Research Data Support: a Growth Area for Academic Libraries?

Robin Rice, Data Librarian and Head, Research Data Support, University of Edinburgh (UK) ORCID: 0000-0003-0214-6559.

Abstract

The ‘data revolution’ has impacted researchers across the disciplines. Not only must publications often be made open access, but data sharing is often a prerequisite for a paper to be published. Academic librarians increasingly find themselves ‘upstream’ in the research process, trying to help their users manage unwieldy amounts of data, when their comfort zone is firmly ‘downstream’ in the post-publication stage. Librarians who embrace this shift and develop their data skills should be well-placed in Library Futures.

This paper examines why the ‘upside-down’ and ‘inside-out’ library predicted by some will give research data support pride of place. Good research data management practice, increasingly required by funders, is an indicator of research integrity to which researchers are learning to aspire. They require support and services that may or may not be available through disciplinary channels or ‘in the cloud’; but localised services have many advantages.

The University of Edinburgh has a ‘mature’ research data service which has been growing since 2012 and earlier. The Research Data Support team based in the University Library has a suite of tools and training to meet user needs from data management planning, to active data management and collaboration, to archiving and sharing data beyond the life of the research project. The talk will provide a tour of its proven services as well as its three-year RDM Roadmap and its key milestones, which will hopefully offer new service providers with ideas and inspiration.

Introduction

The ‘data revolution’ has impacted researchers across the disciplines. As if the traditional work of teaching, competing for grants and promotion, doing research and publishing results was not challenging enough, researchers are required to make fundamental changes in the way they do all of these things: teaching must make use of learner ‘analytics’; bureaucratic decisions must be ‘data-driven’; research must be digital, with transparent methods, making use of ‘big data’; publishing results must be open access; and data sharing is often a prerequisite for a paper to be published.

A similar shift can be seen for academic librarians. Librarians who were taught to meet the needs of their users based on information scarcity now need to retrain themselves to help users deal with information overload. Moreover, librarians increasingly find themselves ‘upstream’ in the research process, trying to help their users manage unwieldy amounts of data, when their comfort zone is firmly ‘downstream’ in the post-publication stage.

Unsettling as it may be, these are exciting developments for the library profession. Researchers who long stopped using reference services to conduct their own searches over the Internet are seeking professional help for their research data management (RDM), and are increasingly finding that expert help from librarians. In this sense every librarian is becoming a ‘data librarian’ in some ways.
Library Futures

Although predicting the future is always a tricky business, there are some clues about the way academic libraries are going from library literature and library leaders. The idea of an ‘upside-down library’ was explored in the 1990s, in the sense of removing the hierarchy of units within libraries, so that each library process was considered equivalent and vied equally for the director’s attention, whose role was also changed to focus primarily externally.¹

However, the focus of this paper is more about how the drive towards open access and science and scholarship are changing (academic) libraries’ attention to the research outputs produced by its own members, even more than the research outputs consumed by its own members.

A 2016 radical experiment resulted from a University-wide task force about the future of the Library at MIT: the entire collections budget was placed under the management of the scholarly communications department. The Library management made this decision as “part of a broader strategic pivot in which research libraries focus more on “inside out” collections — those in fewer collections, often generated by the university, often unique to that university — and less on “outside in” collections — those we buy from external sources to make available locally, and which appear in many universities’ collections.”² Informed by theory provided by Lorcan Dempsey of OCLC, the aim is to target financial and other resources towards collection types that have high ‘uniqueness’ value: including those that are already currently highly stewarded – such as manuscripts and other special collections, but also new information types like research data, which are not currently highly stewarded in libraries.

The University of Edinburgh Research Data Service as example

Although the University of Edinburgh has not taken radical steps with its collection budget like MIT, it serves as an example of a well-invested service focusing on University members’ research data. The University of Edinburgh is a large UK research organisation, making its way up global and UK rankings within the prestigious ‘Russell Group’ of UK research universities. With departments across the spectrum of the disciplines (barring library science, unfortunately), it is particularly strong in medicine and computer science (informatics) and has a cross-disciplinary data science programme aiming to produce “world-class data infrastructure” for data-driven innovation as part of the City-Region Deal funded by UK and Scottish government.³

Edinburgh’s research data management (RDM) programme began before 2011, when its RDM policy came into force, with a focus on requiring data management plans of every new research project (https://www.ed.ac.uk/is/research-data-policy). This helps both to ensure adequate provisions are made for RDM requirements in terms of costs, and that data which can be openly shared are identified early and documented well. The policy outlines the researcher’s own responsibilities when it comes to research integrity and managing research well as part of that, and also the role of the institution in supplying tools and support in RDM across the data lifecycle, from the creation stage, to the active storage and analysis stage, to the archiving and sharing stage. At the same time training and awareness raises the visibility of both the need for RDM and the services available to the researchers. This is communicated to researchers both through an extensive research lifecycle diagram (Figure 1) and a simplified data lifecycle (Figure 2), in order for them to find the right tool at the right time of their project, or ‘user journey’.

*Figure 1: University of Edinburgh Digital Research Services lifecycle approach*

*Figure 2: University of Edinburgh simplified data lifecycle*
Governance

In order to ensure the aspirational policy became a reality, an academic-led steering group was set up to oversee the activities of the service providers in Information Services (teams in Library, IT and user support). An RDM Roadmap was drawn up (https://www.ed.ac.uk/is/research-data-roadmap) to plan for the development of the service over successive three year periods. The Roadmap is a living document that is expected to change over time, with new versions created as targets are realised. The service is currently working to the 2017-2020 Roadmap, with 32 themed and prioritised objectives and accompanying milestones and actions, as documented on the website. According to Information Services’ service management framework, the steering group chair is also the business service owner (known in other frameworks as the senior user).

Another role, the service owner, is a member of the support staff who is responsible for the liaison between the steering group and the service team, communicating requirements and feedback in both directions and reporting progress against the Roadmap and other service targets. The service operations manager has deep technical knowledge of all the systems that make up the service and is responsible for keeping them running smoothly, working closely with the service owner and other members of the service team, regardless of their organisational location.

The service is funded by a combination of core salaries budgets (e.g. existing staff who contribute to the service), RDM dedicated annual revenue (funding additional staff), capital investment for hardware, software and consultancy, development project funds that are , and a cost recovery mechanism based on charging for extensive use of services (such as more than 0.5 TB of storage space per research
project) or wholly charged services that must be costed into research grant proposals.

**Before (Create a data management plan)**

In addition to having a policy requiring data management plans (DMPs), the University provides support for creating such plans. The first port of call is the local research support and IT staff in the department; but the Library’s Research Data Support team is also available to offer assistance with writing DMPs, even at very short notice (e.g. before a grant proposal deadline). The team is familiar with UK and EU funder requirements, which are also available for researchers to look up from the service website. The University makes use of the Digital Curation Centre’s DMPOnline tool ([https://dmponline.dcc.ac.uk](https://dmponline.dcc.ac.uk)), by providing pointers from its website and also by customising the tool for Edinburgh researchers, in the form of guidance notes and a university template. The researcher can choose this as an alternative to the funders’ templates, or use for student and unfunded research). However, the tool is not necessary for writing a data management plan. The support team also collect successful DMPs (DMPs that were part of successfully funded research grants) in a variety of disciplines for reference. These are shared with university researchers through the permission of the creator, but are not published.

The importance of data management planning, in addition to complying with funders’ requirements, becomes apparent in two scenarios. First, projects that intend to collect large amounts of data (‘big data’), or data with large file sizes (such as medical imaging), must resource RDM sufficiently – including short and long-term storage costs and including a staff role for data management within the project. Similarly, costs for high performance computing and software must be included, though this is not normally considered part of RDM. Second, projects collecting personal or sensitive data must ensure they plan for adequately safeguarding the data during the life of the project (which may entail more expensive storage environments or more rigorous organisational procedures which need to be monitored), and also for what will happen to the data when the project comes to a close – whether it needs to be destroyed or retained, for how long, and who may have access to the data, or whether an anonymised version of the data can be created and openly shared. Many of the horror stories about data loss or data breaches, as well as concerns about data sharing, can be avoided through proper data management planning, hence the policy emphasis on early creation of a DMP, whether or not the funder requires one.

**During (Working with data)**

There are a number of functions a researcher requires during the active phase of the research project. A data management plan ensures these are costed from the start, but support is often needed as a project changes over time, or requirements become clearer. The University of Edinburgh, which has offered a data library service since the early 1980s (now integrated with the Research Data Support team), provides a web portal as well as assistance for discovering and re-using existing datasets, so that data need not always be created from scratch.
A hallmark of the University’s Research Data Service is its storage platform, known as DataStore, which provides one half of a terabyte of free storage for every researcher (defined as academic staff and postgraduate research students) as part of its computing infrastructure. Although storage areas can be shared and pooled within research groups, it is essential for larger research projects to include add-on costs for additional storage in their plans. These are priced to be competitive with the marketplace (cloud services).

A new Data Safe Haven facility provides additional security for projects which require it through remote server technology, and will soon be certified according to the international security standard, ISO 27001. The price for the data safe haven depends on research project requirements, which may include some form of high performance computing.

The service has developed or adopted some tools to help research projects manage data in their active phase. A university licence has been purchased for an electronic lab notebook platform (RSpace), which is integrated into the university’s infrastructure and uses the single sign-on to log in. Each lab wishing to use the platform must purchase a number of seats, then can get help to customise permissions and settings as determined by the lab leader. Jupyter notebook is also available, for projects focused on computation with statistical software or Python. An instance of Gitlab is hosted, to allow projects producing code to document, share and version their software. Researchers may collaborate with non-university based partners by using a tool to provide html access to files on their DataStore, or to sync files between a server and local computers (this is named DataSync, but it is based on the open source software, Owncloud). Research projects dealing mainly with shared documents may choose to use Sharepoint with collaborators; a locally hosted Wiki platform is also available.

**After (Share and archive your data)**

On approaching completion, researchers generally need to select an archive for data that should be kept or shared after the end of the project. This may be a national or international data archive associated with their disciplinary area (such as the UK Data Archive, or Dryad). If they wish to archive their data with the university, two options are available, for open access and closed (or restricted) access.

Edinburgh DataShare is an open access digital repository for multiple forms of research data based on DSpace. It has been available for over ten years and holds over 2,500 datasets from all three colleges. Its features ensure data are highly discoverable – including in Google Scholar and Google Dataset Search – with a quality-assured metadata record, a suggested citation and a digital object identifier (DOI). University researchers or their partners may upload up to 20 GB in any number of files per dataset directly through a drag and drop web interface, or they may receive help to ‘batch upload’ data items up to 100 GB through a back end interface. University researchers may deposit any number of datasets for free; this helps to incentivise open sharing. A default open licence (CC-BY, or Creative Commons Attribution licence) is added, and item and collection user statistics are openly available to see how often a dataset has been viewed or downloaded. Items may be linked to papers or other versions of the data with a URL. An embargo may be
enabled to restrict access for a limited period of time, such as until a corresponding paper is published.

The DataVault is a secure, long-term retention solution for data not suitable for publishing openly, including confidential or sensitive data. It was launched at the start of 2019 to fill a long-identified gap in the service, to help researchers meet their obligations with their funders and human subjects. While it depends on cost recovery (researchers need to include the costs of long-term retention in their grants to pay for storage), it is cheaper than the active disc-based storage, and again, designed to be competitive with cloud-based options. The system is comprised of:

- two separately located tape copies managed in a similar way to disc back-up tapes by the university’s IT Infrastructure at its two data centres;
- a third copy ‘in the cloud’;
- a web interface used to upload and retrieve deposits organised within a research project’s ‘vault’;
- a publicly accessible metadata record describing the contents of the project’s vault with a citable DOI identifier;
- a database that keeps track of relations between files and metadata, as well as encryption keys.

The integrity of the data is monitored through fixity checks. Unlike DataShare, where a copy can both be owned by the university and downloaded by the data creator after they leave the university, there is an assumption that items in the DataVault eventually become university data assets. This is necessary because the long-term nature of the holding (ten years or more) means that the original data creators may have left the university before the end of the retention period. When this occurs a new member of staff, nominated by the original data creator or the head of school, becomes the data owner and makes decisions about the data – whether to honour an access request or to confirm deletion (or not) at the end of the holding period. If the new data owner is unable to decide whether a vault’s contents should be deleted or retained at the end of the holding period, the decision may be escalated to the Library’s Special Collections team, who will decide whether to add the item to their own collections.

One further service completes the archival offerings – the use of the University’s CRIS (Current Research Information System) as a data asset register. As part of advocacy around open science or open research, the Research Data Support team recommends that researchers link their datasets with other research outputs such as articles and create metadata landing pages for their datasets in the CRIS, which is publicly searchable at https://www.research.ed.ac.uk. This applies particularly to datasets archived in external repositories, as records associated with DataShare or DataVault are imported into the CRIS automatically.

**Training and support (across the research lifecycle)**

A training and outreach programme accompanies the service suite, which is essential to raise awareness both of RDM requirements and the services available. The Research Data Support team also acts as a second line team in Information Service’s helpdesk service, so that any user enquiries about RDM are answered swiftly and
expertly. Drilling down, this aspect of the service encompasses the following activities, which are spread across the team, given they are time-intensive.

1. **General RDM support**: Answer enquiries by email, phone or appointment; track through central helpdesk system.

2. **Online training (Research Data MANTRA, Research Data Management and Sharing MOOC)**: Researchers and students learn online at their own pace or with a cohort of peers to earn a certificate through our open educational resources (see figures 3 and 4).

3. **Scheduled and bespoke training**: Students sign up for a scheduled workshop or academic staff may request a special training session tailored to their research group. Courses cover understanding RDM, advanced topics (e.g. working with sensitive data; writing a data management plan), and hands-on sessions learning to handle data with particular tools (e.g. SPSS, NVIVO, OpenRefine, ArcGIS/QGIS). The team also works with others in the university to offer Data Carpentry and Software Carpentry courses on an ad-hoc basis.

4. **The Research Data Service website ([https://www.ed.ac.uk/is/research-data-service](https://www.ed.ac.uk/is/research-data-service))**: A section of the University/Information Services website that links to all of the tools and support available and provides limited additional information (such as service policies) and guidance, including quick guides and video case studies. The aim of the website is to allow users to inform themselves and get started using the services on their own, as much as possible. A secondary aim is to avoid information overload by providing essential information only - in a succinct and easily understood form.

5. **Blog and promotional materials**: The Edinburgh Research Data Blog ([https://datablog.ed.ac.uk](https://datablog.ed.ac.uk)) provides a window on the team’s current activities and offers a channel for writing short, topical posts. There are both internal and external subscribers to the blog. The team has created a print brochure and other promotional materials and handouts.

6. **Regular meetings and special events**: The team meets on a regular or annual basis with those operating in the colleges and schools of the university to update both academic and support staff about the service, and hosts special events such as the annual Dealing with Data day-long event each autumn and the irregular Research Data Workshop series focusing on challenging or innovative topics in RDM.
Figure 3: Homepage of Research Data MANgement TRAining or MANTRA (https://mantra.edina.ac.uk)

Figure 4: Research Data Management and Sharing MOOC, by University of North Carolina and University of Edinburgh (https://www.coursera.org/learn/data-management)
A maturity model for RDM services

The University of Edinburgh’s is an example of a fairly mature research data service, having been built over time, since the 2011 RDM Policy was passed and even before. Librarians wishing to emulate such a service need to build up their expertise and offerings gradually, ensuring their efforts are matched by the expectations and requirements of their users. Andrew Cox and the co-authors of the maturity model pictured in Figure 5 demonstrate how a library can begin at ‘level 0’ by conducting needs assessments using tools such as data audits and user surveys, while simultaneously developing expertise in staff through professional development activities. This then leads to a ‘level 1’ maturity characterised by minimal compliance with funders’ requirements and development of an institutional policy.

Figure 5: A maturity model for RDM services

While maturity of the service is in the development stage, a re-engineering of library structures and staffing occurs, to place additional resource into building librarians’ skills and changing their roles to be able to meet increased demand. The effect on the user community is that of capacity-building, with RDM and data literacy training being offered and the advisory services becoming more sophisticated and responsive. In the highest “extensive” level of the maturity model, the service is characterised by “policy, insight and capability,” wherein the user community experiences cultural acceptance of changed, embedded practices in RDM. The library will focus its energy on data stewardship, typically building a data repository service with associated

---

technical support, and potentially be involved in a wider community of providers through participation in shared services.

**RDM and the Open Science/Research Agenda**

While a research data service may mature over time, it is also the case that the wider research environment changes over time, and the service must adapt its key messages and activities to respond. The Open Science or Open Research agenda has now inarguably become a key driver for good practices in research data management, with an emphasis on data sharing. But Open Science encompasses a whole range of researcher behaviours, which in order to realise the full value in publicly funded research and generate improved public trust in science and scholarship, is seen as needing to change. The European Union funded project, FOSTER Open Science, explains that:

> Open Science is frequently defined as an umbrella term that involves various movements aiming to remove the barriers for sharing any kind of output, resources, methods or tools, at any stage of the research process. As such, open access to publications, open research data, open source software, open collaboration, open peer review, open notebooks, open educational resources, open monographs, citizen science, or research crowdfunding, fall into the boundaries of Open Science. Even though, especially for the library and information domain, the focus is usually placed on two of these movements: Open Research Data and Open Access to scientific publications.\(^5\)

**Open data and FAIR data**

The benefits of sharing the underlying data from a research project, especially those underlying published research results are many (see Figure 6). However individual researchers may not accept that making their data publicly available is beneficial to them. In some senses the benefits accrue to others – funders who do not need to pay for repeat studies; other researchers who can make do with secondary data instead of spending time on collecting their own; citizen scientists; data journalists; even, in some cases commercial companies. This is why it is crucial for librarians to emphasise the benefits of data sharing and not only compliance factors. For example, there is scattered evidence that papers where data are shared are more highly cited.\(^6\) However, as the Open Science movement grows, the distorted academic reward system based on publish versus perish and false metrics such as journal impact factors is meant to be overcome by other values, so for example, researchers can be rewarded for highly cited datasets on their own merit, not only for formal publications.

---

\(^5\) [https://www.fosteropenscience.eu/content/what-open-science-introduction](https://www.fosteropenscience.eu/content/what-open-science-introduction).

But while open data may still have its detractors, a newer concept is emerged which is even more difficult to find fault with: make your data FAIR (findable, accessible, interoperable and reusable). According to the EU-funded GoFAIR project, the acronym can be explained as follows:

**FINDABLE**: “Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services.”

**ACCESSIBLE**: “Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.”

**INTEROPERABLE**: “The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.”

**REUSABLE**: “The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.”

Among other things, this means that data which are not appropriate to be openly shared, such as personal and sensitive data, do not have to be. However, the metadata should be open and discoverable, and the instructions for requesting access should be clear – and preferably even machine-actionable, with full documentation made available in order to be able to reuse the data when a request is approved. For example, the European Commission has described its data sharing

---

policy for Horizon 2020 funded research projects as “open by default”, or “as open as possible, as closed as necessary.”

**How do librarians need to reskill for data support?**

There is some concern that library schools do not know how to prepare the librarians for the future of Open Science and Scholarship, and equip them with the digital and research data skills that library users of the future will expect of them. According to a meta-analysis examining results of similar surveys in four countries (Australia, Ireland, Netherlands and United Kingdom), it is indeed both data curation skills as well as data description and documentation abilities that score high, along with legal, policy and advisory skills, as well as knowledge of a variety of research methods (see Figure 7).

*Figure 7: Changing skills and priorities in academic libraries? (A. Cox, et al)*

This might seem to point to the need for some academic librarians to specialise in data skills, or become dedicated data librarians. Yet in another study by Lisa Federer, librarians who do data-related work were surveyed about their work and educational backgrounds and asked to rate the relevance of a set of data-related skills and knowledge to their work. The surprise here was that this although traditional library skills were not rated highly by this group, ‘soft skills’ were rated very highly:

---


• “Personal Attributes” were the most highly rated category overall (70% respondents ranked Very important +).

• “Library Skills” were the lowest rated category (40%).

• The top five rated elements were “Developing relationships with researchers, faculty, etc.”; “Oral communication and presentation skills”; “Teamwork and interpersonal skills”; “Written communication skills”; and “One-on-one consultation or instruction.”

• The bottom five rated elements were “PhD or doctoral degree”; “Professional memberships”; “Cataloging”; “Graduate degree in a [subject discipline]”; & “Collection development.”

**Getting started in RDM support**

In addition to the learning resources mentioned in this paper, there are some very good starting points for academic librarians wishing to make a start in provision of research data services. These include:

1. A top ten list of recommendations for libraries to get started with research data management from LIBER, [https://bit.ly/2NuUhAs](https://bit.ly/2NuUhAs)

Having mentioned LIBER above (the Association of European Research Libraries), the LIBER 2018-22 strategy offers inspiration in terms of strategic goals for academic libraries to pursue support for new norms of scholarly communication. These are not solely focused on support research data but it is striking how much digital and data skills play a role in their view of future libraries (see Figure 8).
LIBER’s “vision for the research landscape in 2022 is that the role of research libraries will lie in Powering Sustainable Knowledge in the Digital Age:

- Open Access is the predominant form of publishing;
- Research Data is Findable, Accessible, Interoperable and Reusable (FAIR);
- Digital Skills underpin a more open and transparent research life cycle;
- Research Infrastructure is participatory, tailored and scaled to the needs of the diverse disciplines;
- The cultural heritage of tomorrow is built on today’s digital information.”

Given the timeline set by LIBER, there is surely much work to be done by all of us!

---