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 Edinburgh

Interview with Stuart Macdonald, RDM service coordinator

8 January 2015

Interview led and recorded by Jonathan Rans, DCC and Verena Weigert, Jisc

Thanks to Robin Rice, Edinburgh data librarian for additional comments.
Introduction to the University of Edinburgh

The University of Edinburgh is a large, research-intensive institution and member of the Russell Group. It is organised into three colleges – science and engineering, humanities and social sciences, and medicine and veterinary medicine. These three colleges are further sub-divided into 22 schools, all of which enjoy a high degree of autonomy in their governance. As of 2012-13 Edinburgh’s research body contained ca. 6,400 academic staff and ca. 11,300 postgraduate students. Stuart Macdonald is currently on one year’s secondment to research and learning services, working as the RDM service coordinator. In this role he is developing Edinburgh’s research data services, and working with colleagues from across the information services support group.

Policy, strategy, governance and sustainability

Policy and roadmap history

The University of Edinburgh was the first institution in the UK to publish an official RDM policy, making it available online in May 2011. The development of this document came about in response to a research storage paper submitted to the university’s IT committee (2010) and to the data audit framework implementation project’s (2008)’s recommendations that Edinburgh should develop training programmes, online guidance, a suite of technical services, and an official RDM policy. The decision to pursue a policy was taken by a committee of academic and support staff convened by the director of library and university collections, Sheila Cannell, to examine potential RDM services. The feeling of the committee was that some form of overarching governance, owned by central services, was required to direct local practices. The policy was drafted by Chris Rusbridge, former director of the DCC, acting as a consultant post-retirement. This policy was described as aspirational since, at the time that it was published, Edinburgh did not have in place sufficient institutional infrastructure to enable researchers to meet their stated RDM responsibilities.

Some technical infrastructure was in place, however, thanks to two prior services: the Edinburgh Data Library, which has been addressing curatorial issues around data since its inception in 1983, and the Edinburgh Compute and Data Facilities (ECDF) which provided access to both high performance computing and large storage space at cost. In EPSRC terms the most notable existing feature was an open access repository for research data, DataShare, based on the DSpace architecture. ECDF had also conducted a research computing survey in 2007 that identified data storage as a primary concern of Edinburgh researchers. It was not until the RDM programme was implemented that information services (IS) could address this requirement.

Although a policy was close to being completed in April 2011, Edinburgh had not started work on an underlying roadmap for implementation of RDM infrastructure when the EPSRC made public its requirement that one be
produced by May 2012. In fact, it was this announcement that drove the production of a strategy document. However, in line with the approach taken by many institutions, Edinburgh chose to produce a roadmap that would satisfy the EPSRC’s expectations whilst addressing the needs of the wider research community. In addition to a public-facing roadmap there is also an internal project plan document that underpins the roadmap and provides a finer level of granular detail in terms of project milestones and individual and divisional responsibilities across IS.

**Governance apparatus**

The driving force in the development of Edinburgh’s response to RDM requirements was the vice-principal for knowledge management, Jeff Haywood. He recognised the need for a high-level steering committee with cross-divisional membership that could ensure the RDM services that were being put into place met the needs of the university’s researchers without harming research competitiveness. This steering committee is chaired by Peter Clarke, professor of physics, and also includes the deans of research for each of Edinburgh’s three colleges. Jeff Haywood considered it vital that the steering committee be academic-led and that it should have strong representation from across Edinburgh’s academic community to ensure that the group decisions were practical and credible to academics.

After the development of the RDM roadmap, a second committee was convened that had responsibility for meeting the strategy’s deliverable targets and reported back to the RDM steering committee. This ‘action committee’ comprised members from across the divisions in IS (including the user services division) and also members from library and university collections, the IT infrastructure division, EDINA and the data library. The group is chaired by the director of library and university collections and meets on a monthly basis, while the steering committee convenes quarterly.

**Resourcing**

Leading on from the work on RDM policy, a business case was written to support Edinburgh’s RDM activities. It underpinned the RDM roadmap and project plans, and was developed in parallel with them. The business case framed the first phase of RDM service delivery as an IT infrastructure project, and secured c. £1.2 million of internal funding to resource requirements gathering, human infrastructure development and the piloting and embedding of technical services. Initially, the funds were roughly split three-quarters to fund infrastructure and one-quarter for staffing costs, spread across three years from summer 2012 to August 2015. The conclusion of this phase of the project coincides with the deadline for implementation of the EPSRC roadmap but there is an expectation that service development will continue after this date.

The funding has enabled the creation of several new posts and the partial secondment of others to create a cross-departmental RDM team. At present, that team comprises two FTEs from library research and learning services, two FTEs from IT infrastructure, and 1.5 FTEs from the data library. It is hard to predict what the shape of the RDM team will be in the future although there is an expectation that it will remain in some form.

As more infrastructure elements come on-line, there will be a need for support from the user services division, which will take some responsibility for advocacy and the promotion of services. Then, as these elements mature and become embedded in ‘business as usual’, there is likely to be a devolution of responsibilities for certain aspects of RDM and an expectation that support will be absorbed into existing posts.
Resource has also been drawn from within departments by engaging academic support librarians in a programme of re-skilling. There are currently 14 academic librarians across the institution’s three departments who have each received tailored training in RDM, delivered by data library and user services staff. The training programme is at a very early stage, only having been completed three months ago, but already it has enabled academic support librarians to field data management planning queries, with a focus on specific funder requirements, and to engage with researchers as part of an ongoing programme of advocacy. As the ambition for future iterations of RDM training programmes is to provide ever more discipline-specific advice and guidance, these locally-embedded academic support staff could well play a major part in its development and delivery.

Data management support and staff development

Re-skilling academic service librarians is one part of a staff development programme that aims to engage more and more support staff in the process of providing RDM support to academics at all stages of the research lifecycle. In an institution the size and complexity of Edinburgh, depth and breadth of engagement is required to ensure that the EPSRC requirements around awareness raising are met (expectation I).

RDM training at Edinburgh is still in at a relatively early stage. A structured training programme run through IS skills commenced in May 2014 and is aimed at researchers and postgraduate research students. It is closely connected to a programme of advocacy and awareness raising, run by research learning services, which identifies groups that feel the need for further training. This might be a fairly generic introduction to RDM or it might focus on particular aspects of the research life cycle. Consequently, the training is largely bespoke and organised on an as required basis.

These training sessions have been delivered in a broadly inclusive way across the disciplines although there has been varied engagement between subjects. For some, this may be down to how data-intensive the subject is and the relative importance of RDM to these researchers. In other areas, lack of engagement may reflect the extent to which researchers feel their needs are met at a school or departmental level, so that they don’t tend to engage with central services. In the college of medicine and veterinary medicine, which has always been data-intensive and has well-developed local support services and infrastructure, engagement has been slower. This is likely to be a universal challenge across research-intensive institutions and highlights the importance of understanding and accommodating existing workflows, practice and infrastructure.

In parallel with the programme of live advocacy and training events, Edinburgh has also developed a series of online resources for disseminating guidance and best practice. There has been a web portal offering a mix of local guidance, contacts and links to other resources around the web for some time. This recently underwent a major redesign in response to user feedback to present a more accessible, flatter structure.

The data library responded to the demand from researchers for online RDM training resources by developing the MANTRA suite of materials in 2010, originally with Jisc funding. Aimed at postgraduate and early career researchers, the course has proved universally popular and is currently licensed CC-BY, enabling its use outside of the institution. In 2012-13 it was endorsed by the RDM steering group as a basis for training in the RDM programme.

---

6 Do-It-Yourself RDM training kit for librarians: http://datalib.edina.ac.uk/mantra/libtraining.html
7 Edinburgh RDM training: ed.ac.uk/schools-departments/information-services/research-support/data-management/rdm-training
8 Mantra RDM training resources: http://datalib.edina.ac.uk/mantra/
With all of the training materials, both on and offline, there is a perceived need to provide ever more discipline-specific content as engagement with researchers spreads. There is a danger that, in making guidance too generic and reliant on commonsense without reference to specific disciplinary context, researchers may be discouraged from seeing central services as a credible source of RDM support. With this in mind, Edinburgh has developed a number of data-handling training exercises\(^9\) aimed at researchers using specific software packages (SPSS, ArcGIS) with a view to adding to that list in future. Developing broader discipline-specific RDM advice takes time mainly because, for it to be robust, it needs to respond to feedback and be an iterative process. There are also resource overheads in relation to maintaining and updating training materials in dynamic and fast-changing research environments.

**Data management planning**

Edinburgh has been instrumental in the development of the latest iteration of the DCC’s data management planning tool, DMPonline\(^{10}\), by contributing to an in-depth usability study that informed the ground-up redesign of the product. Edinburgh has customised the tool by developing funder templates and localised guidance that can be targeted to specific research areas. The ability to use DMPonline to provide researchers with tailored advice and contacts is seen as a possible route to connect with those who may not readily engage with central services. There is likely to be a continuing thread of development in this area with increasingly specific disciplinary content added to the Edinburgh templates.

---

9 Data handling tutorials: [http://datalib.edina.ac.uk/mantra/softwarepracticals.html](http://datalib.edina.ac.uk/mantra/softwarepracticals.html)

10 DMPonline: [https://dmponline.dcc.ac.uk/](https://dmponline.dcc.ac.uk/)
Infrastructure provision

In line with the EPSRC’s expectation VIII, that ‘Research organisations should ensure that effective curation is provided throughout the full data lifecycle’, Edinburgh has concentrated on the delivery of a suite of infrastructure elements that support data management and map onto the research data lifecycle. Some, most notably the DataShare open access repository, were in place or planned prior to the production of the EPSRC roadmap. The primary areas of service development and support are in data sharing, both in terms of active collaboration (through DataStore\(^\text{11}\)) and publishing; providing a richer choice of post-project storage environments; and developing the university’s CRIS, PURE, as a research data catalogue enabling the fulfilment of the EPSRC’s expectation V that appropriate metadata describing their datasets be published online.

Data storage is a universal preoccupation for researchers producing ever-increasing volumes of data. Edinburgh is providing all of its researchers with 0.5TB of free, managed storage for current project data. This amount was reached after gathering figures on current data holdings in selected pilot groups and is expected to exceed the needs of the majority of researchers. Those with excessive data storage requirements tend to be more engaged with RDM and more comfortable with including data storage costs in grant applications, making it feasible to charge directly for storage that exceeds the standard provision. The current price that Edinburgh is charging for storage over and above the initial 0.5TB is £200 per TB per annum.

\(^{11}\) Edinburgh storage and backup solutions:
ed.ac.uk/schools-departments/information-services/research-support/data-management/data-storage
Edinburgh will also soon be trialling a collaborative platform with potential for sharing active project data on DataStore both within and without the institution. This will be called DataSync and the pilot implementation has been built around the open-source, ownCloud platform.

Post-project preservation

EPSRC expectation VII requires that services should be in place to allow researchers to preserve their data securely for a minimum period of ten years. For relevant data that is appropriate to share over a networked connection, Edinburgh provides the DataShare repository, an open access platform based on the DSpace architecture. Recently, Edinburgh subscribed to the British Library’s DataCite service and is using its DOIs to permanently identify datasets. Previously this was achieved using the Handle System, and the two will be maintained in parallel for the foreseeable future. The repository is being put through the self-assessment process for the Data Seal of Approval trusted repository accreditation and it has now been added to the Thomson Reuters Data Citation Index.

The DataShare repository has been in operation since 2009, during which time the data library has made a number of additions and modifications to the service. EPSRC expectation VII requires that datasets held within the institution should be kept for a minimum of 10 years after the date of last access. DataShare is now capable of meeting that obligation by capturing download statistics, which are made publicly available through the interface. In addition, there have been improvements made to the depositor interface, a faceted browsing facility has been added and batch ingest of files is supported. In a drive to meet the needs of the whole academic body, improvements have been made to the upload functionality of the repository and it is now capable of accepting datasets of up to 10GB.

Not all data will be appropriate to share openly, and so there are plans to provide a dark storage option that will offer offline, secure storage, freeing up expensive spinning disc storage more appropriate for active data. This is still in the design stage and a requirements gathering exercise has already identified a lot of elements that need to be incorporated into the user interface. Questions being addressed include the authentication of access, what metadata to capture on ingest, how version control should be handled and how retention and deletion should be managed. It is this element, the need to finesse the interaction between researcher, data and archive that has slowed the project down. Compared to the ingest and user interface, the back-end storage architecture is relatively straightforward and could be provided by local hardware or might be purchased from a commercial company such as Arkivum.

Data catalogue

EPSRC expectations III-VI require research organisations to keep track of their data holdings and make available to the public certain information regarding access conditions, access restrictions and associated metadata. To address these requirements, Edinburgh is developing a research data catalogue.

Edinburgh’s data catalogue system (or data asset registry) will be run on the PURE platform taking advantage of the data content type feature (although the data upload facility will be disabled). This decision was partly informed by the fact that the research office is already using it as their CRIS. Despite the fact that this will allow existing experience to be tapped into and should ease the integration of the data catalogue and CRIS, it was not
an easy decision to make. Two requirements-gathering exercises were conducted to inform the choice and there is still some uncertainty over whether using a proprietary platform is the best option. However, given that researchers are familiar with the environment and are already using it for their profiles, reporting, grant awards and so on, it seemed sensible to position datasets in that space too.

At the moment, systems that are up and running and systems that are being piloted operate in a state of semi-isolation and addressing this is one of the major preoccupations of the development team. They are currently working with the CRIS system to identify the most efficient way of transferring metadata records across the various systems.

**Future challenges**

It is this question of scalability and interoperability that is likely to dominate the next stage of infrastructure development at Edinburgh as new services are added to an RDM suite at various stages of maturity. Having been at the vanguard of this movement meant that there were no templates to follow and few examples to learn from. Current services will need to be evaluated and re-evaluated as each new element is added. In parallel, there is a need to foster communication between departments that have typically operated in isolation, as cross-institutional workflows will be integral to the success of new RDM infrastructure. Providing a seamless user experience is a central ambition for the Edinburgh RDM service, and will need both of these elements to work for it to be realised.

Disciplinary advice is likely to be another major preoccupation moving forward and there is recognition that the scale of the problem is uncertain and potentially quite large. Having said that, there is likely to be a wealth of experience and advice already available in the different schools that could well form the basis of a suite of subject-specific resources. Potentially, a series of targeted case studies could identify what types of resources and support already exists. Edinburgh benefits in this area from having the Edinburgh Data Science initiative[^13], which operates across disciplines to foster communication and promote interdisciplinary collaboration in data science.

**Tips for other institutions developing RDM support services**

» Aim to involve academics in all aspects of RDM development from the beginning. There should be strong representation from senior academics on strategic committees, and researchers should be closely involved in ensuring that support services are fit for purpose. Academic champions with a passion for RDM can be instrumental in engaging their colleagues with the issues

» Foster communications between departments. RDM support provision is a cross-departmental enterprise and will require parts of the institution that are unused to working together to co-design systems and workflows

» When engaging with academics consider that different disciplines will require varying amounts of resource to meet their RDM training needs. Wherever possible, aim to deliver subject-specific guidance

» Interoperability should be a primary concern when designing RDM services

[^13]: [Edinburgh Data science initiative](http://ed.ac.uk/schools-departments/data-science)