DOI: https://doi.org/10.14324/000.learn.15

Case Study 14

The Research Data Storage Service at UCL – A LEARN Case Study

Author: James A J Wilson (Head of Research Data Services UCL)
Email: j.a.j.wilson@ucl.ac.uk

14.1. DATA STORAGE DURING THE 'ACTIVE' PHASE OF RESEARCH

The active phase of a research project comprises the generation or collection of data, its processing, and its analysis. If data is well managed during this phase then it can considerably simplify the job of preparing data for longer-term preservation and access after the end of a project, but it is not easy for institutions constructively to intervene in many elements of research, which are often highly specific to the requirements of a particular project. Data storage is the one element that almost all researchers depend upon, and where institutions can offer a generic central service. However, even in this realm there is a wealth of options available to most researchers, from laptop hard drives and memory sticks, to commercial cloud services such as Dropbox.

By providing researchers with a storage service that is both easy to use and includes helpful collaboration mechanisms, an institution can however gain some measure of control over how their data assets are managed, and facilitate the smooth path of data and associated metadata through the research data lifecycle.

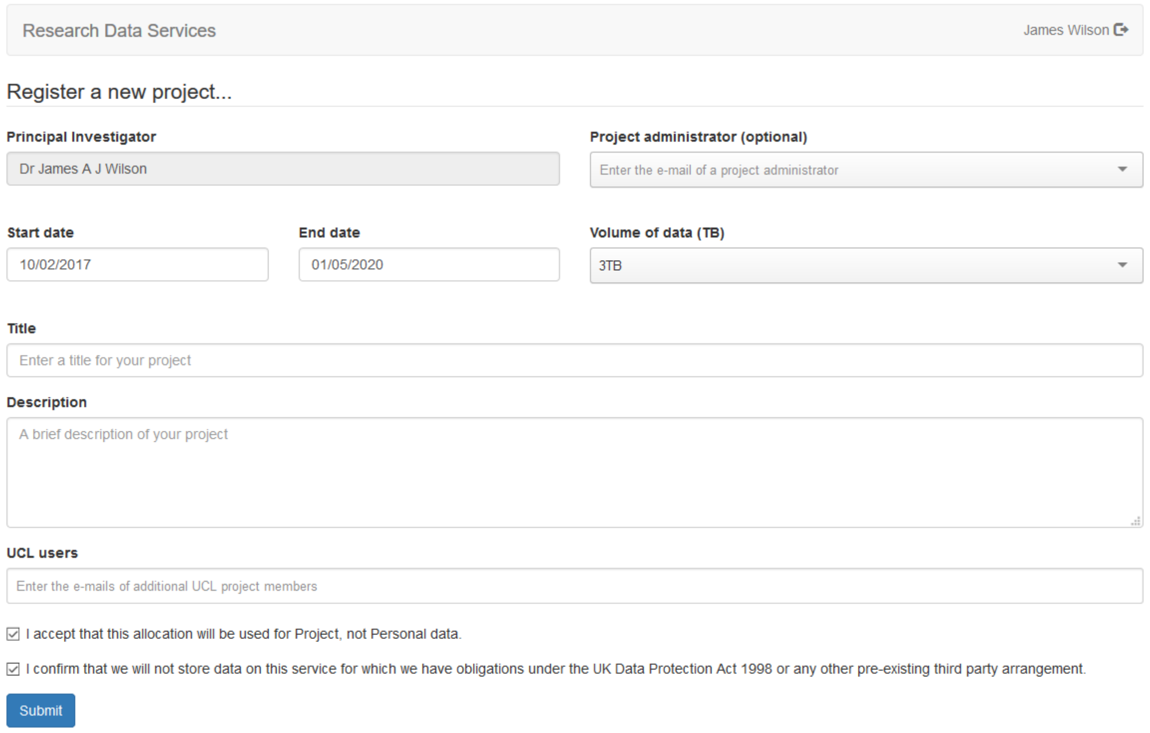
14.2. THE RESEARCH DATA STORAGE SERVICE AT UCL

The development of the Research Data Storage (RDS) Service at UCL was motivated from the outset by the necessity of assisting researchers to comply with the requirements of research funders. UCL sought to develop a data storage service that had the 'resilience and disaster recovery to assure the safety of research data'; 'multiple and intuitive user interfaces to meet a broad set of user experiences', a 'service wrap to make the Service useful to more users', and the 'capacity to increase the user base across UCL'. A tender for physical storage to enable the objectives of a data storage service was issued in 2012, and the service opened to researchers in June 2013.

Use of the service has grown exponentially since that time. As of December 2016, the service hosts approximately 760 TB of research data before replication and redundancy, 1.791 PB in total. All faculties at UCL have at least one project that is using the service.

The service is offered to research projects, rather than individual researchers. This helps with the assignment of useful metadata, as projects can be cross-referenced with administrative information held in grants databases and other UCL information sources. In practice, the service does not prohibit the creation of unofficial projects, as that would effectively proscribe the use of the storage by 'unfunded' research, a mode of working which is common in the humanities and social sciences.

When signing up for an allocation of project storage space, the authorisation of a Principal Investigator is required. The PI must vouch that no personal data (as opposed to research data) is held in the system, and that they recognise their legal obligations under the UK Data Protection Act 1998 and otherwise.



Research Data Services James Wilson

Register a new project...

Principal Investigator: Dr James A J Wilson

Project administrator (optional): Enter the e-mail of a project administrator

Start date: 10/02/2017

End date: 01/05/2020

Volume of data (TB): 3TB

Title: Enter a title for your project

Description: A brief description of your project

UCL users: Enter the e-mails of additional UCL project members

☑ I accept that this allocation will be used for Project, not Personal data.

☑ I confirm that we will not store data on this service for which we have obligations under the UK Data Protection Act 1998 or any other pre-existing third party arrangement.

Figure 14.1 The RDS New Project Registration Page

To be assigned a new project, the PI must also provide a start/end date for their project and some basic descriptive metadata. Projects can request between 1 and 5 TB storage, or contact the service directly if they need more. The minimum allocation of 1TB reflects the fact that the service was originally developed with large-scale data users in mind, as this community was least well served by alternative solutions, although the Storage Service is available to all UCL researchers however much data they anticipate generating.

14.3. UNDERLYING INFRASTRUCTURE

There are two different storage technologies under the bonnet of the RDS Storage Service: *General Parallel File System* (GPFS) block storage; and Web Object Storage (WOS). This was seen as a good combination, as the fast GPFS component can cater for users who require data to be staged to UCL's high-performance computing facilities, whilst the highly scalable object storage provides a cost-effective way of managing the bulk of UCL research data. The *Integrated Rule-Oriented Data System* (iRODS) is used as the management layer for data in the object store.

14.4. SUPPORT REQUIREMENTS

Besides the need to keep the infrastructure up to date and ensure that the service is running smoothly from a technical perspective, the RDS team works with UCL Library Services to assist researchers with interesting use-cases to make the most of the service by ensuring their workflows are rationalised. At present, some common administration processes, such as changing permissions in project groups, are also still a semi-manual process, although web interfaces are being developed to allow users to do more of this themselves.

14.5. COSTS AND PRICING

At the time of writing the Research Data Services team consists of 4 full-time employees (4 FTE), although not all of this staffing resource is dedicated to keeping the storage service ticking over. Monitoring, patching, bug-fixing, service communications, support and consultancy, and service management take about 2.5 FTE at present, with the rest of the time going towards future service development (including a UCL institutional repository), a re-architecting of the present service, and technology monitoring and assessment. An unusually high proportion of staff time over the last year has been spent dealing with issues affecting the object storage. Once the service is more mature, and more of its administrative processes automated, we would expect it to require less staff time to maintain. In addition to the core team, the service requires a small amount of resource from the UCL helpdesk team and the Data Centres team.

Hardware and support costs for a storage service will vary according to the specific deal arranged with the supplier(s). The current RDS capacity was achieved via two purchases: an initial purchase of just under a petabyte of GPFS storage and 240 TB of WOS storage, plus servers, support, and other small items of equipment, for around £740,000 in 2012; and an expansion of 2.88 PB of WOS for a little under £600,000 in February 2014. 1.2 PB of this was later converted to GPFS.

The service itself is currently offered free of charge to UCL researchers, although those with particularly large requirements (>10TB) are asked to contribute to costs if they are able. As the service scales up, this model is unlikely to remain viable, so a new pricing model is currently under development to ensure long-term sustainability.

The new pricing model will almost certainly allow a storage allocation up to a certain point free of charge, with charges applying for quantities beyond this as yet unset level. This should enable small and unfunded projects to continue using central storage, with all its benefits both to researchers and institution in terms of being able to manage data over the long term. More data-intensive projects, on the other hand, will be expected to include their required data storage capacity in their grant applications – passing their exceptional costs on to the research funder.

Although demand for the service is anticipated to continue to grow exponentially, the costs are expected to be offset in part by the falling price of storage. We are seeking to move to a purchasing strategy of buying storage according to more of a just-in-time model in future, as it makes little sense in owning constantly depreciating capacity standing idle. It is possible that some sort of cloud capacity will be used as well, but it is recognised that the costs of cloud storage add an unpredictable and potentially expensive component to the service model.

14.6. FUTURE REQUIREMENTS

At present, the RDS Service is a push-in / pull-out service. However, many of our users want to be able to use their allocated storage space as though it were available as a mounted drive. This prospect is challenging given the large file sizes the service needs to cater for, but various technologies are being assessed for suitability.

Other improvements and functionality that users have requested include:

a. File versioning
b. Dropbox-like sync and share functionality
c. The ability to add non-UCL collaborators to projects (which is currently possible, but only by adding the collaborator as an honorary member of UCL, which is a bureaucratic process)

As of December 2016, the RDS is engaged in a major project to expand capacity and better address user requirements.

14.7. LESSONS LEARNED

Some things to consider when setting up a storage service for active research data:

- Ensure that your choice of underlying storage technology is mature and reliable – this is a situation where being an early adopter is not necessarily a good strategy;
- Have a clear policy as to what the service can and cannot offer;
- Ensure a daily back-up is in place;
- Run induction sessions to understand new users and their requirements;
- Communicate clearly the benefits of institutional storage over personal storage;
- Recommend a single graphical interface for less technical users, plus programmatic access for the more technically adept;
- Invest time in developing a clear reporting system that is independent from the underlying infrastructure;
- Understand how your institution’s identity and group management systems work;
- Have a plan B for if something goes catastrophically wrong!

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 654139.