Postgraduate Research Data: a New Type of Challenge for Repositories?

Gareth Cole, Jill Evans and Hannah Lloyd-Jones, Open Exeter Project, Library and Research Support, University of Exeter.

The Open Exeter project[1] at the University of Exeter is funded by JISC under the Managing Research Data Programme 2011-2013.[2] The project builds on a 2010-11 pilot, working with Biosciences and Medicine, to build a DSpace repository for the secure, long-term storage of large datasets. This paper will focus on the particular issues involved in incorporating postgraduate researchers' (PGRs) data in an established institutional repository.

Open Exeter is working to extend the pilot data repository service, the Exeter Data Archive (EDA[3]), to all disciplines across the University. Additionally, we intend to merge our primary DSpace research outputs repository, ERIC[4], with EDA so that the published research and the underlying data can be accessed from the same record.

There are clear challenges in establishing an institutional data repository, for example, finding a metadata schema that works on some level for all disciplines of necessity involves some loss of granularity and specialisation. There are also technical challenges around the merging of two repositories which, although both run on a DSpace platform, operate different versions, run two handle subscriptions, and use slightly different (though Dublin Core-based) metadata schemas. In addition, we need to retain the Communities/Collections structure of ERIC within EDA, because it is tied to our CRIS, Symplectic[5].

In order to take account of the diverse 'human' and cultural factors that will impact on our multi-disciplinary approach, we first need to gain a better understanding of research data management (RDM) practice throughout the University: what data is being created, what is being done with/to it, how is it stored, how is it shared, and so on. We are doing this through an online survey - an adapted version of the Data Asset Framework (DAF[6]), through follow-up interviews with researchers and PGRs, and a small number of case studies. Add to this our work with a group of seven PGRs, following their interactions with their data over a 12-month period.

Out of all of these investigations, one finding is beginning to emerge: questions and concerns about what happens to PGR data when students leave the university or move on to other projects.

We are obtaining quantitative and qualitative evidence that is beginning to show that both PGRs and their supervisors are unaware of what steps to take, or what options are available, to make students' data available for future generations. The most consistent comment in data extracted thus far is that since there are no relevant

[1] http://as.exeter.ac.uk/library/resources/openaccess/openexeter/
[2] http://www.jisc.ac.uk/whatwedo/programmes/di_researchmanagement/managingresearchdata.aspx
[3] https://eda.exeter.ac.uk/repository/
[4] https://eric.exeter.ac.uk/repository/
[5] http://as.exeter.ac.uk/it/systems/symplectic/
[6] http://www.dcc.ac.uk/resources/tools-and-applications/data-asset-framework

institutional or departmental policies or guidelines, data is most likely to sit on a hard drive or external drive in an office somewhere until either the device fails or no-one can figure out how to access the files again. For PGRs this is a problem for a number of reasons:

1. Students would like to receive recognition for their work and feel it is being valued and reused to contribute to building knowledge in their academic field. To quote one PGR, 'I feel quite strongly that I have spent literally years acquiring my data, and once I have completed my PhD I should be delighted if someone else wishes to use it and add to it in later years. I don't want it to be all hidden away on my own computer'[7].
2. If the data is more accessible, it will have greater impact and enhance the student's career development. As data citation becomes more standard, PGR students will benefit from the visibility of their thesis data.
3. Typically PGR research data is unavailable for incoming students to build on; they will be aware that the research has taken place but due to the lack of policy on recording and storing PGR data, they (and their supervisors) have no way of locating it.
4. What if research findings are challenged? How can the now early career researcher prove their findings if the data they collected have been destroyed by their previous institution?

The consequences of the above points will be well known to any researcher: there will be a duplication of work which will delay the progression of any new research, and valuable research data may be destroyed or lost. All of this is detrimental to the research community in general but to the PGR student in particular, whose career is becoming more reliant on the ability to gain funding grants and publication of the resultant findings.

Increasingly, PGRs are essential elements in research projects. This is not only in the Sciences where research groups have been long established; Humanities' funding is increasingly being awarded for large scale research projects with at least one, if not more, studentships attached. Consequently, instances of the solitary PGR student, creating data for his or her own work, are disappearing.

PGR students funded by bodies that require Open Access to research data will have to meet these obligations. By the end of May 2012 universities that receive funding from the EPSRC are required to have completed a roadmap outlining how they will comply with EPSRC expectations of research data management. Expectation four states that 'all of their researchers or research students funded by EPSRC will be required to comply with research organisation policies in this area'.[8]

Until this point in the project, we had not considered that there might be a role for Open Exeter in providing stable access to PGR data. However, there is clearly a (relatively) quick win opportunity for us here: we already mandate thesis deposit to ERIC, which, as noted, we will integrate with our data archive; we already allow deposit of supplementary files, such as video and audio when they're an integral part

[7] Taken from Open Exeter RDM online survey results, as yet unpublished.
[8] http://www.epsrc.ac.uk/about/standards/researchdata/Pages/expectations.aspx

of the thesis. It is only a comparatively small next step then to permit (or even mandate?) deposit of underlying data.

However, it is not merely a question of depositing PGRs' research data in our merged repositories; there are a number of issues surrounding the decision to do so that must be considered.

Firstly, we will need to review the current system by which PGRs deposit their research outputs. Currently PGRs themselves upload their thesis to the ERIC workflow where it is managed and checked by the Postgraduate Administration Office. Whereas deposit of a single PDF of a finished thesis is a relatively straightforward task to manage, ensuring standard practice in the deposit of data is more complicated. For example:

- Appropriate, preferably open, data formats or alternatives should be used where at all possible.
- Enough quality metadata should be attached to describe the content and nature of the data – this is particularly important in cases where data is zipped.

One of the aims of the Open Exeter project is to embed good practice in RDM in existing training programmes for PGR students, such as the Researcher Development Programme[9], the doctoral supervision course[10] and PGR handbooks. Depending on the content and coverage of such training, PGRs could be "skilled up" in order to be responsible for the curation, description and deposit of their research data. This training could help to embed the concept of data deposit in the PGR research lifecycle and also make them aware of funder requirements, which would be useful for them as early career researchers applying for funding.

However, the University would inevitably need to provide a support service for PGRs, checking the data deposited for correct use of metadata and compliance with copyright, as well as for IPR and confidentiality issues. Other questions that would have to be addressed by the University would be:

- Who decides, imposes and maintains criteria for metadata submission?
- Impact on the network if large datasets are uploaded by multiple users at particular points in the academic year.

This leads us onto the topic of who has the right skills to assess the integrity, accessibility and reusability of data deposits and metadata and to support the upload of large datasets. The skills of support staff (probably repository curators and subject librarians) would have to be developed in order to assist the PGRs in the upload of their research data. Providing this support would impact on the workloads of these support staff as well as on their job description and role within the University. There are questions around who would be an appropriate first point of contact for PGR queries associated with data upload.

[9] http://as.exeter.ac.uk/support/development/researchstudents/erdp/
[10] http://as.exeter.ac.uk/support/admin/staff/supervisionofresearchstudents/

Finally, there are the technical and financial aspects of encouraging or mandating the upload of PGRs' research data to an institutional repository. The question of whether there is a need to impose a limit on the amount of data that can be deposited (in terms of file size) would need to be agreed. The cost involved in the storing and curating of the data would have to be covered somehow. In some cases, the cost could potentially be covered by a funding body, but this would depend on how the student is funded. Addressing these complex questions brings in experts from outside the Library: from Research and Knowledge Transfer, IT, Graduate Employability, and the Legal Office.

Despite the difficulty of tackling the issues discussed, the benefits of allowing deposit of PGR data alongside theses are obvious:

- Thesis and associated data will be accessible from a single point of access.
- The thesis/data record can be linked to PGR publications resulting from. research, and to publications and data originating in the PGR's research group.
- Secure, long-term storage of PGR data facilitates sharing and reuse, use of data for teaching research methodologies and analysis, taking forward of research ideas and outputs by future students.
- Greater visibility and impact for students about to embark on an academic career.

During the lifespan of the Open Exeter project we hope to be able to address some of the questions raised in this paper by working with PGR students, developing and piloting training for them on good practice in RDM and how to prepare their data for deposit and reuse. At the same time, we hope to raise awareness of the benefits of providing Open Access to research data for early career researchers as well as setting up a support service in the University comprising staff who will have the skills needed to advise and assist with data deposit to our repository.