compliance reporting have also been suggested (by University of Glasgow) and these should be in the most recent release of ReCollect (e.g. a field for last access date). For the University of Leeds, one further driver for the future would be guidance on the national data catalogue requirements, for example, work to guarantee that use of the Recollect plugin would ensure interoperability with a national service.

The University of Leeds has an account with DataCite and can assign DOIs. It is likely that DOIs minted against the university prefix will be managed centrally to begin with. ForestPlots.net is a pre-existing repository of data, affiliated with the University of Leeds, with data from all over the world. ForestPlots.net has provided a case study of assigning DOIs, working with the British Library (BL) and DataCite. Leeds requested a dedicated prefix for ForestPlots.net and DOIs have been minted against this. The data is held in ForestPlots.net’s own database. So far, DOIs have been minted via DataCite’s xml form rather than using a plugin but Leeds plans to test the EPrints plugin and review how to create DOIs in a more automated way.

Non-digital data

The N8 consortium metadata group is looking at scenarios for non-digital data, addressing how to identify what non-digital data exist, and what basic information to capture. EPrints will be providing the data catalogue for Leeds, and it does allow fields for non-digital data; it has been agreed that the data repository and data catalogue will utilise the same metadata specification. One outstanding question is how/whether the DataCite metadata fields can accommodate physical location and access information for non-digital data. Rachel would welcome moves to resolve this issue with the community.

15 Leeds Metadata Specification for Research Data: [http://blog.library.leeds.ac.uk/blog/roadmap/post/277](http://blog.library.leeds.ac.uk/blog/roadmap/post/277)
16 Jisc Research at Risk initiative: [jisc.ac.uk/rd/projects/research-at-risk](http://jisc.ac.uk/rd/projects/research-at-risk)
17 DataCite Case Study: ForestPlots.net at the University of Leeds: [bl.uk/aboutus/stratpolprog/digi/datasets/ForestPlot_CaseStudy_ForBL.pdf](http://bl.uk/aboutus/stratpolprog/digi/datasets/ForestPlot_CaseStudy_ForBL.pdf)
General comments and tips

» At institutional level, buy-in to research data management and supporting policies and governance depend a lot on a senior level champion. The University of Leeds was lucky in having an interested and supportive PVC research who chaired the research data steering group for around three years. Since the EPSRC is a key funder at the university, the RDM requirements were taken very seriously from the first letter to the VC. Before the EPSRC published its RDM policy there was already an understanding within the university of research data as an asset that needs management and of the concept of risk around data loss.

» There has been little resistance from professional support staff or from researchers so far to the underlying principles of sharing data. This may well increase in the future but there’s sufficient buy-in to identify early adopters.

» Planning infrastructure and storage is a generally challenging activity. Getting the repository up and running has taken more time and effort than originally anticipated. Researchers will have questions about storage of active data and it is important to have coherent answers regardless of who provides the storage.

» A key challenge in a large institution is the sheer number of service areas and pre-existing policies and practices that are relevant to RDM. There was a need to identify key pre-existing policies and their owners, trying to secure them as allies, seeing RDM as a shared concern.

» Looking towards the future, Rachel thinks that there could be a potential new role for the library in helping researchers to identify data for reuse. This is out of scope of the current RDM service offer but it’s possible to envisage a service that can help researchers find, choose and reuse data for secondary analysis. This overlaps with what the library does in terms of helping its users to navigate resources. As services mature, the focus will be more on impact and on how data are being re-used in practice.
List of service areas at Leeds as of January 2015

1. RDM policies
2. Reporting and communication planning
3. DMP advice, guidance and support
4. DMP templates and costings
5. Training, outreach and community building
6. External repositories and services
7. Ethics
8. Data safeguards
9. Licensing/ legal
10. Data registry
11. Data repository
12. Preservation / workflow
13. Data security (access control)
14. Active storage / IT (with central IT)