Unlocking Thesis Data

University of Leicester case study

Rachael Kotarski

http://orcid.org/0000-0001-6843-7960

Sara Gould

<http://orcid.org/0000-0003-2763-9755>

July 2015

DOI: [10.15123/PUB.4306](http://dx.doi.org/10.15123/PUB.4306)

 © The authors 2015. Licenced under Creative Commons Attribution 3.0 Unported License.

Contents

Overview of Leicester 3

Workflows at Leicester 4

Data and supplementary files 6

Use of identifiers 7

Moving forward 8

Recommendations 9

Workflow overlaid with suggested PID creation 10

Overview of Leicester

1: Interviews

The following people were interviewed as part of the case study:

1. Grant Denkinson, Leicester Research Archive Manager
2. Ian Rowlands, Research Services Manager & Bibliometrician
3. Tom Moore, Leicester Research Archive
4. Konstantinos Georgiadis, Leicester Research Archive
5. Rob Pearce, Library Systems Manager.

2: Summary data

|  |  |
| --- | --- |
| **Summary Table** |  |
| Institution | University of Leicester |
| Higher Degrees Awarded in 2012/13 | 265\* |
| Publications repository | Leicester Research Archive ( DSpace V4) |
| Publication repository IDs  | Handles |
| Data repository | None |
| Data repository IDs | None |
| Theses held in… | Hard-copy and electronic |

\* From HESA Table 18a, Higher Degree (Research) Qualifiers by Institution and Subject of Study

3: Summary profile

The University of Leicester requires deposit of both print and electronic theses. It has had a thesis deposit mandate since 2008 which requires students to agree to a copy of their thesis being held in the institutional repository, Leicester Research Archive. The deposit process is fully mediated, meaning all uploading and metadata creation is undertaken by the relevant library team rather than by the student or Graduate School.

There is currently no data repository but a new role of Data Repository Manager will soon be filled and will take responsibility for developing both a data repository solution, and data management workflows.

Workflows at Leicester

The library first becomes aware of the theses it expects to receive whenever the student gets in touch with an enquiry, or the Graduate School updates its records in the student SITS system. Either of these things can happen very early on in the thesis completion period, and it may be many months before the final thesis is received in both print and electronic formats in the Library.

The Library’s workflows for thesis submission begin when the student has completed and passed the examination and viva, and has been informed by the Graduate School that they have successfully completed their PhD. Students are considered to be fully completed only when they have deposited with the library a printed hard bound copy of the thesis, an electronic copy and the signed Thesis Deposit form. All three of these are delivered to the Library by the student themselves and not by the graduate school. Typically, the printed thesis may arrive direct from the print shop, and the e-thesis sent by the student by email or a Dropbox. A PDF or PDF/A format is requested and the vast majority of theses are submitted in this format. Supplementary files are also sometimes submitted. A small number of PhDs by publication are undertaken, and these might arrive in the library as a set of separate files representing the individual publications making up the final thesis.

Consent forms and embargoes

The consent form includes a section on embargoes, which should have been agreed and signed off by the graduate school before the consent form reaches the library. Leicester has recently tightened its approach to embargo and strong justification is now required before embargoes are given. Around 20 theses a year receive a 'full embargo' meaning both electronic and print copies are restricted and no record is visible in either the repository or library catalogue. These invariably relate to research undertaken with commercial sponsorship and/or where patent applications may be planned. For theses where a monograph publication is planned based on the thesis, full or extensive embargoes are only agreed where the planned monograph publisher is named and written evidence provided that publication of the thesis would jeopardise commercial publication. Otherwise a requested embargo is restricted to 12 months embargo on the electronic copy only.

Additionally, the consent form has an option for the student to indicate their wish for a copy of the thesis to be added to Proquest's "ProQuest Dissertations & Theses" subscription database. This provision is part of a project in which ProQuest digitises older theses at reduced rates or gratis in exchange for permission to add the theses to its products and promote its services to current students. If the student wishes ProQuest to hold a copy, permission is passed to ProQuest to harvest the thesis direct from Leicester Research Archive.

Thinking about how the library knows what theses to expect (how do they know they hold all PhDs awarded?), the Leicester Research Archive (LRA) Administrator has access to the SITS Graduate School student database in which all student study is recorded. The LRA Administrator accesses SITS and notes which doctoral awards have been newly confirmed as awarded, and copy some of the details such as form of name, email address (often the university one and a personal one), department, supervisor, name of award and date.

Tracking deposits

These details are held in a spreadsheet in the library which is then used to track delivery and receipt of the print thesis, the electronic thesis and the deposit form. When all three items have been submitted and received by the Library, the electronic copy is uploaded to LRA. The author is emailed to let them know their thesis has become available on LRA. The Handle identifier is included in the email so the author can use this in any citations or links from that point. The personal email address is often used as the student may have left the university by this point.

The LRA Administrator is the person who creates the metadata and inputs it into LRA. The information is typed or copied from the thesis itself. Keywords are included only if provided by the student already, and sponsor information is added if the Administrator notices it in the general acknowledgements. The library is currently examining the use of RIOXX, but not for theses yet.

The hard copy thesis is stamped and an RFID added, then catalogued. The catalogue record includes a link back to LRA so users will always be routed to the electronic copy in preference to print. EThOS-digitised theses are processed in the same way. ProQuest-digitised theses are initially selected for digitisation from catalogue records and a spreadsheet of metadata created; the theses are scanned by the external supplier, and OCR added by ProQuest. On completion, the LRA records are created in bulk from the spreadsheet, the PDFs bulk-uploaded, and abstracts provided by ProQuest also added at this point unless the author opts out.

Data and supplementary files

So far, only occasional additional files have been received by the library as part of the thesis deposit. Files received include computer code, video of gene sequences, image files. The files are generally uploaded into LRA as bit streams; no preservation formats are considered, and the files are normally zipped up and named appropriately. The additional files can be seen in the main thesis record, described in the 'file type' descriptor. There have been no requests so far from students to provide access to data or files held elsewhere such as in an external data repository. No checks are made to see if the files deposited are identical to those deposited in print, or that the deposited files formed part of the examination submission.

Thinking about a future data repository, it is likely a principle will be established that large data sets are stored in external repositories where possible rather than the library, including established subject repositories and academic department storage. The library team is aware the university has large data sets such as that held in the Leicester Astrophysics Data Service, but its IT services manage this provision and the library has not been involved in their management to date.

A plan is now underway to set up a research data repository and service at Leicester. The proposal is that the Library will lead the development of the service to be established as part of the Academic Liaison team. The proposed service would include:

* Data management planning: advice on existing data sources for re-use; advice on technical solutions to support different research methodologies.
* Support for data management planning: storage & software available; legal and regulatory compliance; licensing of data for re-use; standards, including file formats & metadata; options for long term storage and access to selected data at end of project. Support for budgeting to include cost of data resources/facilities not covered by institutional overheads.
* Managing active data: practical support and training at an early stage to establish consistent approach to file naming, storage, metadata, version control, security, legal & regulatory compliance.
* Dissemination and end of project: support for data selection for long term storage and access (data and/or metadata only); transfer of data either to a national data centre or university repository as appropriate; data registration; anonymisation of data as necessary; allocation of DOIs; confirmation of access rights/policy; use of preservation file formats.
* After the project: Data repositories & Data catalogues: long term access to the data; liaison with Information Assurance on access to restricted data as necessary.

Use of identifiers

Since Handles are used well in LRA, the question is asked if another identifier - DOIs - is also needed. Initial views are that DOIs bring advantages in their established position and effective promotion; they look more official and less "ugly" than Handles! On the other hand use of DOIs requires financial commitment. The decision at Leicester is likely to be made as part of the Research Data Management development plans, alongside decisions on DOIs for data.

ORCiD is not yet part of Leicester's student records. Potential drivers for uptake include Wellcome and NIHR funder requirements for researcher ORCiDs. The ideal scenario is that there will be one place of record for all student data including ORCiD, and at this stage this is likely to be the SITS, the student database. The Current Research Information System (CRIS) will be able to hold ORCiDs and the DSpace LRA repository would easily be able to have an extra field created to hold the ORCiD identifier.

Moving forward

The institution also has a CRIS – a Symplectic Elements implementation called IRIS - managed by the Research Support Office. Academic staff have direct access to IRIS, and input their own research papers into the system. There are currently no plans to open up its use to PhD students, but this would provide opportunities - and challenges - to migrate thesis deposit to the student, many of whom already use IRIS to input other types of research papers.

The university has a number of institution-published journals. There are plans to start to use OJS, the journals platform, which would bring an opportunity to assign DOIs to the journals using CrossRef services associated with OJS. The CrossRef service might also then become an option for DOIs for theses, but it is too soon to think in detail about the various DOI options.

The university is also considering introducing a new requirement to submit PhD theses through TurnItIn. This would mean theses would arrive in the library not direct from the student but via the VLE, requiring a revision of the upload and metadata creation workflows.

Recommendations

1. Review the content of the email that goes to the student when the thesis is added to LRA. Does it encourage the author to use the Handle identifier and explain why this is a good thing? Would this be an opportunity to suggest the author register for an ORCiD?
2. Since deposit into LRA is a mediated process, it was suggested that the LRA Admin team could, with appropriate permission, add a DOI to the thesis itself at the same time as including it in the metadata. This would require opening the PDF of the final thesis and pasting the DOI into an appropriate place in the title page or front matter. Guidance on formatting the title page of the thesis could be updated to allow for a placeholder into which the DOI would be added. This would have the additional benefit of alerting the student to the fact that the final electronic version of their thesis would have this additional identifier included just before making it available. We suggest UTD looks into this idea in more detail.
3. UTD should ensure we examine the question of DOIs for “fully embargoed” theses where both print and electronic copies are restricted, and no record is publically visible in either the catalogue or repository. Leicester currently processes up to 20 such theses a year.
4. UTD should consider how the use of DOIs might need to take account of a situation where a second copy of the thesis is ingested (with permission) from the Institutional Repository by a commercial third party, such as ProQuest.

Workflow overlaid with suggested PID creation

