

Wikidata thesis toolkit

Contents page

[Introduction](#)

[Objectives \(or what are we doing and why\)](#)

[Strategic value and impact](#)

[Getting started in Wikidata](#)

[Introduction to Wikidata basics](#)

[Creating a Wikidata account](#)

[Developing familiarity with Wikidata](#)

[Wikidata hints and tips](#)

[Overview of process](#)

[Outline of process flow to upload theses to Wikidata](#)

[Outline of tasks](#)

[Glossary](#)

[Preparation](#)

[Task 1 - project identifier and project page](#)

[Institutional liaison](#)

[What is OpenRefine?](#)

[Preparation of OpenRefine software](#)

[Preparation of data export from source](#)

[Import of data into OpenRefine](#)

[How to edit author and thesis entries](#)

[Editing in OpenRefine](#)

[Editing author name](#)

[Editing thesis title](#)

[Editing date](#)

[Editing thesis type](#)

[Reconciling against Wikidata in OpenRefine and export to Wikidata](#)

[Task 2 - author \(1\)](#)

[Task 3 - thesis title](#)

[Task 4 - author \(2\)](#)

[Review](#)

[How to create supervisor entries and link to author entry](#)

[Editing supervisor entry in OpenRefine](#)

[Preparation of supervisor spreadsheet](#)

[Task 5 - supervisor](#)

[Task 6 - author \(3\)](#)

[Review](#)

[Annual uploads](#)

[Additional options](#)

[Adding external identifiers to theses entries](#)

[EThOS](#)

[ProQuest](#)

[DART Europe](#)

[CORE](#)

[Wikipedia citations](#)

[Exploration of data use](#)

[Querying Wikidata](#)

[Measuring value and impact](#)

[Potential future actions](#)

[Contact details](#)

[Further information](#)

[Appendix](#)

[LSE schema and data modelling](#)

[LSE author schema](#)

[LSE thesis schema](#)

[LSE supervisor schema](#)

[Data modelling](#)

[University of York schema and data modelling](#)

[Example of York author schema \(JSON text\)](#)

[University of York author schema](#)

[University of York theses schema](#)

[University of York supervisor schema](#)

[Data modelling](#)

[Additional appendix \(added July 2023\)](#)

[Reconciling against Wikidata in OpenRefine and export to Wikidata](#)

Introduction

This toolkit is built on the initial work of Helen Williams (London School of Economics Library) of developing a process to upload theses metadata into Wikidata (WD). Subsequent ongoing work by Ruth Elder at the University of York Library has helped to

refine the process further with the aim of creating a process flow ready to roll out to staff undertaking the work as part of “business as usual.”

Following wider interest across the UK higher education (HE) library sector this toolkit aims to reduce the development burden for other institutions looking to establish similar projects. The toolkit is **not** produced as a step by step guide. However it does do its very best to signpost the most direct way to complete the upload of theses metadata to Wikidata.

Moving forward it is hoped that this document will be a foundation to a growing community of practice amongst UK HE institutions who are interested in developing Wikidata work and sharing experience with one another.

Disclaimer: the information within the toolkit should be considered as an active work in progress and is presented with best intent with skills, knowledge and experience at this point in time.

Helen Williams, Metadata Manager, LSE Library

“Back in 2019 I noticed that Wikidata was a growing topic of conversation in the world of metadata. Its power to aid the discovery of unique content for global audiences, create links and show relationships between entities stirred my interest and I was keen to investigate potential benefits to the Library and the wider institution. Over the course of 2020 I began from scratch, reading online content and watching presentations to teach myself the basics of Wikidata and learn about the tools that would be useful to work with it; it was a steep learning curve! I experimented with adding various LSE entities and content types to Wikidata and, with the help of colleagues, settled on LSE’s digitised theses collection as our first Wikidata project. It was challenging, demanding and immensely enjoyable, with a lot of trial and error, such that I’ve described it as our ‘adventures in Wikidata-land: tears and triumphs down the rabbit hole’. Later on in our work I discovered the New Zealand thesis project and their lead was very generous in sharing expertise which has particularly impacted the development of our project page. I was keen to share what I’d learnt, and wrote some articles and spoke at various events to that end, but the real goal was to produce something which would reduce the development burden for other universities to undertake similar projects, and the idea of the toolkit was born. I was delighted to discover Ruth had a similar desire and it’s been a privilege to collaborate on this toolkit together.”

Ruth Elder, Collections Management Specialist, University of York Library

“I first started experimenting with editing in Wikipedia autumn 2021. The Wikidata work of Helen at LSE interested me, and I started work on developing a similar process flow to best meet the needs of York. Like Helen I am entirely self taught, and have relied on the advice of others, the work completed at LSE, by the New Zealand

Thesis project, and on the excellent training materials produced by the Wiki team at the University of Edinburgh.

It has been a truly iterative process of testing, failing and testing again and I continue to adapt and develop processes.

These instructions are for guidance and to hope that they will assist others in avoiding some of the cul-de-sacs that I have visited!

And it has been a treat to work with Helen on this toolkit.”

Objectives (or what are we doing and why)

The Wikidata Thesis Toolkit aims to support projects promoting institutional original research to the widest possible audience through signposting to electronic theses accessible via an existing database or repository. By creating entries on Wikidata with unique identifiers for individual doctoral theses, the resources and research of the University and the entities within their data become part of the linked open data ecosystem.

Links and relationships are established between entities, and connections are made with unique identifiers from a variety of external knowledge systems. This assists machines in interpreting library resources and creates bridges between previously siloed domains, in turn impacting search engine results by providing a fuller picture of globally available data.

Uploading institutional thesis metadata to Wikidata is an example of an “inside out”¹ collection model in that resources which may be unique to the institution are promoted to an audience which is both local and external. The extent of the anticipated increase in reach and engagement with doctoral thesis can be evaluated and reported when appropriate.

This work illustrates the manner in which metadata can extend the access and visibility of scholarly content by breaking down the walls of institutional silos to increase discoverability as a provider of open knowledge to local and global audiences. In addition, the project can help to develop staff digital skills, confidence and experience.

Strategic value and impact

In order to provide library resources (people and time) to undertake a Wikidata project - either as a “proof of concept”, pilot project or as “business as usual”, it will be essential to clearly articulate the value of Wikidata work to non-metadata colleagues and/or senior leadership colleagues.

LSE example

Definition of work

- Develop a process for bulk uploading content to Wikidata using LSETO theses as a dataset.
- Establish regular upload of theses content to Wikidata to maintain the dataset.

¹ Dempsey, Lorcan, Outside-in and inside-out, (2010)

<https://www.lorcandempsey.net/orweblog/outside-in-and-inside-out/>, accessed December 16, 2022.

- Connect unique LSE theses content with external identifiers to increase accessibility online.
- Provide a foundation from which to explore the value of further Wikimedia engagement to the Library and the School.

Strategic aims

[Demonstrate links to relevant Library/institutional strategic aims for institution in question].

Reasons and benefits

Wikidata is a structured database operating as the central data store for all Wikimedia projects. It is a free and open knowledge base that can be read and edited by humans and machines and is multilingual, supporting global access to information (www.wikidata.org). Sharing LSE Library metadata with Wikidata allows us to mint unique identifiers for our content, pulling it into the Linked Open Data ecosystem, and connecting it with unique identifiers from external knowledge systems. Google Knowledge Graphs, digital assistants such as Alexa and Siri, and Wikipedia Infoboxes are all populated with information harvested from Wikidata so the content we add will impact search engine results, create bridges across currently siloed domains and make Library data more widely accessible which enables new connections and discoveries to support global research.

LSETO theses data has been selected as the dataset with which to begin experimental work with Wikidata:

- Boundaried dataset of just over 4000 records.
- Offers value to early career researchers and alumni by promoting their work.
- Opportunity to expose DOIs via Wikidata.
- Increasing traffic to and downloads from LSETO is an anticipated output.

University of York example

Introduction

In the past academic Library and Archive collections were traditionally ringfenced and designed for discovery as locally maintained collections for a local internal audience.² The focus is now increasingly on the academic library as a provider of Open Knowledge, through refocusing skills and resource to organise information in a manner to expand the accessibility and visibility of the institution's unique data and research to reach beyond the local to an external, (and potentially global), audience.

This move is reflected in the University of York Library, Archives and Learning Services (2022 - 2026) objectives of

“engagement with, and exploration of, our collections and research ... enabling new open research and the wide reuse of York’s world-leading research output”

and supporting our global community through

“Improve(ing) access and discovery of our collections online, enabling local and global audiences to discover, use and enjoy them, using an ‘open as possible’ approach”

all of which are enablers to the primary objective of the University of York as

“a University for the public good.”

² Dempsey, Lorcan, Outside-in and inside-out, (2010)

<https://www.lorcandempsey.net/orweblog/outside-in-and-inside-out/>, accessed December 16, 2022.

Introduction to Wikidata

Wikidata is a free linked database ...
... thus influencing Web-based information discovery.

Wikidata was recognised in a white paper produced by the Association of Research Libraries in 2019 as a “means of documenting and surfacing researchers, publications and research data” and “sharing faculty scholarship on an open and accessible platform, enabling the expansion of accessibility and visibility of academic research.”³

Doctoral Dissertations at the University of York

One of the collections of particular value and significance to an academic library and its wider University institution is that of doctoral dissertations. This collection reflects the independent, original research conducted in academic departments, contributing to the development of the field of knowledge in the specific field and is potentially impactful at a global scale, in addition to influencing departmental funding at an institutional level. The doctoral dissertations of the Universities of York, Leeds and Sheffield are available through the White Rose eThesis Online repository (WREO.)

University of York Wikidata Thesis Project

The University of York Library Wikidata Thesis Project aims to promote York original research to the widest possible audience through signposting to electronic theses accessible via WREO. By creating entries on Wikidata with unique identifiers for York doctoral theses, the resources and research of the University and the entities within their data become part of the linked open data ecosystem⁴...

...The Wikidata Thesis Project is an example of the “Inside out” collection model in that resources which may be unique to the institution are promoted to an audience which is both local and external.⁵ The extent of the anticipated increase in reach and engagement with York doctoral thesis will be evaluated when appropriate.

This project illustrates the manner in which metadata can extend the access and visibility of scholarly content by breaking down the walls of institutional silos to increase discoverability as a provider of open knowledge to local and global audiences. In addition, the project can help to develop staff digital skills, confidence and experience.

³ Association of Research Libraries, ARL white paper on Wikidata: opportunities and recommendations (2019)

<https://www.arl.org/wp-content/uploads/2019/04/2019.04.18-ARL-white-paper-on-Wikidata.pdf>, accessed December 16, 2022.

⁴ Clark, Jason A., Williams, Helen K.R., and Rossmann, Doralyn. ‘Wikidata and Knowledge Graphs in Practice: Using Semantic SEO to Create Discoverable, Accessible, Machine-readable Definitions of the People, Places, and Services in Libraries and Archives’. 1 Jan. 2022 : 1 – 14.

<https://content.iospress.com/articles/information-services-and-use/isu220171>, accessed December 16, 2022.

⁵ Dempsey, Lorcan, Outside-in and inside-out, (2010)

<https://www.lorcandempsey.net/orweblog/outside-in-and-inside-out/>, accessed December 16, 2022.

Getting started in Wikidata

Introduction to Wikidata basics

Knowledge of Wikidata and its data input conventions will be essential prior to beginning a thesis project.

The following resources will provide the basics:

- Introduction to Wikidata
<https://www.wikidata.org/wiki/Wikidata:Introduction>
- WikiEdu Introduction to Wikidata
<https://dashboard.wikiedu.org/training/wikidata-professional/introduction-wikidata-professional>

Note that this toolkit assumes knowledge from the above resources.

Further useful reading:

- Association of Research Libraries, ARL White Paper on Wikidata, (2019)
<https://www.arl.org/resources/arl-whitepaper-on-wikidata/>, accessed March 24 2022
- Evans, Jason, Library data as linked open data, Catalogue & Index 199, (2020)
https://cdn.ymaws.com/www.cilip.org.uk/resource/collection/5F814B6D-500C-42B2-9D5F-E6E3C550C24A/C&1199Evans_Linked_open_data.pdf, accessed March 24 2022
- Lemus-Rojas, Mairelys and Lydia Pintscher, Wikidata and Libraries: facilitating open knowledge (2018)
https://scholarworks.iupui.edu/bitstream/handle/1805/16690/Lemus-Rojas_Pintscher_Wikidata_2017-07-03.pdf, accessed March 24 2022
- OCLC, Hanging Together: Wikimedia, (2020)
<https://hangingtogether.org/category/wikimedia/>, accessed March 24 2022
- Poulter, Martin and Nick Sheppard, Wikimedia and universities: contributing to the global commons in the Age of Disinformation. Insights 33 (1): 14, (2020)
<https://insights.uksg.org/articles/10.1629/uksg.509/>, accessed March 24 2022
- Sheppard, Nick, Wikimedia in universities, Leeds University Library blog, (2019)
<https://leedsunilibrary.wordpress.com/2019/02/01/wikimedia-in-universities/>, accessed March 24 2022

Creating a Wikidata account

Anyone working on the project will require an individual Wikidata account.

- Create an account from the Main page and select *Create account*



- Select *Preferences*, followed by *Gadgets*, and in addition to the automatically selected gadgets those below may be useful:
 - Duplicate References (enables copying a reference to add to other statements on Qid)
 - Drag n Drop (enables dragging and dropping of references from Wikidata or Wikipedia)
 - Current Date (automatically adds the date to 'retrieved' references)
 - Recoin (a relative completeness indicator identifying relevant but absent properties)
 - Rearrange values (enables values of a property to be re-ordered)
- Create user page by clicking on your username hyperlink, bringing you directly to your user page and the option to create it - eg <https://www.wikidata.org/wiki/User:HelsKRW>
 - Important to declare institutional affiliation, demonstrating awareness of conflict of interest.
- Create a commons.js page to add scripts which will give you additional functionality when using Wikidata – eg <https://www.wikidata.org/wiki/User:HelsKRW/common.js>. Further information is available here https://www.wikidata.org/wiki/Help:User_scripts
 - The following scripts may be useful (copy and paste to common.js page)

```
importScript('User:1Veertje/identifierInput.js');
importScript( 'User:Magnus_Manske/mixnmatch_gadget.js' );
importScript('User:Vvekbv/recoin_id.js');
importScript( 'User:Btwashburn/iiif-mirador.js' );
importScript ( 'User:Magnus_Manske/author_strings.js' );
importScript( 'User:Abbe98/copy-Qid.js' );
importScript( 'User:Bargioni/UseAsRef.js' );
importScript( 'User:Bargioni/moreIdentifiers defaultconf.js' );
importScript( 'User:Bargioni/moreIdentifiers.js' );
```

The Wikidata Affinity Group has a recording on gadgets and user scripts which provides further information and demonstrations.
<https://docs.google.com/document/d/1QQn1jS89TWGiLcXOpbRGXCcDVxWKSgEgq6Pj8zXxt9w/edit>

Developing familiarity with Wikidata

Before beginning a full scale project it can be helpful to do some basic editing to develop familiarity with Wikidata. Dan Scott's blog post '[creating and editing libraries in Wikidata](#)' is a good starting point.

LSE basic record creation and editing examples

LSE Library <https://www.wikidata.org/wiki/Q2371017>

Creation of new items for component parts of the Library

- LSE Digital Library <https://www.wikidata.org/wiki/Q96354844>
- LSE Press <https://www.wikidata.org/wiki/Q74433811>
- LSE Research Online <https://www.wikidata.org/wiki/Q96373472>
- LSE Library online catalogue <https://www.wikidata.org/wiki/Q82487943>
- Link everything together using reciprocal 'has part' and 'part of' statements

Creation of new items for different content types

- Blogs <https://www.wikidata.org/wiki/Q98204952>
- Online exhibitions <https://www.wikidata.org/wiki/Q105555006>
- Organisation represented in archives <https://www.wikidata.org/wiki/Q82749481>
- OA journal issue <https://www.wikidata.org/wiki/Q110636818>

York basic record creation

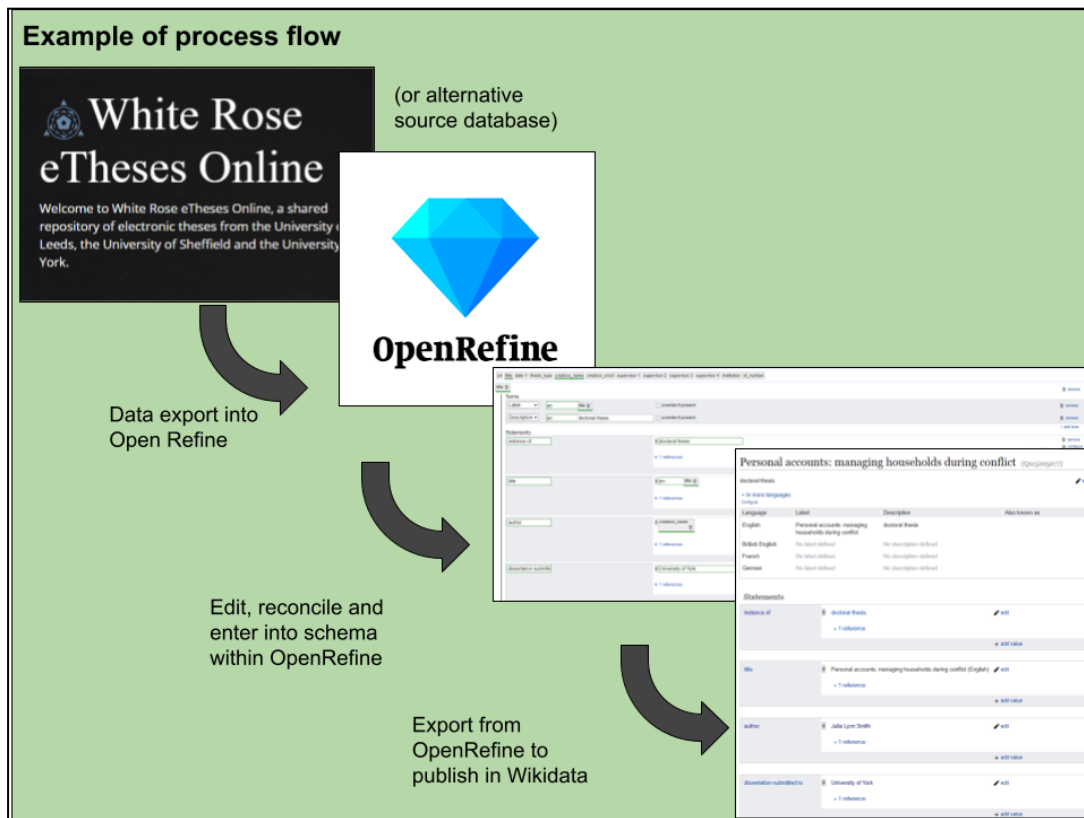
- University of York Library <https://www.wikidata.org/wiki/Q24753366>
- White Rose University Press <https://www.wikidata.org/wiki/Q74433041>
- White Rose Etheses Online <https://www.wikidata.org/wiki/Q24779464>

Wikidata hints and tips

- Selecting 'a' on the keyboard adds a new statement if short cut keys are enabled in Gadgets.
- Double click on a Qid to highlight for copying - or use the script ['User:Abbe98/copy-Qid.js'](#) to display the 'Copy Qid' button.

Overview of process

Outline of process flow to upload theses to Wikidata

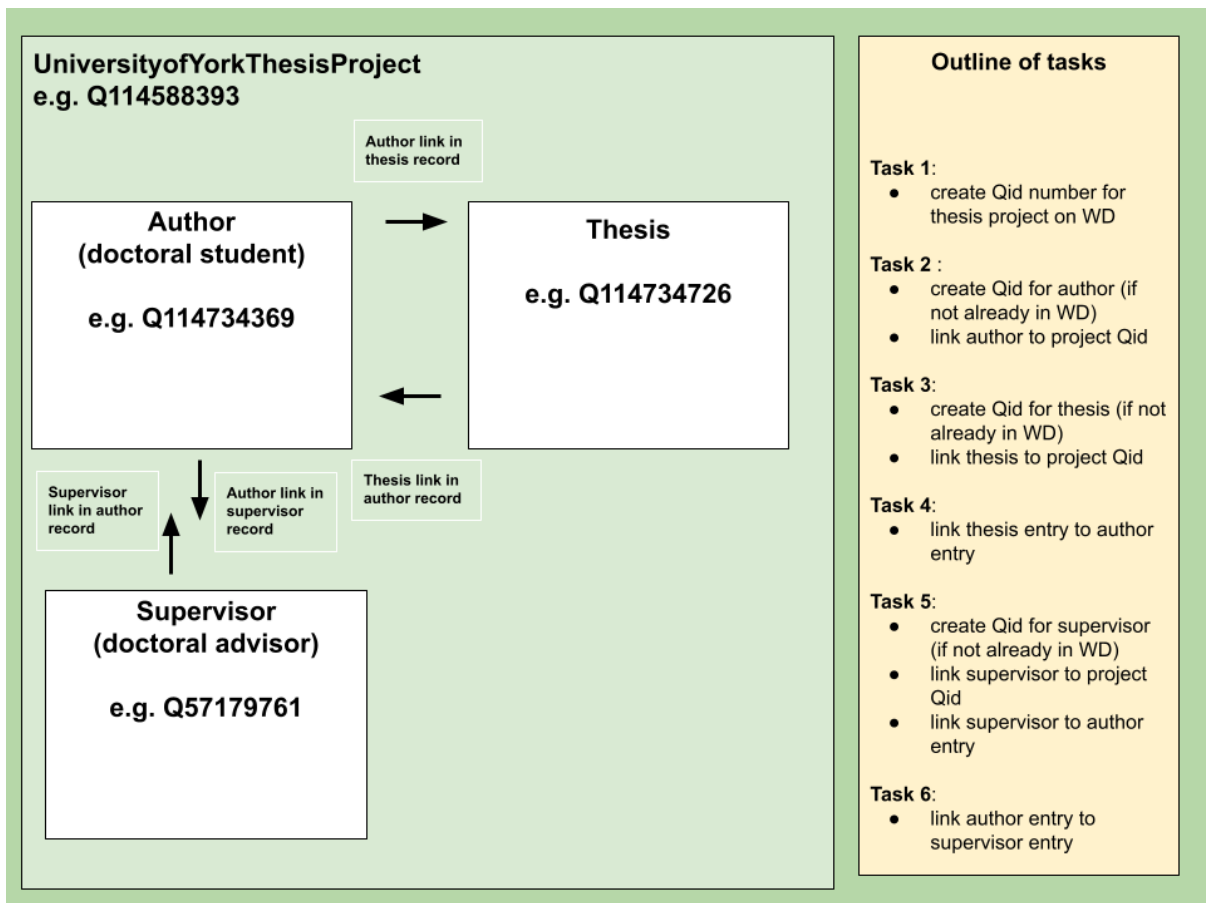


Outline of tasks

The diagram below gives an outline of the tasks which need to be completed to link the thesis statement, author statement, and supervisor statements together in Wikidata, and to identify them as part of a specific project.

The order set out here was based on the premise that the addition of the thesis and author information to Wikidata is the primary action of the project in order to promote university research. The addition of the supervisor entry and linking it to the author entry in Wikidata gives additional information around research relationships.

As mentioned earlier, this work is still in development and this is not a definitive process order. It could be that the tasks around creating the supervisor entries are completed first rather than last as listed here (see page 28 for an example of this workflow). As skills and experience are developed, it may be appropriate to review the best order to work in for the specific individuals and institution involved. However for clarity, this toolkit follows the task order set out below (task 1 - 6.)



Glossary

| Term | Explanation |
|--------------------------|---|
| OpenRefine | OpenRefine (OR) is software developed to assist in the editing and manipulation of large quantities of data. |
| reconcile/reconciliation | Action to match data (names or titles in this example) uploaded to OpenRefine against existing data in Wikidata. Example: checking to see if a doctoral supervisor already has an entry (a Qid) in Wikidata. |
| schema | This is the metadata framework designed to input specific statement information into Wikidata in an acceptable format. See appendix for examples. |
| Qid | Unique identifier for each entity entered into Wikidata shown as a Q number. Example: Mary Garrison: Q87339749 https://www.wikidata.org/wiki/Q87339749 |
| Project statement | Item (Qid) entered (in this example) into each author, title and |

| | |
|--|---|
| | <p>supervisor record to ensure that the entries can be searched and linked together as associated with this project. To be entered into “on focus list of Wikimedia project” statement. Example: LSEThesisProject https://www.wikidata.org/wiki/Q112895606</p> <p>UniversityofYorkThesisProject https://www.wikidata.org/wiki/Q114588393.</p> |
|--|---|

Preparation

Task 1 - project identifier and project page

Through the creation of a unique project identifier all entities linked to the project become searchable through the list name. This is useful for managing the data, and writing and analysing queries further down the line.

Example:

LSEThesisProject

<https://www.wikidata.org/wiki/Q112895606>

UniversityofYorkThesisProject

<https://www.wikidata.org/wiki/Q114588393>

| | |
|-----------|--|
| Objective | To create a unique project identifier and page to represent the project |
| Process | <p>Decide on the project name. In Wikidata enter the project name in the search box to check that the name is not already in use. The screen below should show.</p> <div data-bbox="405 913 1390 1386" style="border: 1px solid black; padding: 5px;"> <p>Special page</p> <h3>Search results</h3> <p>To search for Wikidata items by their title on a given site, use Special:ItemByTitle.</p> <p>Q UniversityofElsewhereThesesProject <input type="button" value="Search"/></p> <p>Advanced search: <input type="button" value="Sort by relevance"/></p> <p>Search in: <input type="button" value="(Main)"/> <input type="button" value="(Property)"/></p> <p>There were no results matching the query. You may create a new item for "Universityofelsewherethesesproject".</p> <p>Privacy policy About Wikidata Disclaimers Mobile view Data access Developers Statistics Cookie statement</p> </div> <p>Click on create new item. Using the examples for LSE and York given above, create a new item in Wikidata using the same statement fields. Publish. Search for project name in Wikidata search box.</p> <p>Set up a Wikidata project page, following this example. https://www.wikidata.org/wiki/Wikidata:WikiProject_LSEThesisProject This page can record the work of the project and will allow all project links and data to be collected in one location. Ensure addition of the category "Wikidata thesis project" so that it can be found along with all other thesis projects.</p> <p>On the Wikidata item for the project link to the newly created project page under "Multilingual sites".</p> |
| Result | Any author, thesis or supervisor associated with the project can be readily identified and queried. |

Institutional liaison

Institutions have different internal policies and risk appetites. If your institution takes a cautious approach then you may wish to discuss a Wikidata project with the institutional records/data protection manager before dealing with names data. Author and supervisor names within the metadata will already be in the public domain, but a thesis project will bring data together in new ways. Consequently your institution may advise a Data Protection Impact Assessment (DPIA). You may also wish to create an internal policy for the creation of Qids for living people to include mandatory, optional and disallowed properties, a complaints procedure, a sample response and an escalation hierarchy if concerns are raised. (To date this situation has not arisen for LSE or York). You may also wish to update thesis submission forms and/or repository FAQs to make it clear that metadata will be included in Wikidata.

Libraries need to confirm that the deposit agreements of the theses allows the sharing of the metadata (and if there are any restrictions to this.) Copying data from the open repositories to Wikidata simply exposes data already available within the public domain and enhances the discoverability of the content.

University of York

“Metadata in White Rose Research Online can be re-used in any medium without prior permission, for not-for-profit purposes, provided the OAI Identifier or a link to the original metadata record are given.”

LSE

“LSE Theses Online may incorporate metadata or documentation into public access catalogues for the Work(s) and the Author(s), including, but not limited to, Wikidata, CORE, DART-Europe and ProQuest Dissertations and Theses. A citation/citations to the Work(s) will always remain visible in the repository during its lifetime.”

It will be helpful to know whether your institution already has any theses in Wikidata. It might be that another Wikimedian or library has already entered one or more for referencing (for LSE just one had been added prior to our project).

Institutions with theses already in Wikidata can be viewed on the SourceMetadata table https://www.wikidata.org/wiki/Wikidata:WikiProject_Source_Metadata/Theses_by_institution or a SPARQL query can be run to obtain a list of theses from a particular institution. In this query <https://w.wiki/6cBA> swap Q174570 for your institution.

What is OpenRefine?

OpenRefine (OR) is software developed to assist in the editing and manipulation of large quantities of data. It is particularly useful in preparing data for input to Wikidata through various Wikidata related add-ons which are now available as part of OR.

Preparation of OpenRefine software

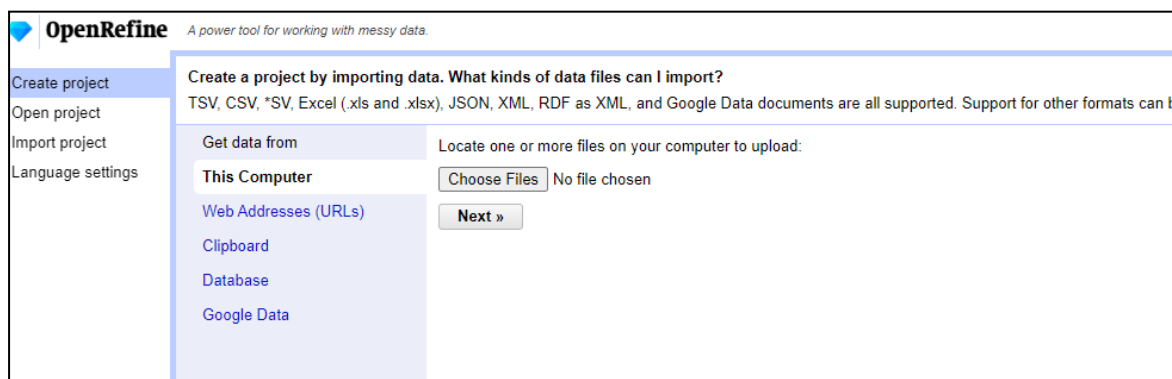
Download the most recent version of OpenRefine.

If your IT set up does not already have java, or will not allow download of java then use OR version with embedded java.

LSE have worked with OR on laptops. York have saved OR to a memory stick to save laptop memory.

Open OR from start menu or memory stick (click on OpenRefine exe.)

Black script screen will open up, and within a few moments an additional screen will appear with OR showing on tab.



Preparation of data export from source

This will be dependent on your data source/repository, and the manner in which data can be exported from it. The amount of editing required in OpenRefine prior to upload to Wikidata will be dependent on the specific content of the export. If possible, it is worth exploring if there is a way to request or create a customised export which can greatly reduce the amount of data editing required in OpenRefine.

University of York

Data is exported from the White Rose eThesis Online repository (WREO), which holds the doctoral theses from the Universities of Leeds, Sheffield and York.

<https://etheses.whiterose.ac.uk/>

Example of customised export for Wikidata Thesis Project:

[WREO Wikidata Export](#)

LSE

Data is exported from LSE Theses Online, which holds born-digital and digitised theses from LSE. <https://etheses.lse.ac.uk/>

LSE did not have a customised export and instead downloaded data one year at a time, starting with the oldest data where the datasets were smaller. We carried out the following steps in Excel, prior to edits in OpenRefine.

- Delete all columns except repository id, creator names, title, pages, date, qualification name, thesis type, DOI, supervisors.
- Concatenate first and surname columns to create single author name column.
- Repeat for supervisor names.
- If there are cells without supervisor data use clear contents tool on blank cells before uploading to OpenRefine (otherwise OpenRefine detects the formula on a visually empty cell and tries to reconcile it).
- Create base URL column and concatenate with repository id to create URL column.
- Insert new column labelled 'Description'. Concatenate phrase 'doctoral thesis by' with full author name to create metadata for description statement.

Example of edited export for LSE Wikidata thesis project:

[LSETO Edited Export Template](#)

When first learning to manipulate data within OpenRefine and Wikidata, it is suggested to start working with small data sets.

(After exporting data from the original dataset, consider editing down into sets of 10 in separate spreadsheets to work through the process. Increase set size as knowledge, skills and confidence grow. It is easier to manually check through 10 entries rather than 100.)

Import of data into OpenRefine



The screenshot shows the OpenRefine web interface. The header includes the OpenRefine logo and the tagline "A power tool for working with messy data." The left sidebar contains navigation options: "Create project" (highlighted), "Open project", "Import project", and "Language settings". The main content area is titled "Create a project by importing data. What kinds of data files can I import?" and lists supported formats: "TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other" (partially visible). Below this, there are options for "Get data from": "This Computer" (selected), "Web Addresses (URLs)", "Clipboard", "Database", and "Google Data". The "This Computer" option is expanded to show a "Choose Files" button and the text "No file chosen". A "Next »" button is also visible.

In OpenRefine click on *Create project*.
 Select *Choose Files*.
 Choose the appropriate file and select *Next*.
 File will import and open showing a preview of the data.

OpenRefine A power tool for working with messy data.

Create project: « start over | Configure parsing options | Project name: Test 1.xls | Tags: | Create project »

Open project | Import project | Language settings

| url | title | date | thesis_type | creators_name | creators_orcid | supervisor | institution | id_number |
|---|--|------|-------------|----------------------|----------------|---|--------------------|--------------------|
| 1. https://etheses.whiterose.ac.uk/26049/ | Refugee Teachers, Pedagogy and Professional Development: The Pursuit of Quality Education in UNRWA, Jordan. | 2019 | PhD | Lucy Janet Atkinson | | Jacob Eriksson; Nina Caspersen; Lucy Atkinson | University of York | |
| 2. https://etheses.whiterose.ac.uk/17379/ | Personal accounts: managing households during conflict | 2016 | PhD | Julia Smith | | Tony Heron; Ignacio Jurado | University of York | uk.bl.ethos.714401 |
| 3. https://etheses.whiterose.ac.uk/20349/ | The Question of Agency and the Reintegration of Former Child Soldiers into Civilian Society - A Case Study of Colombia | 2016 | PhD | Alexandra Mária Kiss | | Nina Caspersen | University of York | |
| 4. https://etheses.whiterose.ac.uk/22242/ | Using Cultural Heritage as a Tool in Post- | 2014 | PhD | Katherine Bishop | | Liam Clegg | University | uk.bl.ethos.759868 |

Parse data as

- Excel files
- JSON files
- Line-based text files
- CSV / TSV / separator-based files
- Fixed-width field text files

Worksheets to Import: Select all | Deselect all | Test 1.xls#Theses 16 rows

Ignore first 0 line(s) at beginning of file
 Parse next 1 line(s) as column headers
 Discard initial 0 row(s) of data
 Load at most 0 row(s) of data

Store blank rows
 Store blank cells as nulls
 Store file source
 Store archive file

Update preview | Disable auto preview

Rename if wanted.
 Click on *Create Project*.
 Project file will open in OR.

15 rows | Extensions | Wikidata

Show as: rows records | Show: 5 10 25 50 100 500 1000 rows | « first | < previous | 1 | next > | last »

| All | url | title | date | thesis_type | creators_name | creators_orcid | supervisor | institution | id_number |
|-----|--|---|------|-------------|------------------------|----------------|---|--------------------|--------------------|
| 1. | https://etheses.whiterose.ac.uk/26049/ | Refugee Teachers, Pedagogy and Professional Development: The Pursuit of Quality Education in UNRWA, Jordan. | 2019 | PhD | Lucy Janet Atkinson | | Jacob Eriksson; Nina Caspersen; Lucy Atkinson | University of York | |
| 2. | https://etheses.whiterose.ac.uk/17379/ | Personal accounts: managing households during conflict | 2016 | PhD | Julia Smith | | Tony Heron; Ignacio Jurado | University of York | uk.bl.ethos.714401 |
| 3. | https://etheses.whiterose.ac.uk/20349/ | The Question of Agency and the Reintegration of Former Child Soldiers into Civilian Society - A Case Study of Colombia | 2016 | PhD | Alexandra Mária Kiss | | Nina Caspersen | University of York | |
| 4. | https://etheses.whiterose.ac.uk/22242/ | Using Cultural Heritage as a Tool in Post-war Recovery: Assessing the Impact of Heritage on Recovery in Post-war Dubrovnik, Croatia | 2014 | PhD | Katherine Bishop | | Liam Clegg | University of York | uk.bl.ethos.759868 |
| 5. | https://etheses.whiterose.ac.uk/6712/ | The influence of research on state building policy with special reference to security sector reform: the case of Sierra Leone | 2014 | PhD | Andrea Edoardo Varisco | | Sultan Barakat; Adam White | University of York | uk.bl.ethos.617241 |
| 6. | https://etheses.whiterose.ac.uk/5207/ | The Neglected Pillar of Recovery: A Study of Higher Education in Post-war Iraq and Libya | 2013 | PhD | Sansom Milton | | Barakat Sultan; Hardman Frank | University of York | uk.bl.ethos.605191 |
| 7. | https://etheses.whiterose.ac.uk/4613/ | The State of Democratisation in Afghanistan: An examination of the state-citizen relationship in a fragile context. | 2013 | PhD | Anna Larson | | Sultan Barakat | University of York | uk.bl.ethos.581721 |
| 8. | https://etheses.whiterose.ac.uk/3960/ | Violence and Fragility: A Study of Violent Young Offending in Yemen and Other Fragile States | 2013 | PhD | Alexandra Lewis | | Sultan Barakat; Matt Matravers | University of York | |
| 9. | https://etheses.whiterose.ac.uk/8980/ | An Investigation of Corruption and Clientelism in Afghanistan Hindering Law Enforcement: With a Specific Case Study in the Afghan National Police | 2012 | MPhil | D Singh | | | University of York | |
| 10. | https://etheses.whiterose.ac.uk/1971/ | Addressing the needs of families of the Missing: A test of contemporary approaches to transitional justice | 2011 | PhD | Simon Robins | | Sultan Barakat | University of York | uk.bl.ethos.547371 |

How to edit author and thesis entries

| | |
|-----------|---|
| Objective | <p>Task 2: to check for existing entities on Wikidata for the author, and create a new entity if this does not already exist.</p> <p>Task 3: to check if the thesis title already exists in Wikidata. If not, a new entity is created to represent it.</p> <p>Task 4: to link the thesis entry on Wikidata back to the author entry in order to establish the relationship between the two.</p> |
|-----------|---|

To complete these tasks as efficiently as possible, some preparatory editing of the data needs to be completed in OpenRefine to enable the most accurate matching in Wikidata.

Editing in OpenRefine

The editing required in OpenRefine will depend on whether you used an edited export, like LSE, or a customised export, like York. The notes below should help you in the editing process, whatever your export looks like.

Initial editing LSE

- Delete unnecessary columns, such as base URL, author first names and surnames, by selecting *Edit column—Remove this column*.
- Remove full stops from the ends of title fields using *Edit cells—Transform—and entering value.replace(/.\$/, "")*
- Transform DOIs to uppercase (necessary for Wikidata) by selecting *Edit cells—Common transformations—To uppercase*.
- Check date entry shows as single year by selecting *Facet—text facet* and editing any outliers in the left hand facet pane.
- Reconcile title column against doctoral thesis Q187685, or thesis Q1266946.
- In Facet/Filter pane check the title:judgement box to see if any titles matched or are all identified as 'none'. Unless your data has already been added, most results should be 'none' for no match.
- Select any titles that have matched. If titles are matched correctly delete from OpenRefine project by starring, selecting *Facet—Facet by star* in the All column. Select true in Facet/Filter pane, then in the All column select *Edit rows—Remove matching rows*. Reset to get back to full dataset.
- If matches are false, select *Choose new match* in title column cells.

Initial editing York

Check all text is in appropriate format - ie - not capitalised through using *Facet/Text Review* option.

Where the name only shows initials, go into the electronic copy of the thesis and look for additional details, and edit into author column. This will ensure a more accurate matching against items

Editing author name

Open project in OR.

Check that the formatting is correct throughout the column, using *Facet/Text Review* option.

Edit out any excess white space.

Edit cells/Common transformations/trim leading and trailing whitespace + collapse consecutive whitespace.

Editing thesis title

Check all text is in appropriate format - ie - not capitalised through using *Facet/Text Facet* option.

(To change a name appearing all in capitals select *Edit cells/Common transformations/To title case* - which capitalises the first letter of every word of the title.)

Edit out any excess white space.

Edit cells/Common transformations/trim leading and trailing whitespace + collapse consecutive whitespace.

Editing date

The date entry should show as a single year - ie - 2011.

If needed, use the split column option:

To separate into separate cells select *Edit column/Split into several columns*.

Select *Separate* and enter "." (full stop or whatever is in place as separator.)

Anything to the right of the full stop will appear in a separate column, (which can then be deleted.)

Edit out any excess white space.

Edit cells/Common transformations/trim leading and trailing whitespace + collapse consecutive whitespace.

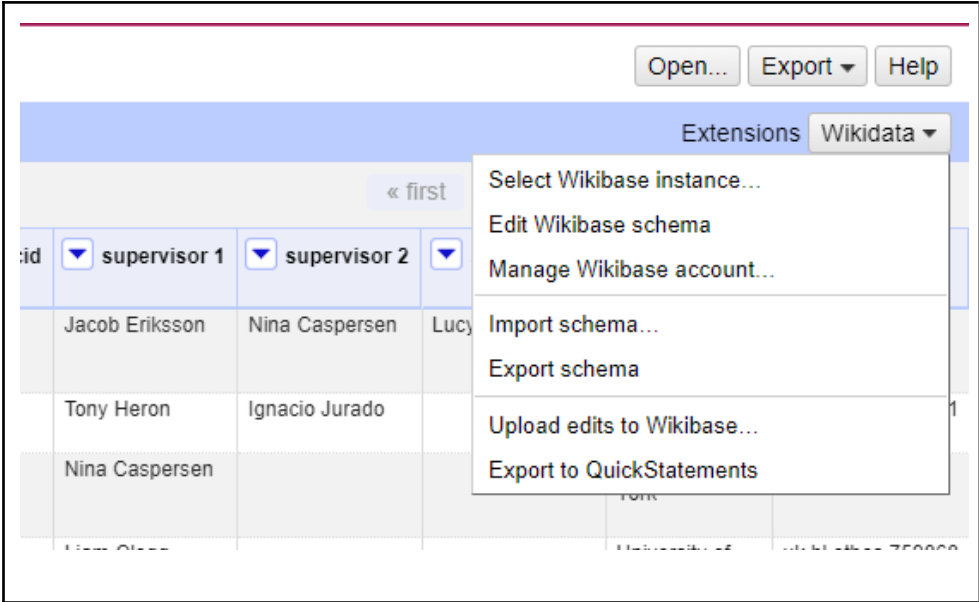
Editing thesis type

Through *Facet* option, edit out any items which are not listed as PhD.

(See "How to edit out rows in OpenRefine" [Link](#))

Reconciling against Wikidata in OpenRefine and export to Wikidata

Task 2 - author (1)

| | |
|-----------|--|
| Objective | <p>To create Q identities (Qids) for authors who do not already have a listing in Wikidata.</p> <p>To give all authors of University of Elsewhere theses entered into Wikidata the UniversityofElsewhereProject statement (which allow them to be searchable, and to be identified as part of this project.)</p> |
| Process | <p>Select authors name column and select <i>Reconcile/Start Reconciling</i>. Click on <i>Wikidata reconc link (en)</i>. Reconcile against Q5 (human.) Select second method of reconciling by using author Orcid number. Select <i>Start reconciling</i>. (The larger the set, the slower the process. This may take a few minutes.) After reconciliation there will be some names reconciled against names/Qids which are already in Wikidata (P50s). The majority will probably not. To identify these select <i>Reconcile/Facets/By judgement type/true or false</i>. Look at those reconciled against existing Qids (highlighted in blue and linking to Wikidata entry.) Check these for false positives.</p> <p>For those that are not already listed in Wikidata, create new items for author names. Select these items through <i>Reconcile/Facets/By judgement/true or false</i>. Then <i>Reconcile/action/create new items for each cell</i>. Select all items in the set.</p> <div data-bbox="400 1335 1382 1935" style="border: 1px solid black; padding: 10px;">  </div> <p>Select <i>Extensions Wikidata</i>, then <i>Import schema</i>.</p> |

| | |
|--------|---|
| | <p>Export into schema: import <i>author schema</i>. Toolkit: Copy of Preferred Creator Schema (University of York) (Also see Appendix for screenshot) “Example of York Creator/Author Schema (Import text to OpenRefine)” Select <i>Edit Wikibase schema</i>.</p> <p>Check in <i>Preview</i> and make any alterations or corrections as needed. Export to <i>Upload edits to Wikibase</i>. author names should now appear with individual Qids in Wikidata. Check through “contributions” in your personal Wikidata account to see if all looks correct.</p> |
| Result | All authors associated with project have a Qid and this is linked to the <i>UniversityofElsewhereProject</i> statement. |

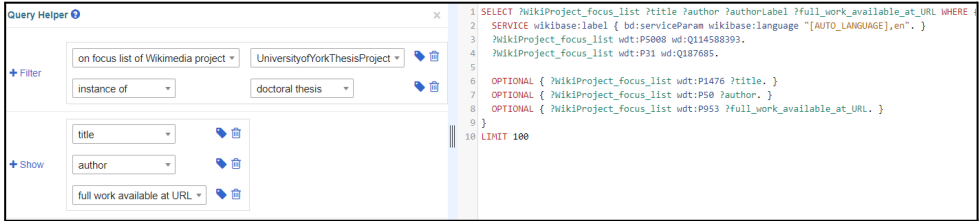
Task 3 - thesis title

| | |
|-----------|---|
| Objective | To create Qids for the thesis title, and to link the title to the author record (created in task 2.) |
| Process | <p>Reconcile all names against author if not already completed (should be a match for every entry.) Reconcile title against doctoral thesis. (In the vast majority of cases there will not be a match.) Identify those titles which need a new item creating for them (to give them a Qid in Wikidata.) To identify these select <i>Reconcile/Facets/By judgement/true or false</i>. Create a new item for each thesis without a match. <i>Reconcile/action/create new items for each cell</i>. Select <i>Extensions Wikidata</i>, then <i>Import schema</i>. Export into Wikidata schema page:import Toolkit: Copy of Preferred Title Schema (University of York) (Also see Appendix for screenshot) Make sure publication year (date) and project statement are included. Check in <i>Preview</i> and make any alterations or corrections as needed. Export to <i>Wikidata Upload</i>. Export into Wikidata. Check through “contributions” in your personal wikidata account to see if all looks correct.</p> |
| Result | <p>All thesis titles will now have a Qid and have an author statement, and a project statement. (At this point there is no link between the author’s record and the thesis Qid - i.e there is no mention of the thesis in the author’s Qid record.)</p> |

Task 4 - author (2)

| | |
|-----------|---|
| Objective | To ensure that the author record Qid has a statement in it linking to the title Qid. |
| Process | <p>Reconcile on title (there should be a match for every entry from the work completed in task 3.)</p> <p>Reconcile on author (there should be a match for every entry from the work completed in task 2.)</p> <p>Export into author schema (including thesis and project statements.)</p> <p>Toolkit: Copy of Preferred Creator Schema (University of York)</p> <p>Export into Wikidata.</p> <p>Check through “contributions” in your personal Wikidata account to see if all looks correct.</p> |
| Result | <p>Author record shows title live link.</p> <p>Title record shows author live link.</p> <p>Author and title records both show project statement link.</p> |

Review

| | |
|--------|--|
| Review | <p>Take time to check that all the information is showing correctly in both author and thesis records.</p> <p>Run query to check all appears correctly in relation to project statement. (Link to query (UniversityofYorkThesisProject))</p> <p>Note: replace Qid for UniversityofYorkThesisProject with the Qid for your project statement (created in task 1.)</p>  |
|--------|--|

How to create supervisor entries and link to author entry

Creating a doctoral supervisor/advisor entry and linking it to the author entry in Wikidata gives additional information around research relationships. This can be undertaken to directly follow on from tasks 1-4 (listed above), or completed at a separate point in time as a stand-alone piece of work.

Some preparatory work needs to be completed around supervisor information prior to matching (or creating new items) in Wikidata.

Editing supervisor entry in OpenRefine

In the export listed in OpenRefine multiple supervisor entries may appear in the same cell as one another.

To separate into separate cells for supervisor 1 and 2 (or more), select

Edit column/Split into several columns.

Select *Separate* and enter “;” (semicolon.)

Each name will appear in a separate column.

Check that the formatting is correct throughout each column.

Edit out any excess white space.

Edit cells/Common transformations/trim leading and trailing whitespace + collapse consecutive whitespace.

Preparation of supervisor spreadsheet

From the initial export, extract details of supervisor and author (doctoral student) associated with the supervisor.

Suggested process (York)

(probably not the best - but perhaps not the worst either)

From WREO export edited above, select *Export* option from top of OpenRefine page, and then *Excel*.

Save and name document as appropriate.

Copy and list all supervisor names into a single column in a separate tab.

Select *Data/Remove Duplication*.

Select *Sort/Filter/Sort A-Z*.

A manual matching process is then required to match the appropriate doctoral student to the correct doctoral advisor.

Return to the first tab and copy authors name column and insert adjacent to each supervisor column.

Sort the supervisor column A-Z

Work down the list of supervisors and note the names of their doctoral students on the second tab (there may be several students to each advisor.)

Complete for each supervisor column on tab 1 until all doctoral students are listed.

Complete the rest of spreadsheet as shown in this example:

[Example of supervisor spreadsheet \(York\)](#)

Include role description, URL link to biography page on University website, ORCID number if available, and years working at the University if not still employed.

This will involve checking departmental websites, and looking wider across Google. It can be useful at this point to check if the supervisor already has an entry in Wikidata (and noting the Qid.)

Note

(The initial load of historic theses with associated supervisor information is an investment in effort and time, but once a dataset is created the updating of it will take a relatively short amount of time to update annually.)

Task 5 - supervisor

This task requires input from the pre-prepared spreadsheet listing doctoral supervisors and doctoral students.

Note that any spelling mistakes in manual inputting of names, (or slightly different format of names), will mean that they do not match in the reconciliation process.

| | |
|-----------|--|
| Objective | <p>To ensure that all the individuals listed as supervisors have a Qid record in Wikidata.</p> <p>Also to ensure that the statements fields appropriate to this project are displayed (including the project statement) if the individual already has a Wikidata entry.</p> <p>To link the supervisor (task 5) to the doctoral student name (author) from task 2.</p> |
| Process | <p>Import the supervisor spreadsheet and edit if necessary.</p> <p>Reconcile against the doctoral student (author) name (showing in Wikidata from task 2 action.)</p> <p>Complete for each column of doctoral student (there may be multiple doctoral students to each supervisor.)</p> <p>Reconcile against supervisor's occupation.</p> <p>Reconcile against supervisor's name (and Orcid number if listed.)</p> <p>Check for false positives.</p> <p>Create new item if necessary for supervisor name and export into Wikidata schema.</p> <p>If the supervisor is no long employed by the University, and start and end dates can be identified, note on spreadsheet.</p> <p>This information can be added as <i>Qualifiers</i> in the employer statement.</p> <p>If you have start date for current University employees this can be added if desired.</p> <p>Import supervisor schema.</p> <p>Toolkit: Copy of Preferred Supervisor Schema (University of York)</p> <p>Export into Wikidata.</p> <p>Check through "contributions" in your personal wikidata account to see if all looks correct.</p> |
| Result | <p>All supervisors associated with the project have a Qid in Wikidata, and are also attached to the project statement.</p> <p>All supervisors have a statement showing the doctoral student associated to them.</p> <p>(There will be queries showing at this point saying that the doctoral student record needs to link to the doctoral supervisor record. This will be resolved in task 6.)</p> |

Task 6 - author (3)

| | |
|-----------|--|
| Objective | To create a statement in the author record (task 2), detailing the associated doctoral supervisor. |
| Process | <p>Return to the author/thesis list in OpenRefine. Reconcile Supervisor column(s). (All should now show a Qid following task 5 work.) Reconcile author list if necessary. Export author list into author schema ensuring inclusion of doctoral advisor statement. Toolkit: Copy of Preferred Creator Schema (University of York) Export into Wikidata. Check through “contributions” in your personal Wikidata account to see if all looks correct.</p> |
| Result | <p>The author entry should show statement for doctoral supervisor linking to appropriate individual. This will reflect the opposing link in the doctoral supervisors Qid linking to doctoral student. (NB: These links are person related - not thesis - so cannot appear in the thesis Qid.)</p> |

Review

| | |
|--------|--|
| Review | <p>Take time to check that all the information is showing correctly in both author and supervisor records.</p> <p>Author should link to thesis title, and thesis title to author. Author should link to supervisor (if listed), and supervisor should link back to author. Author, thesis title and any supervisor entry should all have project statement listed within them.</p> <p>Run query to check all appears correctly. https://w.wiki/5rwD (Insert appropriate project Qid into query in place of York number.)</p> |
|--------|--|

Annual uploads

When beginning with a Wikidata thesis project both LSE and York worked on author data, followed by thesis data and then supervisor data. When your initial thesis project has been completed you will want to establish an annual data upload once the thesis submission date for the previous year has passed. It is likely that for annual uploads there will only be a small number of new supervisor names. LSE has developed a refined workflow which has been more efficient now that we are working with ongoing annual uploads only.

LSE annual upload

1. Export and manipulate thesis data for upload.
2. Reconcile supervisors in OpenRefine and create Qids for any not already existing in Wikidata.
3. Reconcile authors in OpenRefine. Beware of false positive matches as it is highly likely all new authors will need Qid creation. Use LSE author schema (see appendix) which will create inverse relationships between authors and supervisors as part of upload.
4. Create Qids for theses. Use LSE thesis schema (see appendix) which will additionally populate Author Qids above with newly created thesis titles.

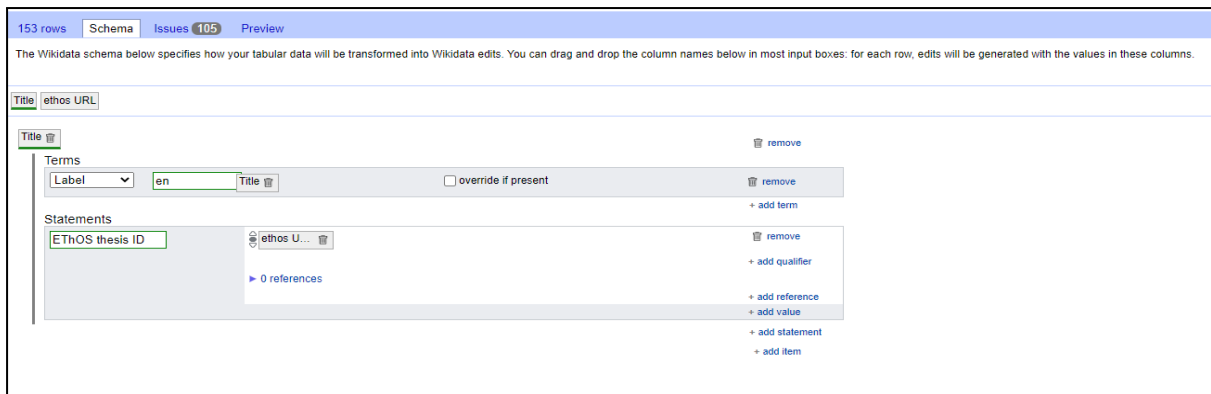
Additional options

Adding external identifiers to theses entries

Wikidata operates as an identifier hub, bringing together identifiers stored in external knowledge systems. Adding external identifiers to institutional theses creates links between these institutional and external sources making the content more verifiable for search engines and supporting discovery. We have been given some helpful feedback that if you add an OpenRefine column with the name of the WikiProject you can add that to the reconciliation process to be used as "on focus list of" property. This will weed out matches between thesis titles and subsequent published papers from it, or mismatches from generic thesis titles.

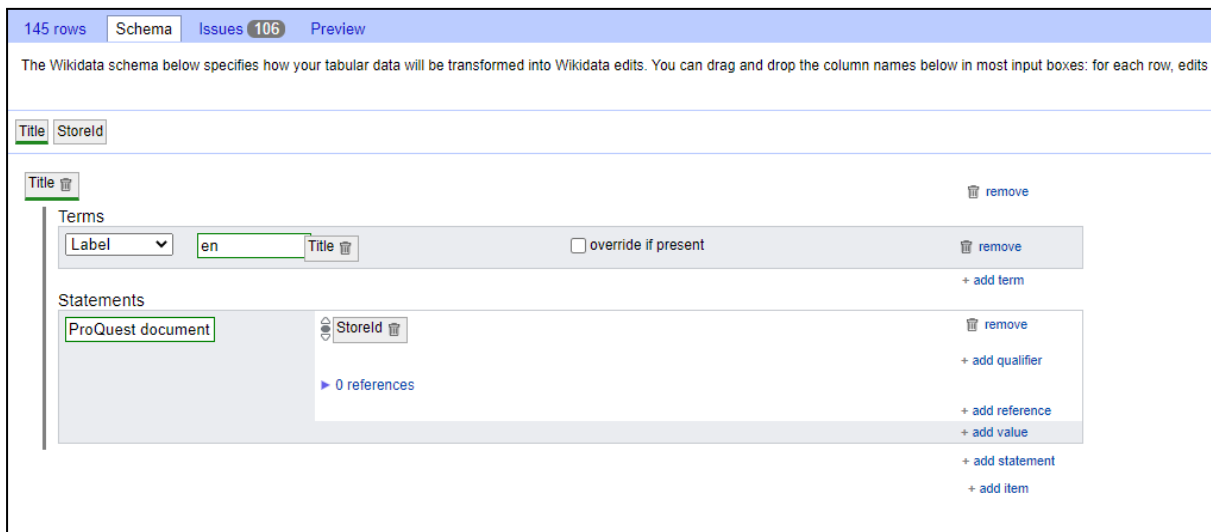
EThOS

- Using the British Library's EThOS dataset extract ids and titles for institutional thesis data into a new spreadsheet. <https://www.bl.uk/ethos-and-theses/re-using-ethos-data>
- Create a project in OpenRefine and reconcile title column with Wikidata.
- Use *Facets/Filter* to select matched titles.
- Create a brief schema, following the example below, to add Ethos URLs.
- Upload to Wikidata.



ProQuest

- Log into ProQuest Dissertations and Thesis Global.
- Use Advanced Search to find institution.
- Select doctoral dissertations and search. If not possible to extract all ids at once work by year and use date range in search.
- Export results and delete all columns except title and store id.
- ProQuest may have some duplicate title records which should be deleted from local spreadsheet. Save file, create OpenRefine project, reconcile by title, select matches and upload to Wikidata using schema example below.



DART Europe

Currently unable to download ids in bulk from DART. Once you have ids use schema to upload.

1012 matching rows (4094 total) Schema Issues 1013 Preview

The Wikidata schema below specifies how your tabular data will be transformed into Wikidata edits. You can drag and drop the column names below in most input boxes: for each row, edits will be generated

title record_id url

title remove

Terms

Label en title override if present remove + add term

Statements

DART-Europe thesis record_id remove + add qualifier + add reference + add value + add statement + add item

0 references

CORE

Go to <https://core.ac.uk/searchAssets/docs/#!/articles/searchArticlesBatch> to run an API search. Results are limited to 100 per page, so it is best to work by year.

- Use the POST batch operation for searching through articles and in query parameter enter

```
[
  {
    "query": "repositories.id:(xxx), year: yyyy",
    "page": 1,
    "pageSize": 50
  }
]
```

Enter repository id and edit year as appropriate

- PageSize = results per page. Although results limit is given as 100 we encountered errors going above 50, so reduce limit and repeated as necessary to extract data for each year.
- Page = which page of search results should be retrieved. Edit as appropriate.
- Enter API and click on *Try it out* to run search.
- Copy and paste response body to a JSON to CSV converter, such as <https://json-csv.com/> and download as a spreadsheet.
- Delete all fields except id and data title, save and create a project in OpenRefine.
- Repeat steps as with other identifiers and upload to Wikidata.

110 matching rows (112 total) Schema Issues 110 Preview

The Wikidata schema below specifies how your tabular data will be transformed into Wikidata edits. You can drag and drop the column names below in most input boxes: for each

data_id data_title

data_title remove

Terms

Label en data_title override if present remove + add term

Statements

CORE ID data_id remove + add qualifier + add reference + add value + add statement + add item

0 references

Wikipedia citations

Discovery of a thesis may be improved if it appears in Wikipedia. This query <https://w.wiki/5q97> will retrieve a list of institutional theses where the first results will be those where the authors have a Wikipedia page (change institution Qid and P953 URL). Use this list to edit Wikipedia to include the thesis title and a citation to it in the institutional repository. The cite thesis template https://en.wikipedia.org/wiki/Template:Cite_thesis can be used for this.

The New Zealand thesis project have kindly shared their 'notable pairs' query with us and given us a top tip for extending Wikipedia coverage. [https://w.wiki/5\\$vc](https://w.wiki/5$vc) By editing this query to contain your institutional project Qid you can obtain a list of names where both the author and their supervisor have a Wikipedia page. Wikipedia editors may use this information to support statements about notable person a being notable person b's student/advisor.

Exploration of data use

Querying Wikidata

SPARQL is a programming language for querying linked data stored on the web. It is essentially a set of commands that allow you to find exactly the data you want. By learning to use SPARQL you will be able to query the information stored in Wikidata, (plus any other data sources which use a SPARQL query service.)

To start with however the queries can look rather intimidating and technical, so for now take advantage of queries which others have created, and edit them to give you the information you are looking for.

In addition this will help you to understand how SPARQL queries are built. Look at queries run by:

New Zealand Thesis Project

<Wikidata:WikiProject NZThesisProject>

<https://www.wikidata.org/wiki/Q111645234>

London School of Economics Thesis Project

<Wikidata:WikiProject LSEThesisProject>

<https://www.wikidata.org/wiki/Q112895606>

University of York Thesis Project

<Wikidata:WikiProject University of York thesis project>

<https://www.wikidata.org/wiki/Q114588393>

By replacing the Qid listed for the project name with the Qid for your specific thesis project, you will be able to create queries returning information on the institutional doctoral theses which you have entered into Wikidata.

Measuring value and impact

View [league table of institutions with theses in Wikidata](#) to see how your institution compares with others who have loaded theses data to Wikidata.

The aim of a Wikidata thesis project is to support the impact of PhD research by increasing the reach of theses and engaging potential audiences. Measuring value and impact can be a challenge since putting the data in Wikidata is about enabling other sources to make use of that data, which might not always be immediately visible. Useful tools for assessing value and impact, however, can include:

- Repository downloads
- Traffic to repository
- Twitter mentions
- Increase of author and supervisor data in Wikimedia.

Measuring value and impact at LSE

LSE's Wikidata thesis project was experimental so an interim analysis was carried out part way through the work to assess whether the time being committed to the project was of value to the institution.

Interim analysis showed that LSETO downloads between February and May 2021 were 14% higher than the same period in 2020. By the end of the project, across the whole of LSETO downloads were 16% higher than 2020.

In 2019 only 1% of referrals to LSETO came from Wikipedia. In 2020, while we were working on the project this increased to 3%. In 2021 once all content had been added this increased to 13%.

Between February and May 2020 there were 38 mentions of [etheses.ac.uk](https://etheses.lse.ac.uk) on Twitter, which increased to 74 for the same time period in 2021. The same time period in 2022 showed 71 mentions indicating the initial increase is holding.

At the start of the project just 23% of LSE thesis authors and supervisors existed in Wikidata. Nearly 4000 Wikidata Qids for individuals were created during the project so that 100% of authors and supervisors are now represented in Wikidata. Just 7.7% currently have a Wikipedia page demonstrating that we have contributed a significant amount of unique data to Wikidata which can now be used by search engines and Wikimedia editors.

Outcomes of the project were communicated to:

- Library Leadership and Management teams and all Library staff
- Alumni office
- Alumni directly via Alumni newsletter
- PhD Academy
- LSE Research Bulletin.

Notification that LSETO data is included in Wikidata is given in:

- Data Protection Impact Assessment
- LSETO FAQs <http://etheses.lse.ac.uk/faq.html>
- LSETO submission form for PhD candidates.

Measuring value and impact at York

At York, we started experimenting with linking research theses to Wikidata in the autumn of 2022, and at time of writing have 463 items uploaded. Work is underway to upload the remaining 4775 theses to Wikidata following training of the Metadata team.

Further exploration of statistics from the WREO database is required to identify if upload to Wikidata has impacted on usage statistics. In addition we will review areas highlighted by LSE.

(16.02.23.)

Potential future actions

Ongoing Wikidata work at LSE

- Annual upload of theses once submission deadline for previous year has passed (Metadata team have already been trained to do this as part of business as usual).
- Addition of subject statements to theses to support subject visualisation <https://w.wiki/5aG5>.
- Extension of Wikidata work to other areas of the Library:
 - Initial work has begun with LSE Press and visualisation of article metadata added to Wikidata in Scholia
 - JIED <https://scholia.toolforge.org/venue/Q96715673>
 - LSE PPR <https://scholia.toolforge.org/venue/Q97011661>
 - Currently working on an experimental content to add metadata related to a collection of Suffrage Interviews held by the Library
 - https://www.wikidata.org/wiki/Wikidata:WikiProject_LSESuffrageInterviewsProject

Ongoing Wikidata work at University of York Library

- Create training materials and rollout upload of research theses to Wikidata to Metadata team as part of “business as usual” workload, (including annual upload.)
- Assess impact of upload of theses to Wikidata, (including feedback to White Rose Libraries of Leeds and Sheffield.)
- Currently planning White Rose Libraries event for research staff and students around **“The role of Wikidata within the research lifecycle.”** (Late Spring 2023.)
- Exploring the opportunities around adding the University of York Art Collection metadata to Wikidata, (including working in collaboration with students.)

Contact details

Helen Williams

Metadata Manager, Digital Scholarship and Innovation Group
LSE Library
Houghton Street
London WC2A 2AE

H.K.Williams@lse.ac.uk

ORCID: [0000-0003-1259-7097](https://orcid.org/0000-0003-1259-7097)

Twitter: @HelsKRW

Ruth Elder

Collections Management Specialist
Content and Open Research
Library, Archives and Learning Services
Student and Academic Services
University of York
Heslington York YO10 5DD

ruth.elder@york.ac.uk

Twitter: @ruthelder2

Further information

Clark, Jason A., Williams, Helen K.R., and Rossmann, Doralyn. 'Wikidata and Knowledge Graphs in Practice: Using Semantic SEO to Create Discoverable, Accessible, Machine-readable Definitions of the People, Places, and Services in Libraries and Archives'. 1 Jan. 2022 : 1 – 14.

<https://content.iospress.com/articles/information-services-and-use/isu220171>, accessed December 16 2022

Williams, Helen K. R., Wikidata: what? why? how? Catalogue and Index (203). pp. 28-35. ISSN 2399-9667 (2020)

http://eprints.lse.ac.uk/110987/1/Williams_wikidata_what_why_how_published.pdf, accessed November 30 2022

Williams, Helen K. R., LSE's adventures in Wikidata-land: tears and triumphs down the rabbit hole. Catalogue and Index, 206. pp. 2-6. ISSN 2399-9667 (2022)

<http://eprints.lse.ac.uk/114976/1/>, accessed November 30 2022

Appendix

LSE schema and data modelling

LSE author schema

URL Author Description Author description title pages date_awarded qualification_name DOI thesis_type Sup 1 Sup 2 Sup 3 Sup 4

Author remove

Terms

| | | | | |
|-------------|----|--------------------|--|---------------------|
| Label | en | Author | <input type="checkbox"/> override if present | remove |
| Description | en | Author description | <input type="checkbox"/> override if present | remove |

+ add term

Statements

| | | |
|------------------------|--|------------------------------|
| instance of | human | remove |
| | 0 references | + add qualifier |
| | | + add reference |
| | | + add value |
| academic degree | Doctor of Philosophy | remove |
| | 1 references | + add qualifier |
| | | + add reference |
| | | + add value |
| educated at | London School of Economics and Politic | remove |
| | 1 references | + add qualifier |
| | | + add reference |
| | | + add value |
| doctoral advisor | Sup 1 | remove |
| | 1 references | + add qualifier |
| | | + add reference |
| | Sup 2 | remove |
| | 1 references | + add qualifier |
| | | + add reference |
| | Sup 3 | remove |
| | 1 references | + add qualifier |
| | | + add reference |
| | Sup 4 | remove |
| | 1 references | + add qualifier |
| | | + add reference |
| | | + add value |
| on focus list of Wikin | LSEThesisProject | remove |
| | 0 references | + add qualifier |

Sup 1
 remove

Terms
no labels, descriptions or aliases added

[+ add term](#)

Statements

doctoral student

Author

remove
[+ add qualifier](#)
[+ add reference](#)
[+ add value](#)

[▶ 1 references](#)

[+ add statement](#)

Sup 2
 remove

Terms
no labels, descriptions or aliases added

[+ add term](#)

Statements

doctoral student

Author

remove
[+ add qualifier](#)
[+ add reference](#)

[▶ 1 references](#)

[+ add reference](#)

Sup 3
 remove

Terms
no labels, descriptions or aliases added

[+ add term](#)

Statements

doctoral student

Author

remove
[+ add qualifier](#)
[+ add reference](#)
[+ add value](#)

[▶ 1 references](#)

[+ add statement](#)

Sup 4
 remove

Terms
no labels, descriptions or aliases added

[+ add term](#)

Statements

doctoral student

Author

remove
[+ add qualifier](#)
[+ add reference](#)
[+ add value](#)

[▶ 1 references](#)

[+ add statement](#)
[+ add item](#)

LSE thesis schema

URL Author Author Qid Description Author description title pages date_awarded qualification_name DOI thesis_type Sup 1 Sup 2 Sup 3 Sup 4

type entity or drag reconciled column here

Terms

Label: en title override if present remove

Description: en Description override if present remove

+ add term

Statements

instance of: doctoral thesis remove

▶ 1 references + add qualifier

+ add reference

+ add value

title: en title remove

▶ 1 references + add qualifier

+ add reference

+ add value

language of work or: English remove

▶ 1 references + add qualifier

+ add reference

+ add value

publication date: date_awarded remove

▶ 1 references + add qualifier

+ add reference

+ add value

dissertation submitte: London School of Economics and Politi remove

doctoral advis: Sup 1 remove
 doctoral advis: Sup 2 remove
 doctoral advis: Sup 3 remove
 doctoral advis: Sup 4 remove

▶ 1 references + add qualifier

+ add reference

+ add value

copyright status: copyrighted remove

▶ 1 references + add qualifier

+ add reference

+ add value

online access status: open access remove

▶ 1 references + add qualifier

+ add reference

+ add value

number of pages: pages page remove

▶ 1 references + add qualifier

+ add reference

+ add value

| | | | | |
|------------------------|------------------|--------------|-----------------|--|
| full work available at | URL | 1 references | + add qualifier | remove configure |
| author | Author | 1 references | + add qualifier | + add reference + add value remove configure |
| DOI | DOI | 0 references | + add qualifier | + add reference + add value remove configure |
| on focus list of Wikin | LSEThesisProject | 0 references | + add qualifier | + add reference + add value + add statement remove configure |

| | | | | |
|-----------------|-----------------|--|-----------------|---|
| Author | | | | remove |
| Terms | | no labels, descriptions or aliases added | | |
| Statements | | | | + add term |
| academic thesis | title | dissertation su London School of E | + add qualifier | remove configure |
| | object has role | doctoral thesis | | remove |
| | | 1 references | | + add reference + add value + add statement |

LSE supervisor schema

Sup name Sup description Sup occupation Sup occ URL Sup ac degree Sup Ac degree URL Sup employer Sup employer URL Sup LC id Sup ResearchGate Google Scholar id on focus list

type entity or drag reconciled column here

Terms

Label en Sup name override if present

Description en Sup description override if present

Statements

instance of human

0 references

occupation Sup occupation

1 references

reference URL Sup occ URL
retrieved 2022-10-07

academic degree Sup ac degree

1 references

reference URL Sup 1 Ac degree URL
retrieved 2022-10-07

employer Sup employer

1 references

reference URL Sup employer URL
retrieved 2022-10-07

Library of Congress Sup LC id

0 references

ResearchGate profil Sup ResearchGate

0 references

on focus list of Wikin LSEThesisProject

0 references

Data modelling

Sample LSE data model:

| Statement | Metadata | Notes |
|---------------------------|---|--|
| label | Thesis title | |
| description | doctoral thesis by author name (or supply additional metadata, eg 2010 doctoral thesis by xx at University of xx) | Lower case other than name(s) |
| instance of | doctoral thesis | Use Qid Q187685 Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| Title | Thesis title | Include mandatory language qualifier En. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| author | Author name | Use Qid. If thesis name differs from Qid add stated as qualifier. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| dissertation submitted to | London School of Economics and Political Science | Add qualifiers for supervisors where metadata exists. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| language of work | English | Use Qid Q1860. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| publication date | Year | Year only. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| number of pages | Pages | Add if metadata in LSETO. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| copyright licence | | Add if metadata in LSETO. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| copyright status | copyrighted | Use Qid Q50423863. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |

| | | |
|---------------------------------------|---|---|
| full work available at URL | http://etheses.lse.ac.uk/xxxx | |
| online access status | open access | Use Qid Q232932. Reference URL: http://etheses.lse.ac.uk/xxxx Retrieved date: YYYY-MM-DD |
| DOI | | Add if metadata in LSETO. Letters must be upper case |
| CORE ID | | https://core.ac.uk/ Extract number in URL which Wikidata will hyperlink |
| DART-Europe thesis ID | | http://www.dart-europe.eu/basic-search.php Extract number in URL which Wikidata will hyperlink |
| EThOS thesis ID | uk.bl.ethos.xxxxx | https://ethos.bl.uk/ keep id preface to automatically create hyperlink |
| ProQuest document ID | | Use institutional log in. Extract first number in URL which will be automatically hyperlinked in Wikidata |
| on focus list of Wikimedia project | LSEThesisProject | Use Qid Q112895606 |

Data mapping

Sample LSE data mapping

| Wikidata statement | Source data | Notes |
|--------------------|---|---|
| Label | LSETO: Title | Sentence case |
| Description | LSETO: Creator given name LSETO: Creator family name | Merge creator names in Excel Merge full name with 'doctoral thesis by' |
| instance of | Wikidata: Q187685 | doctoral thesis |
| Title | LSETO: Title | + mandatory language qualifier En |

| | | |
|------------------------------------|------------------------------|---|
| author stated as qualifier | Wikidata: Author Qids | Reconcile LSETO names data with Wikidata in OpenRefine Where LSETO name differs from Qid name format |
| dissertation submitted to | Wikidata: Q174570 | LSE |
| language of work | Wikidata: Q1860 | English |
| publication date | LSETO: Date | Year only |
| number of pages | LSETO: Number of pages | |
| copyright licence | LSETO: Licence | |
| copyright status | Wikidata: Q50423863 | copyrighted |
| full work available at URL | LSETO: URL | |
| online access status | Wikidata: Q232932 | open access |
| DOI | LSETO: identification number | capitalise letters |
| CORE ID | CORE | https://core.ac.uk/ |
| DART-EUROPE thesis ID | DART EUROPE | http://www.dart-europe.eu/basic-search.php |
| EThOS thesis ID | EThOS | https://ethos.bl.uk/ |
| ProQuest document ID | ProQuest | ProQuest institutional log in |
| on focus list of Wikimedia project | Wikidata: Q112895606 | |

Note that LSE's data modelling and mapping does not contain subject data. Our controlled repository subject metadata uses very broad headings which did not appear to map usefully to Wikidata. Any keyword data is very granular and would be time consuming to map to Wikidata manually. LSE has not included subject data as part of automated uploads, but is experimenting with manual addition of subject metadata to further understand the data and consider automated options. Subjects added so far can be seen at <https://w.wiki/5aG5>

University of York schema and data modelling

Example of York author schema (JSON text)

(Import text to OpenRefine)

```
{
  "entityEdits": {
    "type": "wbitemeditexpr",
    "subject": {
      "type": "wbentityvariable",
      "columnName": "creators_name",
      "nameDescs": {
        "name_type": "LABEL_IF_NEW",
        "value": {
          "type": "wbmonolingualexpr",
          "language": {
            "type": "wblanguageconstant",
            "id": "en",
            "label": "en",
            "value": {
              "type": "wbstringvariable",
              "columnName": "creators_name"
            }
          },
          "name_type": "DESCRIPTION_IF_NEW",
          "value": {
            "type": "wbmonolingualexpr",
            "language": {
              "type": "wblanguageconstant",
              "id": "en",
              "label": "en",
              "value": {
                "type": "wbstringconstant",
                "value": "successful doctoral candidate at the University of York"
              }
            }
          },
          "statementGroups": {
            "property": {
              "type": "wbpropconstant",
              "pid": "P31",
              "label": "instance of",
              "datatype": "wikibase-item",
              "statements": {
                "value": {
                  "type": "wbentityidvalueconstant",
                  "id": "Q5",
                  "label": "human",
                  "qualifiers": [],
                  "references": {
                    "snaks": {
                      "prop": {
                        "type": "wbpropconstant",
                        "pid": "P953",
                        "label": "full work available at URL",
                        "datatype": "url",
                        "value": {
                          "type": "wbstringvariable",
                          "columnName": "url"
                        }
                      },
                      "mergingStrategy": {
                        "type": "snak",
                        "valueMatcher": {
                          "type": "lax"
                        },
                        "mode": "add_or_merge"
                      },
                      "property": {
                        "type": "wbpropconstant",
                        "pid": "P69",
                        "label": "educated at",
                        "datatype": "wikibase-item",
                        "statements": {
                          "value": {
                            "type": "wbentityidvalueconstant",
                            "id": "Q967165",
                            "label": "University of York",
                            "qualifiers": [],
                            "references": {
                              "snaks": {
                                "prop": {
                                  "type": "wbpropconstant",
                                  "pid": "P953",
                                  "label": "full work available at URL",
                                  "datatype": "url",
                                  "value": {
                                    "type": "wbstringvariable",
                                    "columnName": "url"
                                  }
                                },
                                "mergingStrategy": {
                                  "type": "snak",
                                  "valueMatcher": {
                                    "type": "lax"
                                  },
                                  "mode": "add_or_merge"
                                },
                                "property": {
                                  "type": "wbpropconstant",
                                  "pid": "P512",
                                  "label": "academic degree",
                                  "datatype": "wikibase-item",
                                  "statements": {
                                    "value": {
                                      "type": "wbentityidvalueconstant",
                                      "id": "Q752297",
                                      "label": "Doctor of Philosophy",
                                      "qualifiers": [],
                                      "references": {
                                        "snaks": {
                                          "prop": {
                                            "type": "wbpropconstant",
                                            "pid": "P953",
                                            "label": "full work available at URL",
                                            "datatype": "url",
                                            "value": {
                                              "type": "wbstringvariable",
                                              "columnName": "url"
                                            }
                                          },
                                          "mergingStrategy": {
                                            "type": "snak",
                                            "valueMatcher": {
                                              "type": "lax"
                                            },
                                            "mode": "add_or_merge"
                                          },
                                          "property": {
                                            "type": "wbpropconstant",
                                            "pid": "P1026",
                                            "label": "academic thesis",
                                            "datatype": "wikibase-item",
                                            "statements": {
                                              "value": {
                                                "type": "wbentityvariable",
                                                "columnName": "title",
                                                "qualifiers": [],
                                                "references": {
                                                  "snaks": {
                                                    "prop": {
                                                      "type": "wbpropconstant",
                                                      "pid": "P953",
                                                      "label": "full work available at URL",
                                                      "datatype": "url",
                                                      "value": {
                                                        "type": "wbstringvariable",
                                                        "columnName": "url"
                                                      }
                                                    },
                                                    "snaks": {
                                                      "prop": {
                                                        "type": "wbpropconstant",
                                                        "pid": "P813",
                                                        "label": "retrieved",
                                                        "datatype": "time",
                                                        "value": {
                                                          "type": "wbdateconstant",
                                                          "value": "2022-11-10"
                                                        }
                                                      }
                                                    },
                                                    "mergingStrategy": {
                                                      "type": "snak",
                                                      "valueMatcher": {
                                                        "type": "lax"
                                                      },
                                                      "mode": "add_or_merge"
                                                    },
                                                    "property": {
                                                      "type": "wbpropconstant",
                                                      "pid": "P184",
                                                      "label": "doctoral advisor",
                                                      "datatype": "wikibase-item",
                                                      "statements": {
                                                        "value": {
                                                          "type": "wbentityvariable",
                                                          "columnName": "supervisor 1",
                                                          "qualifiers": [],
                                                          "references": {
                                                            "snaks": {
                                                              "prop": {
                                                                "type": "wbpropconstant",
                                                                "pid": "P953",
                                                                "label": "full work available at URL",
                                                                "datatype": "url",
                                                                "value": {
                                                                  "type": "wbstringvariable",
                                                                  "columnName": "url"
                                                                }
                                                              },
                                                              "mergingStrategy": {
                                                                "type": "snak",
                                                                "valueMatcher": {
                                                                  "type": "lax"
                                                                },
                                                                "mode": "add_or_merge"
                                                              },
                                                              "property": {
                                                                "type": "wbpropconstant",
                                                                "pid": "P953",
                                                                "label": "full work available at URL",
                                                                "datatype": "url",
                                                                "value": {
                                                                  "type": "wbstringvariable",
                                                                  "columnName": "url"
                                                                }
                                                              },
                                                              "mergingStrategy": {
                                                                "type": "snak",
                                                                "valueMatcher": {
                                                                  "type": "lax"
                                                                },
                                                                "mode": "add_or_merge"
                                                              },
                                                              "property": {
                                                                "type": "wbpropconstant",
                                                                "pid": "P5008",
                                                                "label": "on focus list of Wikimedia project",
                                                                "datatype": "wikibase-item",
                                                                "statements": {
                                                                  "value": {
                                                                    "type": "wbentityidvalueconstant",
                                                                    "id": "Q114588393",
                                                                    "label": "UniversityofYorkThesisProject",
                                                                    "qualifiers": [],
                                                                    "references": [],
                                                                    "mergingStrategy": {
                                                                      "type": "snak",
                                                                      "valueMatcher": {
                                                                        "type": "lax"
                                                                      },
                                                                      "mode": "add_or_merge"
                                                                    },
                                                                    "property": {
                                                                      "type": "wbpropconstant",
                                                                      "pid": "P496",
                                                                      "label": "ORCID ID",
                                                                      "datatype": "external-id",
                                                                      "statements": {
                                                                        "value": {
                                                                          "type": "wbstringvariable",
                                                                          "columnName": "creators_orcid",
                                                                          "qualifiers": [],
                                                                          "references": [],
                                                                          "mergingStrategy": {
                                                                            "type": "snak",
                                                                            "valueMatcher": {
                                                                              "type": "lax"
                                                                            },
                                                                            "mode": "add_or_merge"
                                                                          },
                                                                          "property": {
                                                                            "type": "wbpropconstant",
                                                                            "pid": "P106",
                                                                            "label": "occupation",
                                                                            "datatype": "wikibase-item",
                                                                            "state":

```

```
ments":{"value":{"type":"wbentityidvalueconstant","id":"Q1650915","label":"researcher"},"qualifiers":[],"references":[],"mergingStrategy":{"type":"snak","valueMatcher":{"type":"lax"},"mode":"add_or_merge"}}},"siteIri":"http://www.wikidata.org/entity/","entityTypeSiteIri":{"item":"http://www.wikidata.org/entity/","property":"http://www.wikidata.org/entity/"},"mediaWikiApiEndpoint":"https://www.wikidata.org/w/api.php"}
```

University of York author schema
(as displayed in OpenRefine preview screen)

The screenshot displays the OpenRefine preview screen for the University of York author schema. At the top, a header row contains the following columns: `uri`, `title`, `date`, `thesis_type`, `creators_name`, `creators_orcid`, `supervisor`, `institution`, and `id_number`. The `creators_name` column is selected, and its details are shown below.

creators_name

Terms

- Label: `en` `creators_name` override if present
- Description: `en` successful doctoral candidate at the override if present

Statements

- instance of: `human` [1 references](#)
- educated at: `University of York` [1 references](#)
- academic degree: `Doctor of Philosophy` [1 references](#)
- academic thesis: `title` [2 references](#)
- doctoral advisor: `supervisor 1`
- `supervisor 2` [1 references](#)
- on focus list of Wikin: `UniversityofYorkThesisProject` [0 references](#)
- ORCID ID: `creators_orcid` [0 references](#)
- occupation: `researcher` [0 references](#)

University of York theses schema

| uri | title | date | thesis_type | creators_name | creators_orcid | supervisor | institution | id_number |
|--------------------------------|-------|-----------------|--|---------------|----------------|------------|-------------|-----------|
| title | | | | | | | | |
| Terms | | | | | | | | |
| Label | en | title | <input type="checkbox"/> override if present | | | | | |
| Description | en | doctoral thesis | <input type="checkbox"/> override if present | | | | | |
| Statements | | | | | | | | |
| instance of | | | doctoral thesis | | | | | |
| ▶ 1 references | | | | | | | | |
| title | en | title | | | | | | |
| ▶ 1 references | | | | | | | | |
| author | | | creators_name | | | | | |
| ▶ 1 references | | | | | | | | |
| dissertation submitte | | | University of York | | | | | |
| ▶ 1 references | | | | | | | | |
| language of work or | | | English | | | | | |
| publication date | | | date | | | | | |
| ▶ 1 references | | | | | | | | |
| full work available at | | | uri | | | | | |
| ▶ 1 references | | | | | | | | |
| on focus list of Wikin | | | UniversityofYorkThesisProject | | | | | |
| ▶ 0 references | | | | | | | | |
| ETHOS thesis ID | | | id_number | | | | | |
| ▶ 0 references | | | | | | | | |

University of York supervisor schema

The screenshot displays a web-based schema editor. At the top, there are navigation tabs for different schema elements: Step 1 name, Doctoral Student 1, Doctoral Student 2, Doctoral Student 3, Doctoral Student 4, Doctoral Student 5, Doctoral Student 6, Doctoral Student 7, Doctoral Student 8, Doctoral Student 9, Doctoral Student 10, Doctoral Student 11, Doctoral Student 12, Doctoral Student 13, Doctoral Student 14, Doctoral Student 15, Doctoral Student 16, Doctoral Student 17, Doctoral Student 18, Doctoral Student 19, Doctoral Student 20, Doctoral Student 21, Doctoral Student 22, Doctoral Student 23, Doctoral Student 24, Doctoral Student 25, Doctoral Student 26, Doctoral Student 27, Doctoral Student 28, Doctoral Student 29, Doctoral Student 30, Doctoral Student 31, Doctoral Student 32, Doctoral Student 33, Doctoral Student 34, Doctoral Student 35, Doctoral Student 36, Doctoral Student 37, Doctoral Student 38, Doctoral Student 39, Doctoral Student 40, Doctoral Student 41, Doctoral Student 42, Doctoral Student 43, Doctoral Student 44, Doctoral Student 45, Doctoral Student 46, Doctoral Student 47, Doctoral Student 48, Doctoral Student 49, Doctoral Student 50, Doctoral Student 51, Doctoral Student 52, Doctoral Student 53, Doctoral Student 54, Doctoral Student 55, Doctoral Student 56, Doctoral Student 57, Doctoral Student 58, Doctoral Student 59, Doctoral Student 60, Doctoral Student 61, Doctoral Student 62, Doctoral Student 63, Doctoral Student 64, Doctoral Student 65, Doctoral Student 66, Doctoral Student 67, Doctoral Student 68, Doctoral Student 69, Doctoral Student 70, Doctoral Student 71, Doctoral Student 72, Doctoral Student 73, Doctoral Student 74, Doctoral Student 75, Doctoral Student 76, Doctoral Student 77, Doctoral Student 78, Doctoral Student 79, Doctoral Student 80, Doctoral Student 81, Doctoral Student 82, Doctoral Student 83, Doctoral Student 84, Doctoral Student 85, Doctoral Student 86, Doctoral Student 87, Doctoral Student 88, Doctoral Student 89, Doctoral Student 90, Doctoral Student 91, Doctoral Student 92, Doctoral Student 93, Doctoral Student 94, Doctoral Student 95, Doctoral Student 96, Doctoral Student 97, Doctoral Student 98, Doctoral Student 99, Doctoral Student 100.

The main content area shows a list of properties with their values and actions:

- Terms:** Description: en:doctoral supervisor; Label: en:Step 1 name
- Statements:**
 - instance of: Human
 - employer: University of York
 - on focus list of Wikimedia project: UniversityofYorkThe sisProject
 - reference URL: URL to occupier biography
- Academic degree:** Doctor of Philosophy
- Doctoral student:** Doctoral Student 1, Doctoral Student 2, Doctoral Student 3
- Occupation:** Step 1 description

| | | | Example |
|------------------------|------------------------------------|--|---|
| Supervisor name | | Jo Smith | |
| Terms | | | |
| | Description | Doctoral Supervisor | |
| | Label | (Supervisor name) | Jo Smith |
| Statements | Instance of | | Human |
| | Employer | | University of York |
| | On focus list of Wikimedia project | (Name of project as appears in Wikidata) | UniversityofYorkThe sisProject |
| | Reference URL | (URL to biography) | Example: https://www.york.ac.uk/history/people/jenner/ |
| | Academic degree | (Degree level) | Doctor of |

| | | | |
|--|--------------------|---------------|-----------------|
| | | | Philosophy |
| | Doctoral Student 1 | (Name) | XXXX |
| | Doctoral Student 2 | (Name) | YYYY |
| | Doctoral Student 3 | (Name) | ZZZZ |
| | Doctoral Student 4 | (Name) | AAAA |
| | Occupation | (Description) | Senior Lecturer |

Data modelling

Sample York data model

| Statement | Metadata | Notes |
|---------------------------|-------------------------------|--|
| Label | Thesis title | |
| Description | doctoral thesis | Lower case Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |
| instance of | doctoral thesis | Use Q187685 Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |
| Title | Thesis title | Include mandatory language qualifier En Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |
| author | Author name (if existing Qid) | Use Qid. If thesis name differs from Qid add stated as qualifier Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |
| dissertation submitted to | University of York | Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |
| language of work | English | Use Q1860 |
| publication date | Year | Year only Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |

| | | |
|------------------------------------|---|--|
| Full work available at URL | https://etheses.whiterose.ac.uk/XXXX | Add Reference in form of "full work available at URL" https://etheses.whiterose.ac.uk/XXXXX/ |
| EThOS thesis ID | uk.bl.ethos.xxxxx | https://ethos.bl.uk/ keep id preface to automatically create hyperlink |
| On focus list of Wikimedia project | UniversityofYorkThesisProject | Use Q114588393 |

Additional appendix (added July 2023)

(University of York Library)

Reconciling against Wikidata in OpenRefine and export to Wikidata

An Open Refine update (late Spring 2023), has led to an adaptation of the process used at the University of York library.

This is the updated version (as of July 2023.) As ever - it comes with the disclaimer that the process listed is one of many ways to approach the task, and is not necessarily the best or the most efficient, but is a guide to how it can be approached and is offered with best intent.

Complete upload of data set to OpenRefine and edit as described previously. Ensure that appropriate schema have been created or imported into OpenRefine.

At York we are currently using:

(see end of this supplementary document for additional information.)

- **Correct title - no author - with correct references**
- **Correct author schema - no supervisor entry- correct references**
- **Correct title schema - with author - correct references**

Thesis title

| | |
|-----------|--|
| Objective | To create Qids for the thesis title. |
| Process | <p>Reconcile title against doctoral thesis Q187685. (In the vast majority of cases there will not be a match.)</p> <p>Identify those titles which need a new item creating for them (to give them a Qid in Wikidata.)</p> <p>To identify these select <i>Reconcile/Facets/By judgment/true or false</i>. Create a new item for each thesis without a match.</p> <p><i>Reconcile/action/create new items for each cell</i>.</p> <p>Select <i>Extensions Wikidata</i>, then <i>Import schema</i>.</p> <p>Export into Wikidata schema page:</p> <ul style="list-style-type: none"> • Correct title - no author - with correct references <p>Make sure publication year (date) and project statement are included. Check in <i>Preview</i> and make any alterations or corrections as needed. Export to <i>Wikidata Upload</i>.</p> <p>Export into Wikidata.</p> <p>Check through “contributions” in your personal wikidata account to see if all looks correct.</p> |
| Result | All thesis titles will now have a Qid and a project statement. |

Thesis author

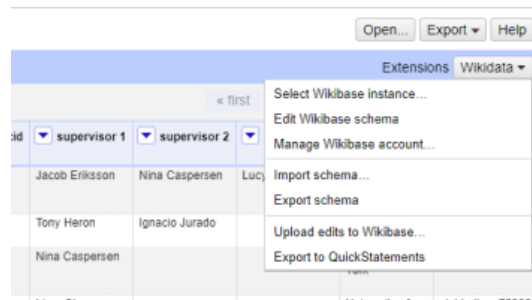
| | |
|-----------|--|
| Objective | <p>To create Q identities (Qids) for authors who do not already have a listing in Wikidata.</p> <p>To give all authors a University of York Thesis project statement, to make them searchable as part of the project.</p> <p>To link the author record to the thesis record.</p> |
|-----------|--|

Process

Select authors name column and select *Reconcile/Start Reconciling*. Click on *Wikidata reconci link (en)*.
Reconcile against Q5 (human.)
Select second method of reconciling by using author Orcid number. Select *Start reconciling*.
(The larger the set, the slower the process. This may take a few minutes.) After reconciliation there will be some names reconciled against names/Qids which are already in Wikidata (P50s). The majority will probably not.
To identify these select *Reconcile/Facets/By judgment type/true or false*. Look at those reconciled against existing Qids (highlighted in blue and linking to Wikidata entry.)
Check these for false positives.

For those that are not already listed in Wikidata, create new items for author names.

Select these items through *Reconcile/Facets/By judgment/true or false*. Then *Reconcile/action/create new items for each cell*.



Select all items in the set.
Select *Extensions Wikidata*, then *Import schema*.
Use schema: **Correct author schema- no supervisor entry - correct references**.
Publish to Wikidata.

Result

There is an entry (Q number) in Wikidata for each author.
Each author record is associated with the University of York Thesis Project.

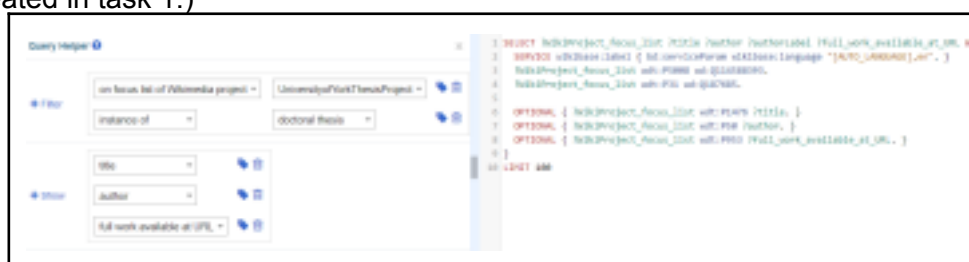
| | |
|-----------|--|
| Objective | To link the title to the author record. |
| Process | <p>Select <i>Extensions Wikidata</i>, then <i>Import schema</i>. Export into Wikidata schema page:import Correct title schema - with author - correct references Make sure publication year (date) and project statement are included. Check in <i>Preview</i> and make any alterations or corrections as needed. Export to <i>Wikidata Upload</i>. Export into Wikidata. Check through “contributions” in your personal wikidata account to see if all looks correct.</p> |
| Result | <p>All thesis titles will have an author statement.</p> <p>Each author will have an entry in Wikidata, and this will link to the thesis entry.</p> <p>The entry for each author and each thesis will have a statement in it associating it with the University of York Thesis Project.</p> |

Review

Review Take time to check that all the information is showing correctly in both author and thesis records.

Run query to check all appears correctly in relation to project statement. ([Link to query \(UniversityofYorkThesisProject\)](#))

Note: replace Qid for UniversityofYorkThesisProject with the Qid for your project statement (created in task 1.)



Mapping of Wikidata - Thesis entry (not inc author)

Last updated: 1211/270623

| | Import /type | Metadata | Header in WREO export | Notes |
|---------------------------|--------------|--|-----------------------|--|
| Item | | title Import title column header from export | title | |
| label | | title Import title column header from export | title | Lower case (other than name(s)) Include language qualifier (En.) |
| description | | doctoral thesis (type doctoral thesis into schema) | | Include language qualifier (En.) |
| Statements | | Metadata | | Notes |
| Instance of | | doctoral thesis (type doctoral thesis into schema) | | Use Qid 187685 Add Reference to entry: Reference URL (P854) + import “url” column header from export Retrieved: YYYY-MM-DD (include dashes) Add todays date. Select copy option |
| Title | | title Import title column header from export | “title” | Include mandatory language qualifier En. Select paste to duplicated reference URL and retrieved information. |
| Dissertation submitted to | | University of York (type University of York into schema) | | Use Q967165 (University of York Qid) Select paste to duplicated reference URL and retrieved information. |
| Language of work | | English (type English into schema) | | Use Qid Q1860 (English West Germanic Language Qid) Select paste to duplicated reference URL and retrieved information. |
| Publication date | | Date Import date column header from | “date” | Year only. Select paste to duplicated |

| | | | | |
|--|--|---|-------------|--|
| | | export | | reference URL and retrieved information. |
| Full work available at URL | | Import url column header from export | "url" | https://etheses.whiterose.ac.uk XXX Use P953 |
| EThOS thesis ID | | Import id_number column header from export | "id_number" | https://ethos.bl.uk/ keep id preface to automatically create hyperlink |
| On focus list of Wikimedia project (Use P5008) | | UniversityofYorkThesisProject (type UniversityofYorkThesisProject into schema) | | Use Qid Q114588393 (Qid for project) |

Mapping of Wikidata - Author entry (not inc doctoral advisor statement)

Last updated: 1005/270623

| | Import /type | Metadata | Header in WREO export | Notes |
|-------------------|--------------|--|-----------------------|--------------------------------------|
| Item | | author Import Creator's name column header from export | Creator's name | |
| label | | author Import Creator's name column header from export | Creator's name | Include language qualifier (En.) |
| description | | successful doctoral candidate at the University of York (type successful doctoral candidate at the University of York into schema) | | Include language qualifier (En.) |
| Statements | | Metadata | | Notes |
| Instance of | | human (type human into schema) | | Use Qid 5 |
| occupation | | researcher type researcher into schema | | Use Q1650915 |
| Educated at | | University of York (type University of York into schema) | | Use Q967165 (University of York Qid) |

| | | | | |
|--|--|--|-----------------|---|
| | | | | <p>Add Reference to entry: Reference URL (P854) + import “url” column header from export</p> <p>Retrieved: YYYY-MM-DD (include dashes) (Add today's date here.) Select copy option</p> |
| Academic degree | | Doctor of Philosophy (type Doctor of Philosophy into schema) | | Select paste to duplicated reference URL and retrieved information. |
| Academic thesis | | Title Import title column header from export | title | Select paste to duplicated reference URL and retrieved information. |
| On focus list of Wikimedia project (Use P5008) | | UniversityofYorkThesisProject (type UniversityofYorkThesisProject into schema) | | Use Qid Q114588393 (Qid for project) |
| Identifier | | Metadata | | Notes |
| ORCID iD | | Orcid iD Import “ creator's ORCID ” column header from export | creator's ORCID | |
| | | | | |

Mapping of Wikidata - Thesis entry (inc author)

Last updated: 1211/270623

| | Import /type | Metadata | Header in WREO export | Notes |
|-------------------|--------------|--|-----------------------|---|
| Item | | title Import title column header from export | title | |
| label | | title Import title column header from export | title | Lower case (other than name(s)) Include language qualifier (En.) |
| description | | doctoral thesis (type doctoral thesis into schema) | | Include language qualifier (En.) |
| Statements | | Metadata | | Notes |

| | | | | |
|---|--|---|---------------------|--|
| Instance of | | doctoral thesis (type doctoral thesis into schema) | | Use Qid 187685 Add Reference to entry: Reference URL (P854) + import “ url ” column header from export Retrieved: YYYY-MM-DD (include dashes) Add todays date. Select copy option |
| Title | | title Import title column header from export | “title” | Include mandatory language qualifier En. Select paste to duplicated reference URL and retrieved information. |
| Author | | (author name) Import creator’s name column header from export | “creator’s name” | Use individuals Qid (once created) |
| Dissertation submitted to | | University of York (type University of York into schema) | | Use Q967165 (University of York Qid) Select paste to duplicated reference URL and retrieved information. |
| Language of work | | English (type English into schema) | | Use Qid Q1860 (English West Germanic Language Qid) Select paste to duplicated reference URL and retrieved information. |
| Publication date | | Date Import date column header from export | “date” | Year only. Select paste to duplicated reference URL and retrieved information. |
| Full work available at URL | | Import url column header from export | “url” | https://etheses.whiterose.ac.uk XXX Use P953 |
| ETHOS thesis ID | | Import id_number column header from export | “id_number” | https://ethos.bl.uk/ keep id preface to automatically create hyperlink |
| On focus list of Wikimedia project (Use P5008) | | UniversityofYorkThesisProject (type UniversityofYorkThesisProject into schema) | | Use Qid Q114588393 (Qid for project) |

(End of document)