IJDC | *General Article*

Establishing a Research Data Management Service at Loughborough University

Gareth Cole
Loughborough University

Abstract

In common with most UK universities Loughborough University needed to be compliant with the EPSRC Data Expectations by May 2015. This paper explains the process the University went through to meet these expectations. The paper also demonstratea how University senior management took the opportunity to look beyond compliance with EPSRC requirements. Project staff were challenged to identify a solution which would help to increase the University's research visibility and reach. The solution to all of these challenges is an innovative and ground-breaking relationship between the University and three external partners. Investment has also been made in professional services staff to help manage and oversee the service. This paper explores the ways in which each element of Loughborough's research data service helps to reduce the burden on researchers, how much of the infrastructure is invisible to the research community, and how the service is being embedded in existing infrastructure and workflows.

Accepted 24 February 2016

Correspondence should be addressed to Gareth John Cole, Pilkington Library, Loughborough University, England LE11 3TU. Email: g.j.cole@lboro.ac.uk

An earlier version of this paper was presented at the 11th International Digital Curation Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

In common with most UK universities Loughborough University needed to be compliant with the EPSRC Data Expectations by May 2015. This paper will explain the process the University went through to meet these expectations. The paper will also demonstrate how University senior management took the opportunity to look beyond compliance with EPSRC requirements. Project staff were challenged to identify a solution which would help to increase the University's research visibility and reach. The solution to all of these challenges is an innovative and ground-breaking relationship between the University and three external partners. Investment has also been made in professional services staff to help manage and oversee the service. This paper will explore the ways in which each element of Loughborough's research data service helps to reduce the burden on researchers, how much of the infrastructure is invisible to the research community, and how the service is being embedded in existing infrastructure and workflows.

Pilot Work

In autumn 2012 Loughborough formed a Steering Committee to examine which technical solutions would best answer the EPSRC expectations and the additional requirements from Senior Management. It was decided from the outset of the decision making process that academics and researchers should be involved in the choice and then the development of any chosen solution. This was essential to ensure not only academic 'buy-in' for any solution but also to enable the completed solution to function as required by the research community, rather than how Professional Services staff felt it should work. In order to be a success at Loughborough the solution needed to work for real researchers. Having this academic involvement has also proved useful in the advocacy of the chosen solution as well as in user testing.

A number of solutions were investigated and the choice quickly came down to two: figshare[1] and Arkivum.[2] However, neither solution offered everything Loughborough wanted. Arkivum was very good at preserving data, but did not have a web presence. As such, it could meet the EPSRC requirement to preserve data but not the additional University requirement of increasing research visibility. Figshare had the necessary web presence but could not offer the same guarantees on preservation as Arkivum. Figshare also had the advantage of being a member of the same portfolio of companies as Symplectic[3] who already provided Loughborough's current research information system (CRIS), locally branded as LUPIN. Faced with the Arkivum or figshare dilemma, the Steering Group decided to try for the best of both worlds and ask the two companies to work together to offer the best possible solution for Loughborough. The first joint meeting took place in September 2014.

Other options considered included extending the functionality of our DSpace Institutional Repository to also act as the data repository and to use in house storage or Arkivum as the storage solution. However, it was clear that DSpace was not as suited to

1 figshare: http://figshare.com/
2 Arkivum: http://arkivum.com/
3 Symplectic: http://symplectic.co.uk/

Loughborough's priorities as regards visibility, and also that the support which IT Services currently provide could not be guaranteed to continue (for example, DSpace support is limited to a couple of IT staff who also have other responsibilities). In addition, as regards storage, Arkivum was not only more scalable than in-house storage but also cheaper as a result. If Loughborough had gone for in-house storage then we would have faced the difficulty of knowing how much to buy: too much storage and the outlay would not be seen as worth it by senior managers, resulting in difficulties developing the service further in the future; too little storage and we would have faced the problem of having to buy additional storage, potentially at greater cost. Additionally, in-house storage would have added to the burden on IT staff.

As stated, we decided to ask the two companies to work together. However, even before we got to this stage we needed to source the necessary funds from the University. Two options were taken to the University Operations Committee responsible for strategic decisions. Option one was that which was ultimately chosen i.e. figshare and Arkivum; option two was to use Arkivum storage and DSpace as the data repository. As can be seen, the use of in-house storage was not presented to the Operations Committee.

Symplectic Elements (LUPIN) is an established system at Loughborough and academics regularly use it to deposit their publications into the Institutional Repository. In order to limit the number of systems academics will have to use it was decided to ask Symplectic to develop the functionality to integrate LUPIN into the data deposit solution. Thus, the proposed solution would involve academics depositing publications and data via LUPIN into the University's institutional[4] and data repositories[5] respectively. This would mean that, regarding deposit, the academic would not have any additional workflows or systems above those they already use.

However, the solution is even more refined than that described above. If they choose to, academics can deposit their data directly into figshare. Elements will then harvest the metadata from figshare and automatically create a record in the depositing academic's LUPIN page. This is possible as the two systems use identical identifiers. Academics will not have to 'accept' the data deposit before it is added to their LUPIN profile (co-authors will still have to accept the record) removing additional workload from their workflow.

Although we haven't yet implemented the deposit via LUPIN functionality, this has not proved to be a problem. Initial conversations with academics and researchers have highlighted that they prefer the feel of the figshare data repository and, as such, even when we implement the LUPIN-to-data repository link the depositors may still continue to use the figshare interface. This may well be because we have deliberately decided to ask for less metadata at the point of deposit and so is seen as an 'easier' and less labour intensive option.

Once data is deposited in figshare it is automatically transferred to Arkivum servers, where it is preserved. This transfer is done without any input from the academic and is invisible to them.

We have recently implemented a curation workflow step. This was an important development for the University and has enabled us to proceed with a full launch of the data repository in January 2016. Before the development of the curation workflow researchers were only able to deposit data if they had been added to the HR feed for the repository. As such, we were able to limit who had access and who could deposit. The reasons for this were:

4 Loughborough University Institutional Repository: https://dspace.lboro.ac.uk/dspace-jspui/
5 Loughborough University Data Repository: https://lboro.figshare.com/

1. As there were no checks in place, as soon as a researcher published their data record the record was publicly available. All the staff involved in the project were wary of having this as a permanent solution because of the risk of improper release of information;
2. The solution was still new and we were working on improving it at regular intervals. As such, the solution's look and feel was changing regularly and we felt this would not be helpful with many users;
3. If something were to go wrong with the solution we would only have to work with a small number of researchers;
4. We didn't know how researchers would react to the solution and so it was useful to know those depositing personally, and to work with them to improve and iterate the solution.

Before the curation workflow was in place, we only enabled researchers who had been to training sessions or had been in extensive communications with the Research Data Manager to deposit into the data repository. As such, we felt that we had adequate protections in place to limit the risk of improper release of information. This scenario worked effectively for the small number of researchers we were working with but was not scalable to the whole institution.

Since December 2015 we have had a curation workflow. This has worked well and has enabled us to include all staff and PGR students on the HR feed fed into the repository. Consequently, we now have around 4,500 members of the University who can deposit if they wish.

During the development of the project there were a number of debates and discussions internal to the University about what checks should be put in place (if any), who should conduct the checks, and what the checks should involve. The researchers the project spoke to had a range of opinions. These consisted of those researchers who wanted absolutely no one to check the data before it was made public. The thinking behind this was that only specialists in the particular field of research would understand the data and the researchers didn't want an administrator or a senior academic in either the Library or their own School checking on their work. These researchers also didn't want a non-expert (as they saw it) "approving", or more critically, "rejecting" their work. We also heard from the opposite viewpoint where researchers would have been perfectly happy to have colleagues in their School or department check the intellectual side of the data and then colleagues in the Library check the metadata, legality etc. of the deposit.

Both of these extremes had disadvantages as far as the project team was concerned. Firstly, if there were no checks then improper information could be released. This was not acceptable to either the project team or senior managers at the University. The second extreme, although it reduced the risk of improper release, also carried with it disadvantages. Colleagues in the academic schools were not clear who could, or would want to make the intellectual checks. The obvious candidates could be the Associate Deans for Research (ADRs – one per School). However, what would happen if the ADR was absent for an extended period of time? How much time could a senior academic actually devote to checking all the deposits for their School? How could an ADR be expert in all the types of data being deposited from their School? Would they actually have any more idea than a research data colleague in the library, IT services or research office? In addition, having two checks before deposit would require additional infrastructure and workload, which the project was trying to avoid where possible.

In the end, the project managed to reach a compromise where it was agreed that a 'sanity check' would be conducted by the Library. Even those researchers who said that they didn't want any checks on their deposits were actually quite happy for someone to check for typos etc. As such, it was agreed that there would be a light-touch 'approval' process. This placated all sides of the discussion. It reduced the possibility of material being improperly released yet at the same time meant that the intellectual content of the material was not challenged or questioned by non-subject specialists.

Consequently, Library staff (predominantly the Research Data Manager) conduct a check on deposits before they are made public. The depth of these checks depends on the type of data deposited and whether the depositor has made numerous deposits. For example, if a depositor included a commercial partner as a funder of the research and the deposit wasn't marked as confidential or embargoed then we would contact the depositor (and potentially the PI of the project if not the depositor) to check whether the data could be released. This is particularly the case if a more junior colleague was the depositor. In addition, researchers would be contacted about data which appears to include human participants to ensure that either consent had been agreed to make the data available or that it had been adequately anonymised. If the data is obviously not either of these categories then the checks made include establishing if there are keywords/tags, checking that the description adequately covers what is being deposited, and establishing if there is an associated article so we can add the DOI.

So far, this curation workflow is proving successful and adequately light touch (but also rigorous) for both the academics involved as well as Library staff.

There are, however, some limitations with the existing workflow. For example, feedback and experience has shown us that academics who deposit datasets that will be updated over time (e.g. an evolving model or a longitudinal study) were unaware that once they had chosen a title they were not able to modify it later down the line. In order to keep the advocacy message clear this was not something we had been including in our presentations or discussions with academics, as we felt it wouldn't be one of their priority concerns. However, because of the experience with the pilot group and early adopters we will be modifying our communications to include this reality.

In addition, we are aware that we may face difficulties in the future with the comparatively limited amount of metadata we are collecting from depositors. We only have five mandatory fields (to ensure compliance with the DataCite schema). These fields are entitled: 'Title', 'Authors', 'Categories', 'Tags', and 'Description'. In addition, depositors need to select from a list of licences before the record can be published (this currently defaults to CC-BY-NC). We also have fields for 'References' and 'Funding' but these are not mandatory. However, most academics are still completing these fields and we check whether any information needs to be added to these fields before a record is made publicly available. As expected, the quality of the metadata (and supporting documentation) has varied. We are hopeful that as the advocacy and training being conducted by Library and Research Office staff increases and reaches more researchers that the quality of the metadata will increase. We do expect there to be some difficulties, for example if a researcher needs to be compliant with a particular schema and wishes to deposit in the Loughborough figshare repository rather than a specific subject repository. We have already received questions along these lines from researchers working with geo-spatial data.

Other functionality within the repository also aims to reduce the workload for researchers. One of the key aspects of making research data available is to link published outputs with the data which supports them. As part of this, many funders

(such as RCUK) require a statement to be included in published outputs about where the underlying data may be accessed. EPSRC go further than this and state that:

> 'Where the research data referred to in the metadata is a digital object it is expected that the metadata will include use of a robust digital object identifier' (EPSRC, 2011).

At Loughborough we work with DataCite to mint DOIs for all of our data records. This allows the data to be easily cited and relevant metrics to be drawn from it. However, the functionality goes further than this. One of the issues researchers have raised in the past is the order in which they need to do things. For example, if the DOI of the associated data needs to be included in the paper but the researcher does not want the data to be published until the paper is published, what should they do? Our functionality at Loughborough provides a couple of solutions to this. Firstly, a researcher can reserve a DOI for a data record without publishing it. This involves the researcher completing a data record (or even just creating a skeleton record) and then clicking on the 'Reserve DOI' button. This means that the researcher can include the DOI of the associated data in the paper but not make any details of the data available. The other two options for the researcher revolve around embargoes. Functionality is provided to enable depositors to embargo either the whole record (i.e. the metadata is also invisible) or just embargo the data files (i.e. the metadata is visible). These three options allow the researcher to include a DOI in the paper before the supporting data is published. Once the researcher has chosen which option to use, it is a very simple process for them to follow and the technical process is hidden from both the depositor and the repository administrator.

Further Development

As can be seen from the discussion above, we have continued to develop the product during our pilot phase. To an extent this was forced upon us. We wanted to be as compliant as we could with the EPSRC deadline of 1st May 2015. As such, it was decided to split the project into two. The first phase was up until 1st May 2015 and the second phase was post the deadline. We worked with figshare to deliver 'what we could' by 1st May (the Arkivum storage solution was in place well before the deadline). This attitude meant that we faced some problems when piloting the repository as outlined above. However, it did mean that we could contact all of Loughborough's EPSRC funded Principal Investigators and inform them that we had a solution in place for them to be compliant with their funder expectations. Perhaps as importantly, having a live product also helped show senior managers that the project was progressing and was delivering results. This ensured that senior managers remained strongly behind the project.

We had a full launch of the data repository on 26th January 2016 (slightly later than initially hoped but it was felt that a post-Christmas vacation launch would have more effect) with the Pro-Vice Chancellor for Research, School Associate Deans for Research, project staff, early academic adopters, and interested Professional Services staff. This launch was the continuation of our advocacy campaign which includes contacting researchers via their school leadership teams as well as directly via training sessions, email communications etc.

Developments around the data repository continue and have been useful in ways we hadn't perhaps realised when the project started. Firstly, a number of researchers have shown interest in the collaborative opportunity offered by the 'Projects' functionality in figshare. This allows researchers or project teams to use space which can then be accessed by users at Loughborough and further afield (so long as they are signed up with figshare). We do not yet know the full potential of this space or, indeed, how researchers may use it in the future, but the potential is exciting. For example, it may be used by PIs to check a research associate's deposit before being sent for publication or for a research group to work collaboratively on a metadata schema for specific outputs from their research, such as images.

An important advantage to having a data repository is that researchers are able to include this in their grant applications. As many funders now require data to be preserved after the end of a project we are able to fulfil this requirement and applicants only need to know that an Institutional offering exists.

Other aspects of the research data infrastructure and service are also invisible to academics. Research data services are embedded in existing tools and programmes which the academic community are used to using. For example, Research Data Management (RDM) training is conducted under the auspices of other programmes including Graduate School training or in consultation with the University Research Challenges.

Data Management Planning is also increasingly embedded into existing solutions and workflows. The University grants management tool has a set of questions which all prospective applicants have to answer. One of these asks whether a data management plan (DMP) has been completed. If the academic answers no to this, they are directed to contact the Research Data Manager for advice and assistance. As such, the Research Data Manager is seen as another source of advice and guidance for the application rather than as a separate entity. In addition, it gives the DMP the same status, on paper, as other elements of the application, such as the Pathways to Impact section.

The relationship between the Research Office and the Library has been further developed by the upgrading of the grants tracking software. Consequently, whenever a new project is funded by an external body (such as the EPSRC) designated Library staff now receive an automated notification along with the relevant researchers, School Dean and ADR. This has helped to merge the Research Data Support Service into existing workflows and procedures.

RDM aspects have also been included in the standard checklist which all researchers need to complete if they are applying for ethics approval. Again, they are notified of the contact details for help regarding RDM in the same manner as for other areas of interest.

By embedding RDM services in existing solutions and programmes we have integrated RDM into existing researcher workflows. There is, of course, potentially additional work which the growth of open data has pushed onto researchers (e.g. cleaning data, writing documentation for an external audience etc.) but with training and advocacy this will hopefully reduce over time and become much more 'business as usual' for the academic community.

Conclusions

The manner in which we have implemented the RDM Service at Loughborough actually helps to highlight its scalability. We have deliberately started with a low key, soft-launch

of the service and with the full launch from January 2016 hope to increase the number of deposits. As such, the service has shown to be useful to a few research groups as well as practical across the whole institution.

The implementation of the service has not been perfect and if we were to do it again there are things we would change. For example, during the development of the data repository there was limited advocacy work around data management requirements. This has meant that we had to advocate around the existence of the data repository and the wider policy environment (i.e. the 'why' as well as the 'how'). However, it was felt that it would be difficult to 'sell' a system which didn't yet exist and that we weren't sure exactly how it would look and fell.

The work is not yet complete and there is still development work needed to further integrate the RDM solutions into the research workflow. Additional work is also required to embed RDM solutions and services in the workflows of other Professional Services staff. Finally, continual advocacy and training is needed, not only for existing researchers but also for new starters.

Once the existence of the data repository is more widely known across the Institution and data deposit is more widespread the advocacy and training message will probably be adapted. Instead of funder requirements and using the data repository, more of the sessions will be focussed on topics such as the importance of metadata, which file formats are best for preservation and when should you include a readme.txt file.

The embedding work will include training more staff to approve deposits and advise on DMPs. This will ensure that there are fewer single points of failure, which is always a danger with a comparatively new service, and will hopefully further reduce the need for a researcher to know exactly who to contact for a particular issue.

Loughborough University is well on the way to integrating RDM packages into existing workflows and systems, being compliant with funder and publisher expectations, and increasing the visibility of Loughborough research. We have tried to complete all this whilst limiting additional workload of academics and where possible by using existing routes of communication and access.

References

Engineering and Physical Sciences Research Council, 2011. Policy Framework on Research Data: Expectations. Retrieved from http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx