
Unlocking Thesis Data

LSE case study

Stephen Grace

<http://orcid.org/0000-0001-8874-2761>

Sara Gould

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

July 2015

DOI: [10.15123/PUB.4303](https://doi.org/10.15123/PUB.4303)



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



Overview of LSE

1: Interviews

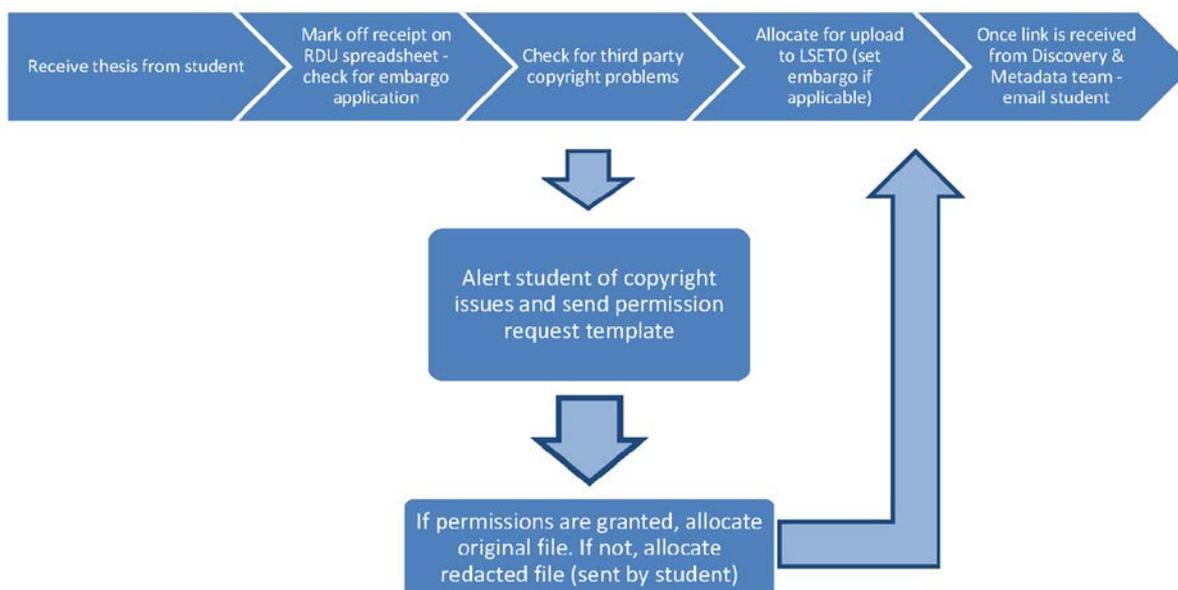
The following people from the Library were interviewed as part of the case study: Dimity Flanagan, Laurence Horton, Alexandra Kohn.

2: Summary data

Summary Table	
Institution	London School of Economics
Higher Degrees Awarded in 2012/13*	200
Publications repository	LSE Theses Online
Publication repository IDs	URLs
Data repository	None currently, may use LSE Research Online
Data repository IDs	[URLs in LSE Research Online]
Theses held in...	Electronic

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

3: Current workflow



LSE case study

The London School of Economics and Political Science (LSE) was founded in 1896 for the “betterment of society”, and is one of the world’s leading social science universities. Its library, the British Library of Political and Economic Science has Designated status from the MAL as being of outstanding national and international importance.

1: The student arrives

Students are registered on MPhil/PhD or MRes/PhD programmes, and a formal upgrade process at the end of the first year allows progression to the PhD. Full time students have a maximum of four years to complete their research, and part timers eight years, unless special permission has been given by the Research Degrees Subcommittee.

2: The thesis is submitted

Arrangements for examining theses are overseen by the Research Degrees Unit (RDU). Students and their supervisors complete an Examination Entry Form when close to submitting their research to examination; this covers nominating examiners for approval by the relevant subject panel. Arrangements for the viva are made and two soft-bound copies of the thesis are submitted via RDU for the examiners (who may also request an electronic version)

After the examination the RDU will confirm the examiners’ decision. This may accept the thesis as is, require minor or major amendments and resubmission, award an MPhil degree or make no award. In addition, it is possible to request another viva without revision of the thesis.

3: The final thesis is processed

When the final version of the thesis is accepted according to the examination outcome, an electronic copy has to be submitted by the student to the library (which converts where needed to PDF format). Students can either remove third-party copyright material (with a standard form of words indicating removed content) or provide instead a redacted version suitable for public access. Library staff on receipt of the file will check for copyright issues (clarifying these with the student if necessary). Students can request an embargo of up to twelve months; especially in area studies, some content is seen as particularly sensitive and may be removed to safeguard the identity of participants. No copy of the original, unredacted thesis is retained.

The thesis is then passed to the discovery and metadata team for uploading and publishing on LSE Theses Online (LSETO). A spreadsheet of relevant information is regularly shared between the RDU and library covering student surname, initials, department, degree and any embargo date, and is updated by the library with the date of receipt for the thesis. Metadata staff also assign LC subject heading but not keywords.

Very few supplementary files (data, software, AV material etc.) are received with theses: only four between November 2014 and May 2015, for instance, one of which contained proprietary data not available for public sharing. These objects are added to the same record as extra files. There is no requirement to submit data as part of the examined work, or on completion of studies.

4: Tentative steps towards data sharing

LSE is in the process of adopting an institutional policy on research data management (RDM), as a concrete step in its move towards supporting its researchers in this area. The policy will apply to staff and research students' data. Currently, the Digital Curation Centre (DCC) is helping the School scope its options for repository support options, including a home for research data requiring long term curation. The library has appointed a Data Librarian with responsibility for Research Data Management Librarian to develop its remit in this area. An RDM support service has been established and is in the process of promoting itself within the school.

Given this context it is unsurprising that students are not advised in the research degree regulations or other guidance to present their data for sharing. At present, if they ask, students would be told about Zenodo, Figshare and subject data repositories as suitable homes for data – the same message offered by the library to academic staff. There are a small number of 'empty' records in LSETO which have no content but a link to data held elsewhere such as Github. It was felt, though, that research students would be a good audience to teach good data management practice, that they could become early adopters of LSE RDM infrastructure and thus that curating thesis data would be a useful focus of activity for the School. In particular, the library is looking to include RDM support messages and training in the School's PhD Academy to be established in 2015/16.

The library would need to develop procedures for handling data, criteria for data selection, checks for anonymisation etc. Even if research degree regulations were changed to encourage data availability, new structures would be needed to respond to students' activity. The School more generally would also need an Intellectual Property Rights policy (a draft is under consideration) to clarify who had rights in data created by research students under supervision at LSE.

5. Reflections on the process

LSE is at an early stage in identifying data as institutional assets to be managed and shared (where appropriate). Thesis data provide a good use case for showing early progress in RDM: provided the IPR issues are clarified, it is clear whose data are under consideration, they are clearly related to a research activity and by virtue of the PhD enterprise in advancing knowledge are worthy of finding a wider audience for potential citation and reuse.

The School's process for publishing theses is efficient, with the library solely responsible for record creation. Since students have to supply their electronic thesis before graduation, there is a strong likelihood of receiving the final thesis before the student departs. There is, though, a distinctive and time-consuming emphasis on checking theses for third-party copyright materials – a process that in other universities is the responsibility of the student. In the absence of a CRIS or other centralised system it is hard to see where other efficiencies could be gained by inheriting existing metadata for use in LSETO.

Applying Persistent Identifiers

LSE could adopt PIDs for students and their theses in the following manner.

ORCID

The School is following Jisc activity in relation to ORCID identifiers before making any decision about using it for staff. There was felt to be scope for assigning ORCIDs to research students, though, even in advance of a School-wide adoption. An appropriate point would be either on registration or on upgrade to the PhD programme, in both cases within the first twelve months, provided the information could be passed from departments/RDU to the library for creating ORCID accounts (which require students to agree to the account).

DOI

The library is actively investigating DataCite membership as part of its RDM scoping work with the DCC. Theses (and their data) would be an early use case for minting DataCite DOIs, though thought would need to be given to DOI suffix and granularity. Presuming LSE became a DataCite client, it could assign DOIs in the following ways.

At present, a DOI could be assigned to the thesis on receipt of the electronic copy post-award as part of the work of adding it to LSETO. Data objects could form part of the same record, either using the same DOI or with their own DOIs but resolving to the same landing page. In general, the library would prefer one record with one landing page for the thesis "collection". If LSE had a separate data repository, it would be open to hosting data objects here provided the link between them and the thesis was clear and explicit. As the nature of LSE's archiving and sharing infrastructure is subject to change in light of the RDM work, it would not be appropriate to make definitive recommendations for minting DOIs: using theses and their data as a specific use case would help to develop the nascent RDM infrastructure with a clear but manageable pipeline of content.

Recommendations

1. Adopt ORCID identifiers as outlined above.
2. Consider DOI minting in the light of developing RDM infrastructure and policy-making, as a specific early exemplar for the School.
3. Consider whether metadata could be inherited by the repositories from other current or planned systems or extracted from them to improve efficiency and reduce delays.
4. Provide further guidance welcoming submission of data/software files with theses, together with examples to stimulate consideration by students and their supervisors.
5. Clarify ownership of student data.

Appendix: Suggested workflow with PID creation

