Mapping the UK thesis landscape

Phase 1 project report for Unlocking Thesis Data

Stephen Grace, Michael Whitton, Sara Gould and Rachael Kotarski

DOI: 10.15123/PUB.4307

Contents

Publication details:

Title	Mapping the UK thesis landscape: Phase 1 project report for Unlocking Thesis Data
Authors	Stephen Grace, University of East London (**http://orcid.org/0000-0001-8874-2671**) Michael Whitton, University of Southampton (http://orcid.org/0000-0001-6838-1100) Sara Gould, The British Library (http://orcid.org/0000-0003-2763-9755) Rachael Kotarski, The British Library (http://orcid.org/0000-0001-6843-7960)
Date	15 July 2015
URL	http://dx.doi.org/10.15123/PUB.43077
Version	1.1

Executive Summary

Unlocking Thesis Data (UTD) is a community-driven project to promote the use of persistent identifiers for theses, their underlying data and their authors. It is a collaboration between the Universities of East London and Southampton and The British Library. UTD is a Research Data Spring project funded by Jisc as part of its Research at Risk programme.

By their very nature, PhD theses break new ground and advance scholarly knowledge. Most make use of newly-created data but these data can be trapped in an appendix or DVD – either unavailable or not suited for reuse. UTD will make data more discoverable and citeable, thereby offering incentives to students to share their data in more appropriate formats, in the context of a sustainable national thesis framework.

This report details the work of the first phase (April-July 2015), where the project carried out a survey of EThOS institutions, interviewed staff at six universities for more in-depth case studies, and synthesised the findings. Overall, there is much appetite for applying DOIs to theses and their data (which includes datasets, software components and other non-textual supplementary files) and ORCiDs to research students. Glasgow, Southampton and East London universities each minted a DOI for an existing thesis, demonstrating the viability of our intent, but the case studies showed there are constraints in both processes and technologies to be addressed before persistent identifiers (PID) for theses can be a nationwide reality in the UK.

The project makes five recommendations for further work:

1. **Hold at least three thesis "clinics" to investigate opportunities and barriers to assigning DOI and ORCiD identifiers in UK universities**
2. **Engage with system suppliers/vendors to identify opportunities for enhancing software with required PIDs**
3. **Consult with EThOS formally to understand what needs to change in EThOS systems and processes to harvest and display PIDs and related metadata for theses and their data**
4. **Evaluate approaches to updating UKETD profile, initially in EPrints, before planning software enhancements**
5. **Investigate requirements and solutions for those institutions that use EThOS as their first-point repository**

Note that these five recommendations cover the second phase of Unlocking Thesis Data. In a third phase we would expect to finalise the metadata requirements, develop software enhancements in specific repository/CRIS products and create a comprehensive toolkit and guidance for institutions implementing PIDs for theses and their data.

Further details are at **unlockingthesisdata.wordpress.com** and the various project outputs are listed at **http://dx.doi.org/10.15123/PROJECT.15**.

Background to the project

The project arose from community interest in the potential of applying persistent identifiers (PID) to theses and their associated data:

1. Initial conversations
2. Community workshops
3. Research Data Spring

1: Initial conversations

Stephen Grace (University of East London) and Wendy White (University of Southampton) were each thinking about applying DOIs to theses, and talked about this approach in the margins of the **Research Data Management Forum** on 20 June 2014. They agreed that it would be best if this was progressed at the national level, and so spoke to British Library staff responsible for EThOS (the national theses service) and for DataCite in the UK. There was great interest, and so a couple of workshops were arranged to explore persistent identifiers.

2: Community workshops

EThOS organised an "exploratory workshop" **A national approach to persistent identifiers for theses**, held at the British Library on 28 November 2014. This was an initial exploratory workshop which kicked around the many reasons why "DOIs for Theses" make sense – and the challenges to reaching a cohesive system across all UK theses. On the one hand the task seemed easy: many institutions already assign DOIs for other material such as journal articles or datasets, so theses might fall easily into existing DOI workflows. On the other, the very nature of PhD theses brings specific challenges: not formally published, held locally within a repository but with possible variant (redacted) versions in EThOS, PQDT, FigShare and elsewhere; wide ranging submission requirements between institutions, or between departments within an institution; print or e-born, or a bit of both; and the involvement of many institutional stakeholders like academic committees. And, the benefits of DOIs for theses are not so obvious outside the library or repository environment. Participants agreed early on that by 'identifiers' they meant DOIs, and that DataCite is likely to be the most appropriate (but not the only) service provider. DataCite can certainly be used for content types other than 'data', including theses. There was a strong desire to agree a national approach. More so than with other research outputs, there is a shared sense of the need to manage the UK's PhD theses responsibly and as a long-term commitment, at a national level. While each institution would probably mint its own thesis DOIs, it would be great to be able to assign DOIs consistently, and follow agreed guidelines and workflows across all institutions.

A second workshop **DataCite, DOIs and Theses** was organised by DataCite UK and held at the British Library on 16 January 2015. This continued the discussion with wider representation from across the UK HEI sector. Stephen Grace presented an outline of how UEL would approach assigning DOIs for doctoral theses, and a lively discussion afterwards considered practical issues for UEL and for other institutions. There was overwhelming support for working together on a nationwide system for thesis DOIs across the two workshops, with at least 60 delegates across the two events. Around this time, Jisc announced its **Research Data Spring** (part of the

Research at Risk programme) and it was agreed to develop a proposal led by East London and Southampton universities and the British Library.

3: Research Data Spring

The project team submitted a proposal "Unlocking the UK's thesis data through persistent identifiers" to Jisc, successfully reaching the evaluation stage at a sandpit workshop held at Aston University on 26-27 February 2015. A dry-run of the workshop, where ideas were pitched, comments were offered and informal votes were cast, was held at the **International Digital Curation Conference** in London on 11 February 2015. At the Aston workshop, the proposal was refined by Stephen Grace (UEL), Dorothy Byatt (Southampton), Sara Gould and Rachael Kotarski (both British Library). Soon after the workshop, Jisc offered a phase one grant to the **Unlocking Thesis Data** team (the name shortened for ease of use and abbreviated to UTD) led by UEL. During April-July 2015, the team organised and analysed an online survey of existing thesis practice, conducted six case studies, and here report on how to take the proposal forward in further funded phases.

Survey

The first task for the project was to understand the current situation as it relates to thesis identifiers, thesis authors, uptake of DOI identifiers in HE institutions for other research outputs, and the management of supplementary data files and raw data produced in the course of the PhD.

The survey aimed to:

- Gather quantitative data on the level of DOI and ORCiD usage so far for outputs other than theses
- Understand typical workflows for the handling of theses in institutions, from student deposit to repository access and citation
- Begin to note 'non-typical' scenarios for which solutions may be needed
- Gauge the level of interest in assignment of unique identifiers to institutions' theses and the UTD project
- Begin to identify how assignment of DOIs might fit into current workflows, and spot potential pinch points
- Gather volunteer institutions for the next UTD Phase 1 task, institutional case studies.

The survey was sent to a named contact at 138 HE institutions which offer research degrees. This list was drawn from the contact list held by the British Library for the EThOS E-Theses Service. The survey was launched on 27 April 2015 and ran for three weeks.

Key findings

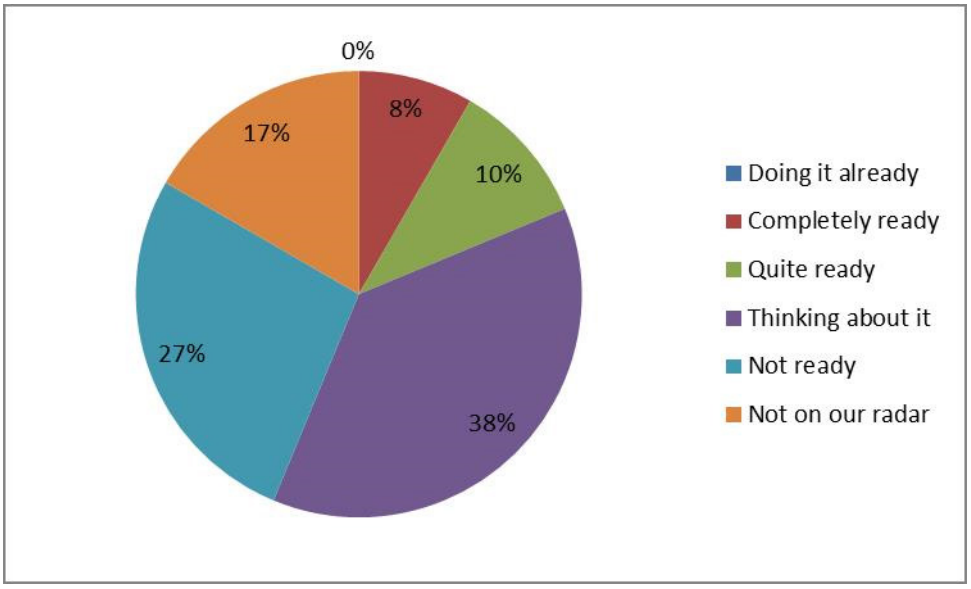
- 49 institutions (35.5%) responded to the survey, indicating that *Unlocking Thesis Data* is of interest to a significant proportion of HE institutions. A list of respondents is provided in Appendix A.

- At the time of the survey, no institution assigned DOI identifiers for their theses, although DataCite DOIs were used by 33% of institutions for their research data.

- Around 59% of institutions require students to submit both print and e-copies of their final thesis, and this often results in double-handling, for example in creating separate records for the catalogue and repository. This may have implications for UTD.

- The most 'typical' scenario is an institution which uses an EPrints repository for its e-theses and supporting files, students must submit both print and electronic versions of their thesis, the thesis is uploaded and the metadata created by the repository staff (though students are a close second), and the institution assigns DOIs for its datasets but not theses. This suggests such a scenario might form the first case study or the core focus for UTD.

- In response to the question *How ready are you to begin assigning DOI identifiers to your theses?* institutions varied from *'Completely ready'* to *'DOIs are not on our radar at all'*:



Q16: How ready are you to begin assigning DOI identifiers to your theses?

The intention is to ask this question again at the end of the full UTD project; this will form a key indicator as to the success of the project.

- Twenty-four respondents (49%) volunteered their institution to be a case study for UTD. The aim is to deliver just six case studies under Phase 1, and we hope those institutions not selected will be willing to host UTD clinics on phase two, become early adopters or have other opportunities to be closely involved.

http://dx.doi.org/10.15123/PUB.4307

Case Studies

Several institutions had indicated their interest in the project at the two British Library workshops, in supporting the Jisc application or at the two Jisc events: as well as UEL and Southampton, the universities of St Andrews, LSE, University of the Arts London (UAL), Northampton and Bristol had signed up by the Aston workshop, and in all 24 institutions were willing to be case studies. UTD chose six institutions based on a combination of technical platforms, readiness for assigning DOIs, institutional profiles and to some extent proximity to London or Southampton. In chronological order based on interview dates they were:

- University of East London
- University of Southampton
- London School of Economics and Political Science (LSE)
- University of the Arts London (UAL)
- University of Bristol
- University of Leicester

The case studies found a range of institutional practices and software. We summarise each institution before drawing out some more general points. The numbers of higher degree awards 2012/13 comes from HESA Table 18a, Higher Degree (Research) Qualifiers by Institution and Subject of Study.

1: University of East London

Institution	University of East London
Higher Degrees Awarded in 2012/13	55
Publications repository software	EPrints (ROAR)
Publication repository IDs	Handles
Data repository software	EPrints (data.uel)
Data repository IDs	DataCite DOIs
When to assign DOI	DOI could be assigned at registration using a naming convention that involves a unique number relating to the student. This could be embedded in the PDF of the thesis added to ROAR.
When to assign ORCiD	Request student establishes account following registration
Case study available at	http://dx.doi.org/10.15123/PUB.4301

UEL is able to mint DOIs by virtue of its DataCite membership using the CoinDOI plugin. It could assign a DOI in advance by using a standard notation; this would not require a placeholder record to be created on the repository, nor a spreadsheet of reserved DOIs. It would, though, have the advantage of letting students know it in advance so that they could embed the DOI in their thesis – ideally, prominently on the title page. This notification could happen on registration (or possibly on transfer from MPhil to PhD). While DOIs could be assigned in advance, they are only created ("minted") when metadata is sent to DataCite and the DOI is returned to the agent making the request.

The same event could trigger a request for the student to create an ORCiD account, and notify the Graduate School and Library of the ID. UEL does not have an institutional ORCiD membership, but is following UK-wide developments. In the meantime, the first batch of eight research students have been contacted to create ORCiD accounts following successful registration.

2: University of Southampton

Institution	University of Southampton
Higher Degrees Awarded in 2012/13	600
Publications repository software	EPrints (ePrints Soton)
Publication repository IDs	EPrints IDs / URIs
Data repository software	EPrints (ePrints Soton)
Data repository IDs	DataCite DOIs
When to assign DOI	On publication of the thesis in ePrints Soton
When to assign ORCiD	On registration using Southampton ORCiD Service
Case study available at	http://dx.doi.org/10.15123/PUB.4302

Southampton, like UEL, already mints DOIs for datasets but on the same repository used for theses and other publications (ePrints Soton). It uses the DataCite Metadata Store (MDS) directly rather than the CoinDOI plugin. Students are given a choice to include the data with their thesis in the same record or to have separate records for datasets with their own DOI. They need supervisor approval to deposit data separately.

There are potentials to improve efficiency by using Pure for handling theses, though it does not at present support creation of DOIs. Southampton like several other universities requires both print and electronic versions of doctoral theses, creating metadata and curating both. With e-theses now well established as a normal or even default form of publication it might be time to revisit this requirement.

Southampton has also established an embedded ORCiD service to simplify the creation of accounts and return the identifier into the Pure-based CRIS.

3: LSE

Institution	London School of Economics and Political Science
Higher Degrees Awarded in 2012/13	200
Publications repository software	EPrints (LSE Theses Online, LSETO)
Publication repository IDs	URLs
Data repository software	None currently, may use LSE Research Online
Data repository IDs	[URLs in LSE Research Online]
When to assign DOI	On publication of the thesis in LSETO
When to assign ORCiD	On registration or upgrade
Case study available at	http://dx.doi.org/10.15123/PUB.4303

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. Theses could be an early area for exemplifying RDM good practice, the Library working with the new PhD Academy to prepare students for the new norms of managing and appropriately sharing data.

Provided it had the means to create DOIs, the Library could assign a DOI to a thesis when it arrives post-examination. Data objects could have their own DOIs but form part of the same record with all DOIs resolving to the same landing page for the thesis "collection". If it had a separate data repository, any data objects that were part of a doctoral thesis would need to have clear links to the text of the thesis in LSETO.

ORCiDs could be assigned on registration or upgrade in advance of any School-wide approach; like many institutions LSE is watching national developments with interest.

4: University of the Arts London

Institution	University of the Arts London
Higher Degrees Awarded in 2012/13	15
Publications repository software	EPrints (UAL Research Online)
Publication repository IDs	URLs
Data repository software	Small internal repository not currently open online nor linked to the publications repository.

Data repository IDs	N/A
When to assign DOI	On publication of the thesis in UALRO
When to assign ORCiD	On publication of the thesis in UALRO
Case study available at	http://dx.doi.org/10.15123/PUB.4304

UAL has some distinctive features by virtue of its nature as an arts and design university. Indeed, "data" might not be universally welcomed as a term at UAL to describe the range of performance-based and documentation objects encompassed in student theses.

Student guidelines require that a record of any live performance or exhibition must be included as part of the document submitted for examination, and is therefore also included in the items submitted to the library. These accompany the printed version (typically as a DVD in the rear of the volume) and also as additional electronic files, so best efforts are made to ensure the full body of works making up the thesis are gathered and recorded.

Given that multi-part works are the norm rather than the exception for theses, UAL could provide interesting guidance to the sector with regard to assigning DOIs. It does not yet have this guidance, nor guidance to students on what might be offered as supplementary data.

UAL is looking to the Unlocking Thesis Data project and other initiatives to advise on good practice with regard to ORCiDs. It could adopt a similar approach to other case studies in requesting students create an ORCiD account and share their details with the university, for embedding in systems and details of the final thesis. More likely, it would request ORCiDs at the point the e-thesis is received for cataloguing.

5: University of Bristol

Institution	University of Bristol
Higher Degrees Awarded in 2012/13	590
Publications repository software	Pure
Publication repository IDs	Handles
Data repository software	CKAN (data.bris)
Data repository IDs	DataCite DOIs
When to assign DOI	On publishing the thesis on Aleph
When to assign ORCiD	On registration through Pure
Case study available at	http://dx.doi.org/10.15123/PUB.4305

Bristol currently has a print-based approach to theses; e-theses are only used as an aid to detect plagiarism using TurnItIn and are discarded when this process is finished. It does have a data repository based on CKAN, so could offer a home to thesis data. Requests for online access to University of Bristol theses are directed to EThOS in the first instance and then passed on to the Library; where agreement has been given by the author the hard copy will be sent to EThOS for digitisation and then made available online via EThOS.

Aleph does not assign persistent identifiers and Pure uses Handles, so Bristol would require system enhancement to support creation of DOIs for theses. It already supports ORCiD, and Bristol has been an institutional member since March 2015. The University sees potential in being able to register or capture ORCiDs as part of student admissions and staff recruitment processes, and will be exploring this over the coming months.

At the same time, it is just beginning an internal change process to move to full electronic submission and handling of theses and the thesis examination process. A project has been proposed and a business case is being developed so that resources can be allocated, with an aim of having electronic submission in place for the 2016/17 academic year. There is an appetite from the Examinations and Library interviewees to create a streamlined end-to-end process for handling e-theses, and further engagement with UTD could help in their planning.

6: University of Leicester

Institution	University of Leicester
Higher Degrees Awarded in 2012/13	265
Publications repository software	DSpace V4
Publication repository IDs	Handles (LRA)
Data repository software	None
Data repository IDs	N/A
When to assign DOI	On adding thesis to LRA
When to assign ORCiD	
Case study available at	http://dx.doi.org/10.15123/PUB.4306

A plan is underway to set up a research data repository and service at Leicester, with the Library leading the development of the service. This will augment Leicester Research Archive (LRA), a DSpace publications repository. Theses (both print and electronic) are received by the library from the student post-award, and there are no plans as yet for students to interact with the new CRIS system (Symplectic Elements). At the same time, Leicester is considering use of TurnItIn, which will mean

theses arrive via the VLE rather than the student. Because the infrastructure is in this state of flux, the project recommendations for assigning PIDs are accordingly tentative.

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 in phase two since it could have applicability elsewhere.

The project should also consider 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. We 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.

Leicester could create ORCiDs on student registration, and adapt LRA to accommodate the identifiers. It is likely that SITS (the student records database) will become the locus of all student data so the university will have to consider the interaction between systems that hold relevant data and systems/services that can generate persistent identifiers. An early consideration would be the justification for changing from Handles to DOIs.

Applying Persistent Identifiers

We review here recent steps in applying PIDs to theses and their subsequent harvesting by EThOS, and consider how these identifiers can be embedded in appropriate systems.

1: Minting DOIs for sample theses

At the DataCite UK client meeting at the British Library on Monday 6th July 2015, three universities assigned a DOI to one thesis in each of their respective EPrints-based repositories. Valerie McCutcheon of the University of Glasgow used the CoinDOI plugin to request and receive back a DOI for a thesis in Enlighten:Theses. The thesis is at **http://dx.doi.org/10.5525/gla.thesis.6423**. Michael Whitton of the University of Southampton uploaded a small XML file directly to the DataCite Metadata Store, and received back the DOI **http://dx.doi.org/10.5258/SOTON/374711** for ePrints Soton. Finally, Stephen Grace of the University of East London used the same CoinDOI plugin to assign a DOI **http://dx.doi.org/10.15123/PUB.3929** to a thesis in ROAR – one which had related data objects (actually two full-length documentary films created as part of the PhD thesis) in data.uel the data repository at UEL

2: Harvesting minted DOIs for inclusion in EThOS

Subsequently, Heather Rosie the EThOS Repository Metadata Manager sought to harvest the DOIs and augmented records for these three theses are now held in EThOS:

http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.650368 (University of Glasgow)
http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.640751 (University of Southampton)
http://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.630167 (University of East London)

The University of Glasgow and University of East London each minted a DOI using the CoinDOI Plugin available in EPrints. The DOIs appear in the record for the thesis on the repository landing page and so are potentially available for harvesting by third parties via OAI-PMH. University of Southampton minted their DOI by submitting an XML file to the DataCite Metadata Store and this DOI has not yet been added to the repository metadata (as of 7th July 2015).

Metadata harvested in the uketd_dc metadata format (the default format used by EThOS) for the University of Glasgow and University of East London theses did not include the newly minted DOI. A harvest of the 'simple' oai_dc format from Glasgow, however, did include the DOI (in a repeated `<dc:relation>` XML tag); and a harvest of the rioxx format for UEL provided the DOI (in the `<rioxxterms:version_of_record>` tag). These DOIs have been added to EThOS.

3: OAI harvesting requirements

The uketd_dc format is the thesis-specific format developed to ensure the highest quality data exchange between the hosting repositories and aggregating services such as EThOS. It includes all the data elements required by the EThOS Service held in individually defined metadata tags. In order for thesis DOIs (and other identifiers such as ORCiDs and ISNIs) to be included in the EThOS service they will need to be mapped from the local repository format to the uketd_dc OAI export format. The uketd_dc format has the following XML elements available for this content:

```
<dc:identifier xsi:type="dcterms:DOI">
<uketdterms:authoridentifier xsi:type="uketdterms:ORCID">
<uketdterms:authoridentifer xsi:type="uketdterms:ISNI">
```

EThOS was adapted ahead of the harvesting exercise to include space for ORCiD, DOI and ISNI numbers relating to a thesis and its author. As Heather Rosie's harvesting showed, it is possible for EThOS to retrieve DOIs from different university repositories. In two cases this was a manual process because relying on the simple oai_dc metadata only worked for Glasgow. The UTD project will need to investigate in a subsequent phase how repositories can offer comprehensive metadata covering PIDs for automated harvesting. This will involve addressing the metadata requirements, discussion with systems suppliers and users, and further testing.

4: Assigning ORCiDs

ORCiDs present a different challenge to DOIs. The identifiers are by design personal ones, owned by researchers rather than their employer or funder. Universities can take advantage of institutional membership (and Jisc has just announced a UK-wide consortium approach for this1), but ultimately the person identified by an ORCiD has to agree to its creation. We have shown in each of the case studies where ORCiDs could be assigned for PhD students. Part of our future work will be to address the challenges of take-up, and we expect to learn from the Jisc/ARMA ORCiD pilot projects as well as current work underway in institutions.

PhD students are increasingly expected to produce other scholarly work as part of their studies (such as conference papers and journal articles), in advance of their thesis. It would make sense to have ORCiD accounts for research students at an early stage to embed them (and the practice of using ORCiDs in scholarly communications) in their research outputs. Early account creation also help students to quote their ORCiD on their thesis; this will help those cases where the repository does not interact or inherit metadata from other systems which might be holding the ORCiD (student register, CRIS, HR, etc.), and of course it is a prominent aid for the readers of the thesis wishing to follow up other publications by the author.

1 https://jisc.ac.uk/news/national-consortium-for-orcid-set-to-improve-uk-research-visibility-and-collaboration-23-jun

5: Adding ORCiDs to EThOS

As ORCiD implementation is starting to take hold in many institutions, EThOS needs to be able to harvest and ingest ORCiDs wherever they are present in the thesis metadata. In June 2015, supported by UTD, the EThOS database was expanded to create new fields for DOI, ORCiD and ISNI identifiers, and the user interface refreshed to present the new fields in the record display. Two examples of EThOS records displaying an ORCiD author ID are **EThOS ID 550256** and **EThOS ID 285773**.

These two example ORCiDs (and 20 more) were identified via a prototype **EThOS-ORCiD importer tool** which allows authors to find their thesis in EThOS and add it to their ORCiD profile. The importer was developed as part of the European ODIN (ORCiD DataCite Interoperability Network) project. It is currently in beta development and is now being further refined by a follow-on European project called **THOR**.

The process used to add these sample ORCiDs was entirely manual, and was carried out simply to produce some sample EThOS records to illustrate the use of ORCiD for thesis authors. The Importer has an Administrator view which allows service owners to view the ORCiD identities of those people who have used the tool for (in this case) EThOS. The ORCiDs were simply copied and pasted from the Importer Admin view and added to the relevant EThOS records.

6: ORCiD ingest as standard

The regular harvesting of over 100 repositories for new and updated thesis metadata by EThOS has so far not resulted in a single ORCiD identifier being exported and ingested into EThOS. There are a number of possible reasons:

1. **Institutions' thesis metadata records do not include ORCiDs.**
   The UTD survey of current thesis workflows (**10.15123/PUB.4274**) invited institutions to provide an example of a thesis record in their repository whose author had an ORCiD which was also present in the record. No such examples were provided, suggesting that no institution yet has a workflow in place for consistent recording of ORCiDs in thesis records.

2. **Some institutions have some ORCiDs present in some thesis records.**
   This may be true – there may be instances where an ORCiD is listed in the thesis metadata, but it would be impossible to spot such ad hoc data during regular mass harvesting of around 3000 new or updated records each month. No institution has yet been in touch with EThOS to indicate that ORCiD data has become available for harvest.

3. **Repository systems are not able to store ORCiD data, or do not output ORCiD data as standard.**

ORCiD plugins exist for both EPrints and DSpace so it unlikely that such repositories cannot hold author identifiers or expose the data in the record. In fact ORCiD identifiers are used for research authors other than PhD students.

4. **Repository systems can hold the data and output it, but the data is not mapped to the output required by EThOS.**
   As described with DOIs, the uketd_dc format is the thesis-specific format developed to support thesis data exchange between repositories and aggregating services such as EThOS. In order for ORCiDs to be included in the data harvested by EThOS, institutions (or their repository support service) will need to map their thesis output from the local repository format to the uketd_dc OAI export format. An alternative solution may be to harvest the RIOXX output from those institutions using RIOXX since RIOXX supports the use and export of ORCiDs alongside other data elements. Those institutions now beginning to export RIOXX data are not currently the same institutions that have implemented ORCiD identifiers for research students.

Recommendations

1. Hold at least three thesis clinics to investigate opportunities and barriers to assigning DOI and ORCiD identifiers in UK universities

The survey has given us a broad understanding of thesis handling in the UK, and the case studies have investigated processes in more detail at individual universities, but UTD would want to engage with more institutions to see how PIDs for theses and their data can become a reality. We propose in phase two to hold a series of clinics in invited universities or groups of universities to see how they could apply PIDs, and explore the opportunities and barriers. We are particularly mindful that our case studies did not involve Scotland, Northern Ireland or Wales, nor offer a wide spread of institutions across England. We would take advice from Jisc about the best way to undertake clinics in phase two (and possibly phase three), and would welcome expressions of interest from those who can bring multiple institutions together.

2. Engage with system suppliers/vendors to identify opportunities for enhancing software with required PIDs

Clinics could be incorporated into user group meetings for some of the common systems involved, and this would have the advantage of presenting requirements to be enacted in specific technologies. But we will also want to talk to those responsible for such systems to prepare the ground for developing software to incorporate creating and handling PIDs; some are open source software but others are proprietary and would require approval, scheduling and development by their owners.

3. Consult with EThOS formally to understand what needs to change in EThOS systems and processes to harvest and display PIDs and related metadata for theses and their data

EThOS acts as a discovery tool for UK doctoral theses, and would thus need to receive, manage and present enhanced metadata for ORCiD and DOI identifiers. At present, there is a combination of automated and manual harvesting actions depending on the set-up of the individual repository. Even where an institution uses the UKETD profile, the full metadata is not always being made available as the harvesting exercise demonstrated. The software and processes should be reviewed in the light of wider development in the repository sector.

4. Evaluate approaches to updating UKETD profile, initially in EPrints, before planning software enhancements

EThOS has been in discussion with Dr Timothy Miles-Board of ULCC about revising the current UKETD profile EPrints plugin to incorporate DOI and ORCID identifiers. EPrints is working towards a major new release (v4) and it might be possible to build into the core metadata fields of EPrints some specific – and now expanded – fields relevant to theses and their data. This would have the advantage of dispensing with the requirement of an extra profile as at present, but would require institutions to use this new version of the software: many are on earlier versions and upgrading will present a challenge.

The work with EPrints will lay the groundwork for considering the approach to take with other platforms. Based on the survey we would expect to consider DSpace, PURE and Fedora/Hydra; there were also responses from institutions using Digitool, Haplo, Converis and Symplectic Elements and a clear trend to incorporate PhD students and their theses and data into new CRIS systems.

5. Investigate requirements and solutions for those institutions that use EThOS as their first-point repository

Several institutions use EThOS as their thesis repository, typically because they do not have an institutional repository or retain a print-based approach. Any national system for enhancing thesis discovery would have to take account of these institutions. Because DOIs must be issued by the organisation that makes the object available (publishes it), the British Library could not undertake minting of DOIs for other institutions – hence the need for the Unlocking Thesis Data project. The British Library would explore whether it could establish a service entity which could act in this publishing role, and consult with potential users of such a system; it may be that the need for a first-point repository is dwindling in the light of REF Open Access requirements (which mandates universities to use an institutional repository) and the growth of CRIS systems.

Note that these five recommendations cover the second phase of Unlocking Thesis Data. In a third phase we would expect to finalise the metadata requirements, develop software enhancements in specific repository/CRIS products and create a comprehensive toolkit and guidance for institutions implementing PIDs for theses and their data.

Appendices

A: Survey design

Unlocking the UK's Thesis Data: Survey of current workflows

This survey on current thesis workflows in your institution will provide baseline information to help the UTD project understand how persistent identifiers might be assigned to UK theses. It will also ask if your own institution has started to think about identifiers, and map the level of interest in this national initiative.

The survey may take up to 20 minutes to complete.

Thank you very much for supporting the UTD Project.

Please read the following carefully before agreeing to take part in this survey. I understand that:

- All views and results from this survey will be treated in the strictest confidence in accordance with the Market Research Society's Code of Conduct
- Individual results or comments will not be shared with any third party unless this is required by law. The project may wish to use some quotations for illustrative purposes, but my specific permission would be sought before any responses are attributed to my institution
- I am free to withdraw from this survey at any time without penalty
- I am free to decline to answer particular questions

[Mandatory tick box]

Q1. Please select your HE institution (tick one only)

[Dropdown list]

If 'Other', please tell us your institution here [free text]

Q2. What are your institution's current requirements for student deposit of doctoral theses? (Tick one)

Mandatory electronic deposit only

Mandatory print deposit only

Mandatory deposit of both print and e-

Mandatory print deposit with voluntary e-deposit

Student may choose any option

Other

Would you like to expand or comment on your response?

Q3. Where are your institution's new/recent *e-theses* held and made accessible? (Multiple tick)

EPrints repository

DSpace repository

Hydra repository

CRIS system – Pure

CRIS system – Converis

CRIS system – Symplectic

British Library EThOS

Other repository or CRIS

E-theses stored locally but not open access

We do not have e-theses

Would you like to comment or expand on your response? [free text box]

Q4. If supplementary data files relating to the thesis are also deposited, where are these held? (tick multiple)

In the same repository or CRIS as the thesis

In a separate data repository and linked to the thesis

In a separate data repository, but not linked to the thesis in any way

Locally in a separate store and linked to the thesis

Locally in a separate store, but not linked to the thesis in any way

Situation has never arisen

Don't know

Please comment or expand on your response [free text box]

Q5. If your institution has more than one repository, please briefly describe your arrangements

e.g. A main repository plus a separate one for theses. A 'publications' repository and a 'data' repository. A single repository, but theses held locally elsewhere. [Free text box]

Q6. *Where* is the initial metadata or a bibliographic record for the thesis first created? (Tick multiple)

In the repository

In the CRIS

In the library cataloguing system

In the student record system

On a standalone system

Other

Would you like to comment or expand on your response? [free text box]

Q7. *Who* normally creates this initial record for the thesis? (multiple tick)

The student

Repository staff

Cataloguing staff

Graduate school staff

Other

Would you like to comment or expand on your response? [free text box]500

Q8. At what point in the entire PhD process is the metadata or bibliographic record for the thesis first created? E.g. *Immediately after award*

Q9. What file formats are accepted for PhD theses in your institution? Please give as much detail as possible. [Free text]

Q10. How are supplementary files such as DVDs handled, and how are they connected to the main thesis in your repository or other system? [Free text]

Q11. What guidance or requirement is there for students to describe the *type or format* of the thesis and the existence of any supplementary material? *e.g. This thesis consists of the main work and two DVDs; This creative studies PhD consists of a published novel plus supporting critical analysis.* [free text]

Q12. What guidance is given to students in writing their thesis *abstract*? If you have online guidance, the URL would be very useful. [free text]

Q13. Which of these describe your current level of activity relating to your research postgraduates and ORCID author IDs? Please tick all that apply (Multiple ticks)

We now require all our research postgrads to have an ORCID

We proactively encourage our research postgrads to register for an ORCID

Student ORCIDs are added to their publications record if available but we don't mandate or encourage their uptake

Students must register for an ORCID at the start of the research degree

Students must hold an ORCID before they deposit their thesis

We are planning to ask research postgrads to get an ORCID

Our repository system has a plug-in which manages the process

Our CRIS system has a registration facility which supports the process

We have institutional membership of ORCID

ORCID implementation is managed as part of student record workflows rather than by the library

We have no plans to implement ORCID workflows for our research postgrads at the moment

We do not use ORCID institutionally at all.

If you can provide an example of a PhD thesis whose author has an ORCID, please provide the URL here [free text]

Q14. What persistent identifiers do you use now for your theses? (Tick multiple)

DataCite DOIs

CrossRef DOIs

Handles

ARKs

ISBNs

Other [please state]

None

Would you like to comment or expand on your response? [free text]

Q15. What persistent identifiers do you use for your research data?

DataCite DOIs

CrossRef DOIs

Handles

ARKs

Other [please state]

None

Would you like to comment or expand on your response? [free text]

Q16. How ready are you to begin assigning DOIs for theses?

We are doing it already

Completely ready, we could start next week

Quite ready, just waiting for guidance from the 'Unlocking Thesis Data' project

At the "thinking about it" stage

Not ready; it will happen at some point in the next five years

DOIs are not on our radar at all

If you have plans for or are thinking about DOIs for theses, please answer Q17 – Q22 if you can. If you have not started to think about DOIs yet, please go to Q21.

Q17. Which departments in the institution have been involved in any planning/discussions about assigning DOIs to theses? e.g. library, repository, academic departments, grad school, DTC … [free text]

Q18. Who / Which department will lead on DOI assignment for theses? [free text]

Q19. What about more complex theses with multiple volumes, supplementary DVDs, or data which is deposited with the thesis? How do you think you might assign DOIs in such cases? [free text]

Q20. What's the first thing that needs to happen, in your opinion, to start getting DOIs assigned to your theses? [free text]

Q21. If you would like to be included in a mailing list to be kept informed of our progress relating to DOIs and theses, please provide your name and email address here:

Q22. Finally we plan to do some case studies of institutions describing the detailed processes relevant to DOI assignment. If you would like to be a case study please indicate here.

No thanks

Yes, we'd like to be a case study institution and I have provided our contact details above.

We won't be able to speak to everyone who volunteers, but will aim for a small range of different repository systems, levels of DOI development and size of institution.

A Included in the thank you pop-up after submission –

Thank you very much for supporting the Unlocking Thesis Data project.

We would now like to gather examples of existing material to help develop templates and a toolkit as part of the project. If you have any flow diagrams or guidance documents describing thesis deposit; thesis management; thesis record creation; or plans for assigning DOIs to theses, please would you share them? Please send any such documents you think would be useful to **Sara.Gould@bl.uk**, or simply send the URLs if they are available online.

Thank you.
Unlocking Thesis *Data project team*

B: Questions for case study interviews

A semi-structured interview schedule was designed to allow flexibility in asking further questions to elicit detail from interviewees. The questions used were

1. **Describe the journey of a PhD thesis at your institution from conception to making available/ publishing.** [ask for any workflow/documentation, if these haven't been shared in advance]
2. **Could you explain which functions in the university handle theses and how, but exclude the examination part (this comes next)?**
3. **What is the process for submission/examination, and what form does the thesis take at this point?**
4. **How does the final form of the thesis get from the student to the repository?** [check: How do you ensure the final form of the thesis is put in the repository? And how do you know it is the final form?]
5. **What guidance is given to students on submission and electronic theses?**
6. **What guidance is given to students on supplementary files (anything that isn't in the main thesis)?** [e.g. Are there limits on formats, sizes, documentation needed to accompany datasets etc.? Are the supplementary files handled differently to the thesis, and if so how?]

C: Related outputs

The outputs for UTD phase one are available at **http://dx.doi.org/10.15123/PROJECT.15**

- Phase 1 survey of UK Higher Education Institutions **http://dx.doi.org/10.15123/PUB.4274**
- Phase 1 baseline survey data **http://dx.doi.org/10.15123/DATA.12**
- University of East London case study **http://dx.doi.org/10.15123/PUB.4301**
- University of Southampton case study **http://dx.doi.org/10.15123/PUB.4302**
- LSE case study **http://dx.doi.org/10.15123/PUB.4303**
- University of the Arts London case study **http://dx.doi.org/10.15123/PUB.4304**
- University of Bristol case study **http://dx.doi.org/10.15123/PUB.4305**
- University of Leicester case study **http://dx.doi.org/10.15123/PUB.4306**

There are also presentations from events held before and during phase one:

- Initial project pitch **http://researchatrisk.ideascale.com/a/dtd/101964-31525**
- Presentation at DataCite client meeting 16 Jan 15 **http://www.slideshare.net/StephenGrace1/fresh-doi-minted-research-20150116**
- Final project pitch at Jisc sandpit workshop 26-27 February 2015 **https://www.slideshare.net/StephenGrace1/unlocking-thesis-data**
- Presentation at DataCite client meeting 6 July 2015 **http://dx.doi.org/10.15123/PUB.4314**

D: Workflows overlaid with suggested PID creation

Workflows for each of the interviewed universities are available in the respective case study:

University of East London case study **http://dx.doi.org/10.15123/PUB.4301**

University of Southampton case study **http://dx.doi.org/10.15123/PUB.4302**

LSE case study **http://dx.doi.org/10.15123/PUB.4303**

University of the Arts London case study **http://dx.doi.org/10.15123/PUB.4304**

University of Bristol case study **http://dx.doi.org/10.15123/PUB.4305**

University of Leicester case study **http://dx.doi.org/10.15123/PUB.4306**