Wikidata thesis toolkit

Contents page

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, resulting in the creation of the University of York Library article by June 2022 (yes - it took a while.)
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.

[1] Dempsey, Lorcan, Outside-in and inside-out, (2010) https://www.lorcandempsey.net/orweblog/outside-in-and-inside-out/, accessed December 16, 2022.

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.
- 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

[2] Dempsey, Lorcan, Outside-in and inside-out, (2010) https://www.lorcandempsey.net/orweblog/outside-in-and-inside-out/, accessed December 16, 2022.

> "*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."*

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 ecosystem[4]...

…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&I199Evans_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

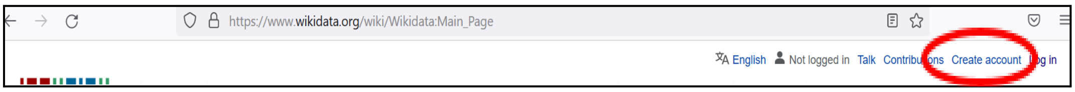
- Poulter, Martin and Nick Sheppard, Wikimedia and universities: contributing to the global commons in the Age of Disinformation. Insights 33 (1): 14, (2020)
  https://insights.uksg.org/articles/10.1629/uksg.509/, accessed March 24 2022

- 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/1QQn1jS89TWGiLcXOpbRGXCcDVxWKSgEqg6Pj8zXxt9w/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

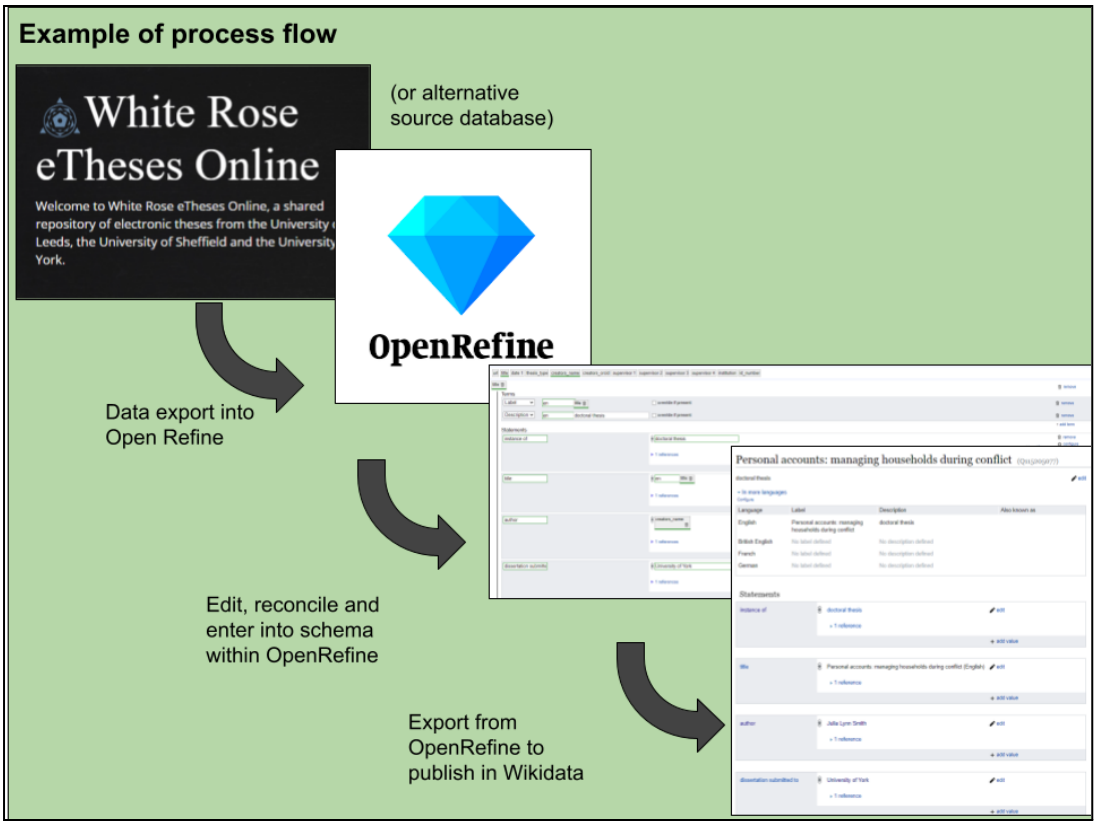
- University of York Library https://www.wikidata.org/wiki/Q24753366
- White Rose University Press https://www.wikidata.org/wiki/Q74433041
- White Rose Etheses Online https://www.wikidata.org/wiki/Q24779464

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



Example of process flow

White Rose eTheses Online

Welcome to White Rose eTheses Online, a shared repository of electronic theses from the University of Leeds, the University of Sheffield and the University of York.

(or alternative source database)

OpenRefine

Data export into Open Refine

Edit, reconcile and enter into schema within OpenRefine

Export from OpenRefine to publish in Wikidata

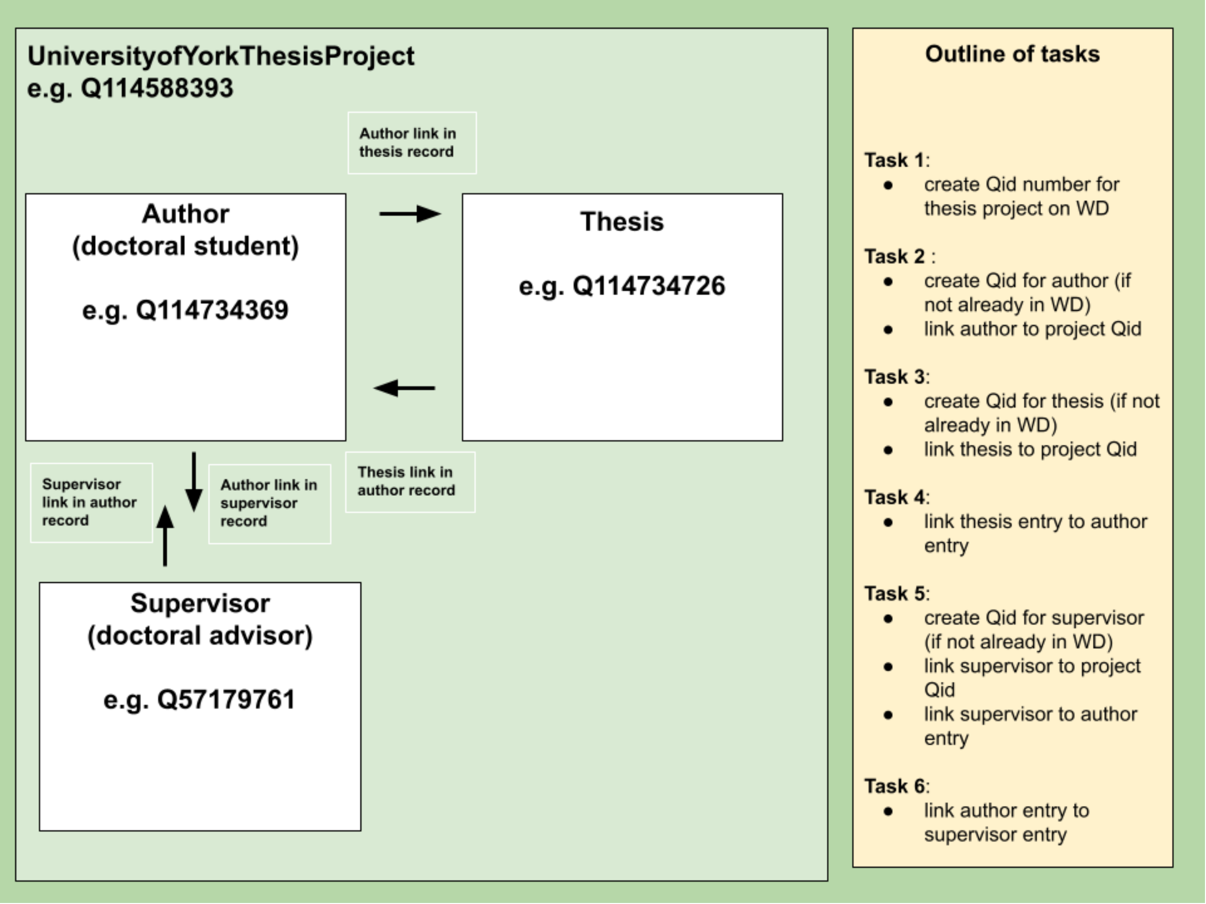
Personal accounts: managing households during conflict

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.)


UniversityofYorkThesisProject
e.g. Q114588393
Author link in thesis record
Author
(doctoral student)
e.g. Q114734369
Thesis
e.g. Q114734726
Supervisor link in author record
Author link in supervisor record
Thesis link in author record
Supervisor
(doctoral advisor)
e.g. Q57179761

Outline of tasks

Task 1:
- create Qid number for thesis project on WD

Task 2 :
- create Qid for author (if not already in WD)
- link author to project Qid

Task 3:
- create Qid for thesis (if not already in WD)
- link thesis to project Qid

Task 4:
- link thesis entry to author entry

Task 5:
- create Qid for supervisor (if not already in WD)
- link supervisor to project Qid
- link supervisor to author entry

Task 6:
- link author entry to supervisor entry

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

	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 **UniversityofYorkThesisProject** https://www.wikidata.org/wiki/Q114588393.

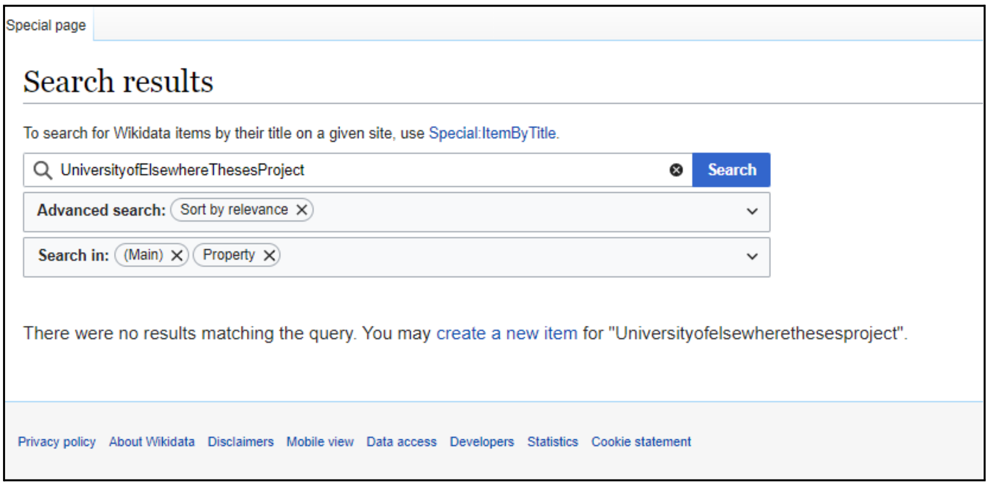
Preparation

Task 1 - project identifier

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 to represent the project name
Process	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.  Special page Search results To search for Wikidata items by their title on a given site, use Special:ItemByTitle. UniversityofElsewhereThesesProject Advanced search: Sort by relevance Search in: (Main) Property There were no results matching the query. You may create a new item for "Universityofelsewherethesesproject". Search Privacy policy About Wikidata Disclaimers Mobile view Data access Developers Statistics Cookie statement 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.
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

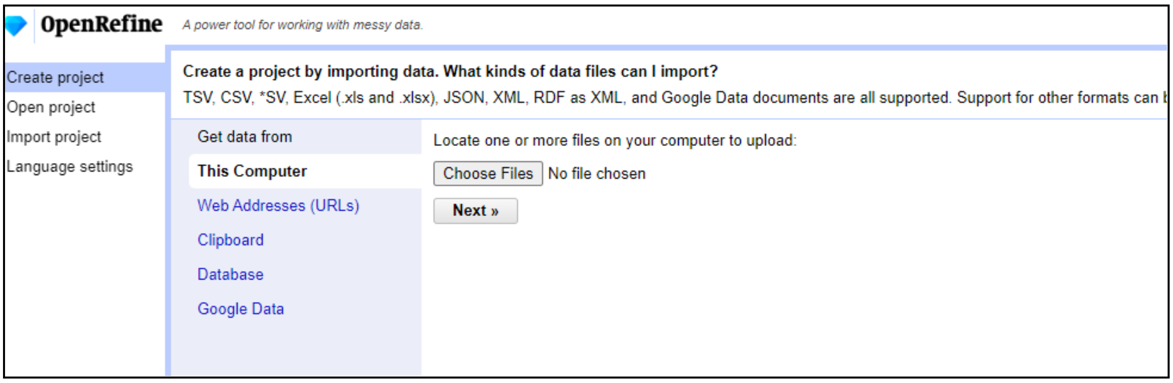
"LSE Theses Online may incorporate metadata or documentation into public access catalogues for the Work(s) and the Author(s), including, but not limited to, Wikidata, CORE, DART-Europe and ProQuest Dissertations and Theses. A citation/citations to the Work(s) will always remain visible in the repository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

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.



Create a project by importing data. What kinds of data files can I import?

TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for other formats can b

Locate one or more files on your computer to upload:

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 (customised)

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.

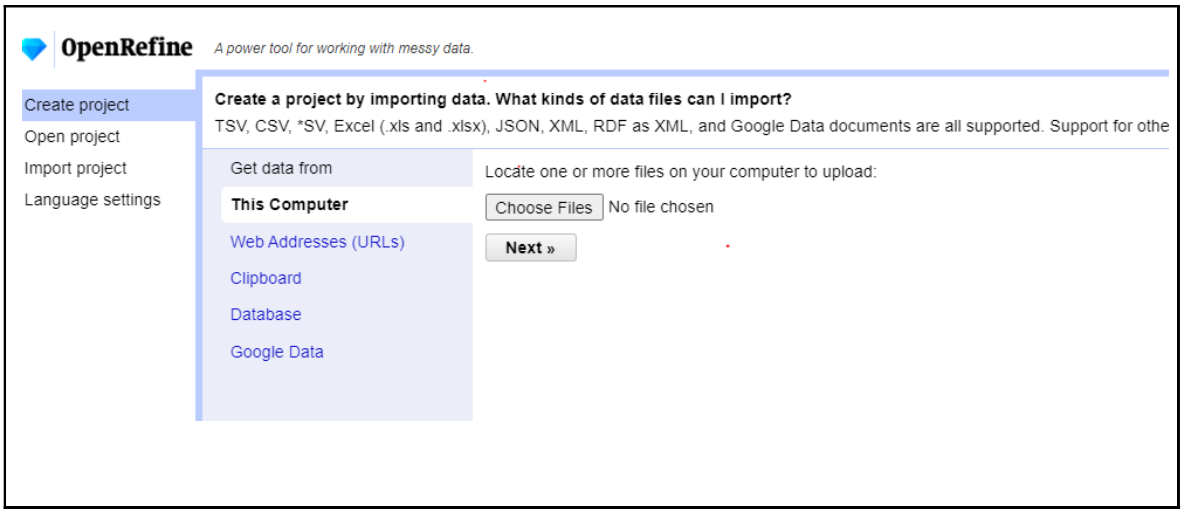
- Delete all columns except repository id, creator names, title, pages, date, qualification name, thesis type, DOI, supervisors.
- Concatenate first and surname columns to create single author name column.
- Repeat for supervisor names.
- If there are cells without supervisor data use clear contents 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:
LSETOEditedExportTemplate

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.)

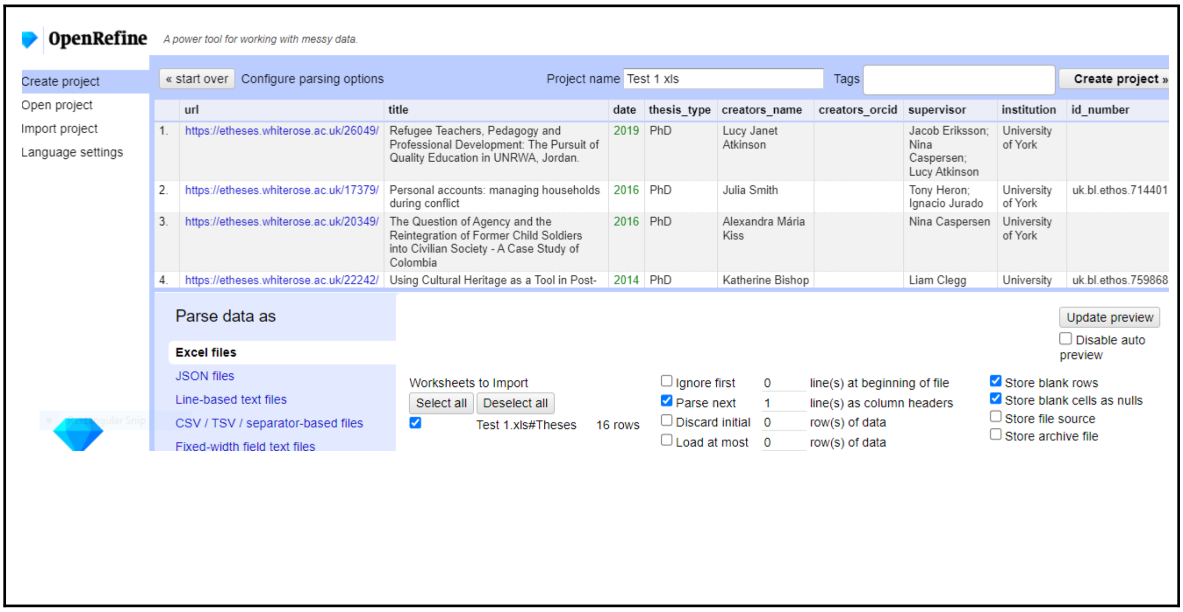
Import of data into OpenRefine



Create a project by importing data. What kinds of data files can I import?
TSV, CSV, *SV, Excel (.xls and .xlsx), JSON, XML, RDF as XML, and Google Data documents are all supported. Support for othe

Locate one or more files on your computer to upload:
Choose Files No file chosen

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.*

Project name: Test 1 xls
Tags:

	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

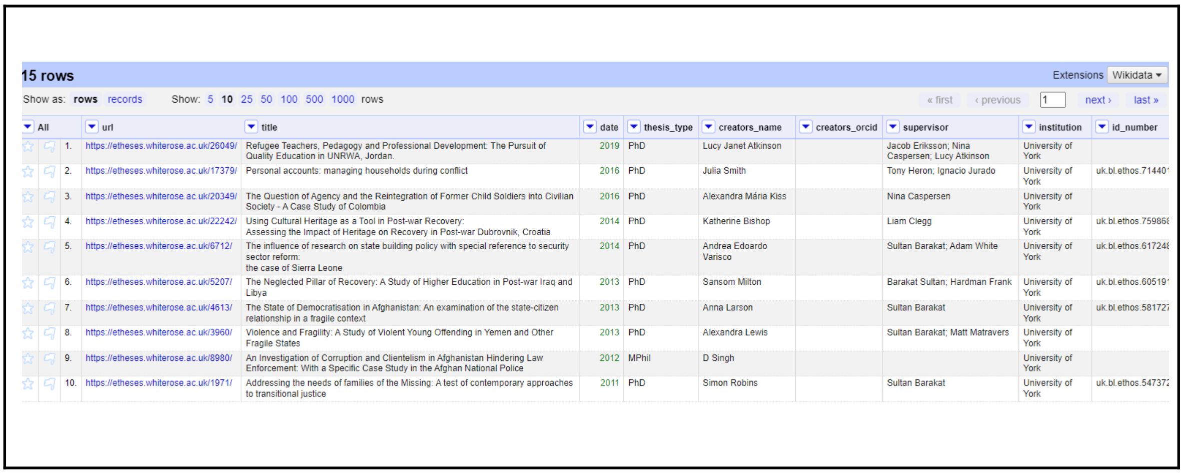
Worksheets to Import
Select all | Deselect all
☑ Test 1.xls#Theses 16 rows

☐ 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

☐ Disable auto preview

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

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



15 rows

Show as: **rows** records   Show: 5 **10** 25 50 100 500 1000 rows   Page 1

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.71440
☆	🏳	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.75986
☆	🏳	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.61724
☆	🏳	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.60519
☆	🏳	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.58172
☆	🏳	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.54737

How to edit author and thesis entries

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

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.
- 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

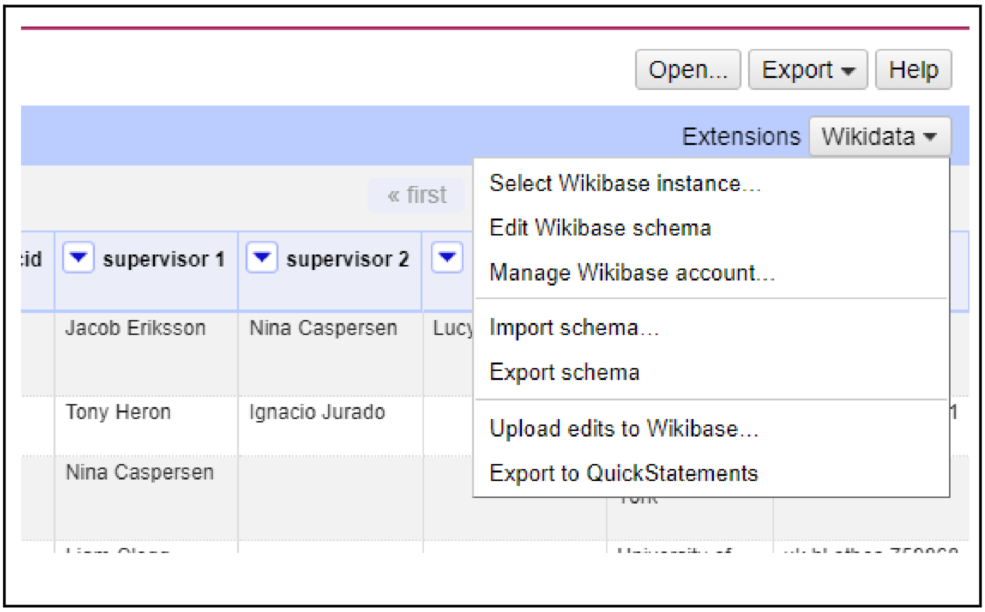
The date entry should show as a single year - ie - 2011.
If needed, use the split column option:
To separate into separate cells select *Edit column/Split into several columns.*
Select *Separate* and enter "." (full stop or whatever is in place as separator.)
Anything to the right of the full stop will 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	To create Q identities (Qids) for authors who do not already have a listing in Wikidata. 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.)
Process	Select authors name column and select *Reconcile/Start Reconciling.* Click on *Wikidata reconici link (en).* Reconcile against Q5 (human.) Select second method of reconciling by using authors 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 judgement 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 judgement/true or false.* Then *Reconcile/action/create new items for each cell.* Select all items in the set.  Open... Export Help Extensions Wikidata « first id supervisor 1 supervisor 2 Jacob Eriksson Nina Caspersen Lucy Tony Heron Ignacio Jurado Nina Caspersen Select Wikibase instance... Edit Wikibase schema Manage Wikibase account... Import schema... Export schema Upload edits to Wikibase... Export to QuickStatements Select *Extensions Wikidata,* then *Import schema.*

	Export into schema: import *author schema*. 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 *Edit Wikibase schema.* Check in *Preview* and make any alterations or corrections as needed. Export to *Upload edits to Wikibase.* author names should now appear with individual Qids in Wikidata. Check through “contributions” in your personal Wikidata account to see if all looks correct.
Result	All authors associated with project have a Qid and this is linked to the *UniversityofElsewhereProject* statement.

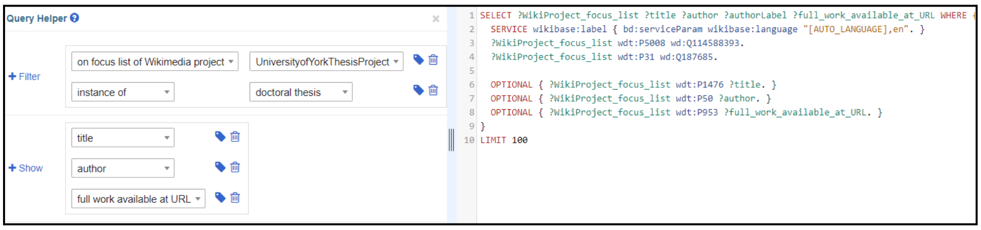
Task 3 - thesis title

Objective	To create Qids for the thesis title, and to link the title to the author record (created in task 2.)
Process	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 *Reconcile/Facets/By judgement/true or false.* Create a new item for each thesis without a match. *Reconcile/action/create new items for each cell.* Select *Extensions Wikidata,* then *Import schema.* 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 *Preview* and make any alterations or corrections as needed. Export to *Wikidata Upload.* Export into Wikidata. Check through “contributions” in your personal wikidata account to see if all looks correct.
Result	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.)

Task 4 - author (2)

Objective	To ensure that the author record Qid has a statement in it linking to the title Qid.
Process	Reconcile on title (there should be a match for every entry from the work completed in task 3.) Reconcile on author (there should be a match for every entry from the work completed in task 2.) Export into author schema (including thesis and project statements.) 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.
Result	Author record shows title live link. Title record shows author live link. Author and title records both show project statement link.

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.) 

+ Filter
on focus list of Wikimedia project: UniversityofYorkThesisProject
instance of: doctoral thesis

+ Show
title
author
full work available at URL

```
SELECT ?WikiProject_focus_list ?title ?author ?authorLabel ?full_work_available_at_URL WHERE {
  SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],en". }
  ?WikiProject_focus_list wdt:P5008 wd:Q114588393.
  ?WikiProject_focus_list wdt:P31 wd:Q187685.

  OPTIONAL { ?WikiProject_focus_list wdt:P1476 ?title. }
  OPTIONAL { ?WikiProject_focus_list wdt:P50 ?author. }
  OPTIONAL { ?WikiProject_focus_list wdt:P953 ?full_work_available_at_URL. }
}
LIMIT 100
```

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

Task 6 - author (3)

Objective	To create a statement in the author record (task 2), detailing the associated doctoral supervisor.
Process	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 neccessary. 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.
Result	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.)

Review

Review	Take time to check that all the information is showing correctly in both author and supervisor records. 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. Run query to check all appears correctly. https://w.wiki/5rwD (Insert appropriate project Qid into query in place of York number.)

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.

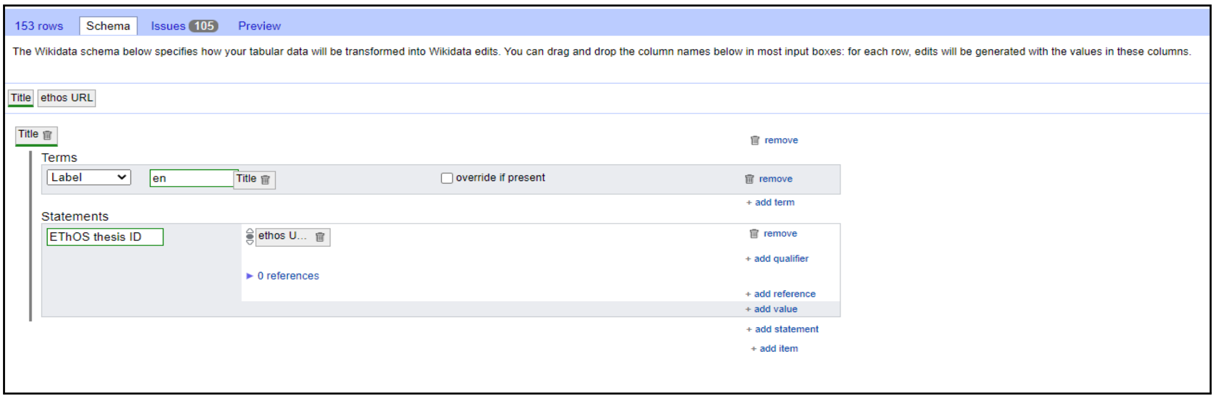
Additional options

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.

EThOS

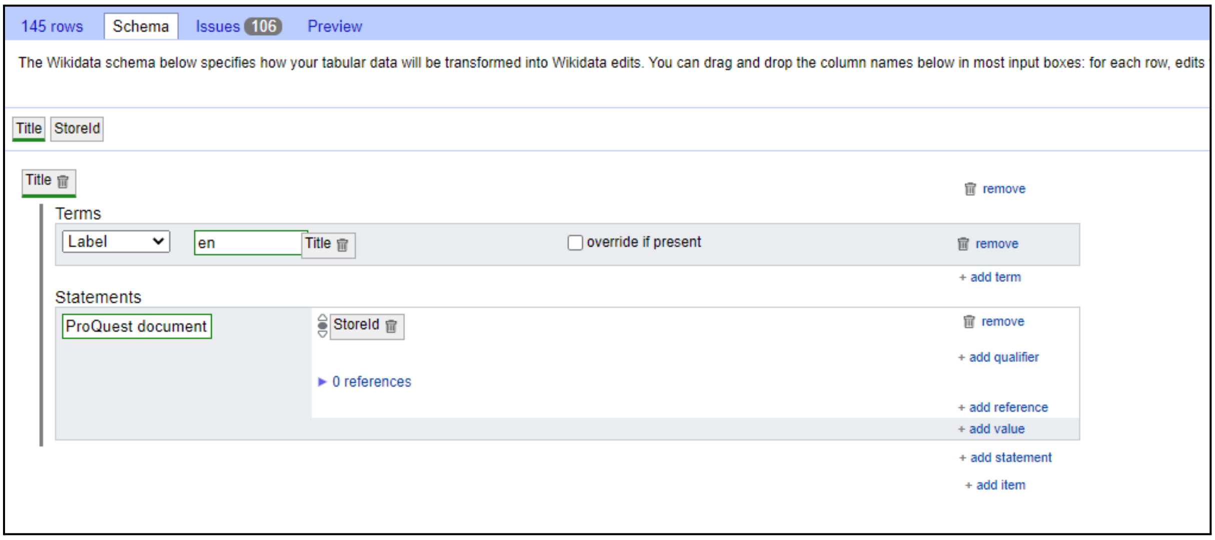
- Using the British Library's EThOS dataset extract ids and titles for institutional thesis data into a new spreadsheet. https://www.bl.uk/ethos-and-theses/re-using-ethos-data
- Create a project in OpenRefine and reconcile title column with Wikidata.
- Use *Facets/Filter* to select matched titles.
-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



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 with the values in these columns.

ProQuest

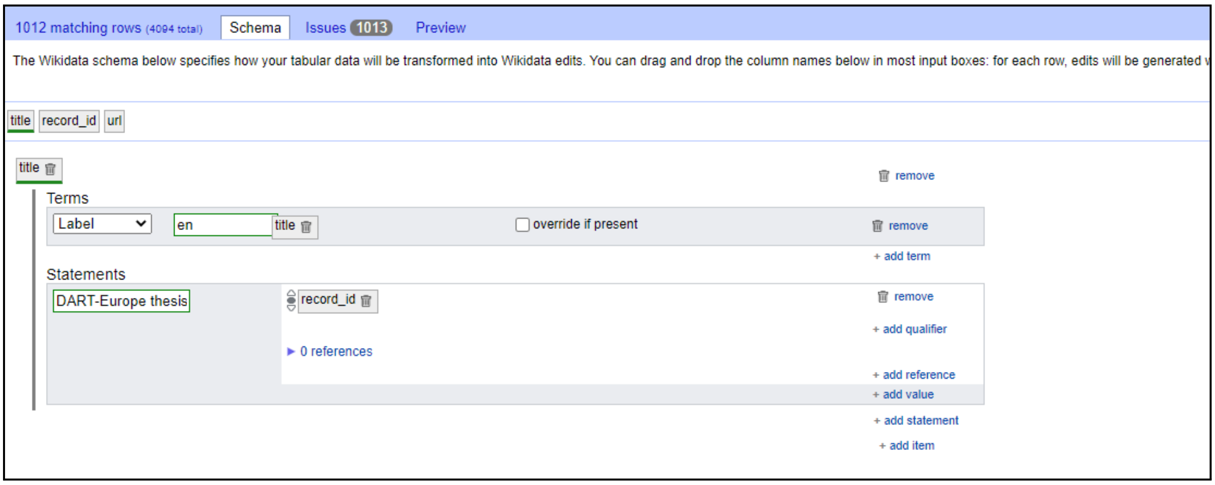
- Log into ProQuest Dissertations and Thesis Global.
- Use Advanced Search to find institution.
- Select doctoral dissertations and search. If not possible to extract all ids at once work by year and use date range in search.
- Export results and delete all columns except title and store id.
- ProQuest 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.



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

DART Europe

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



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 v

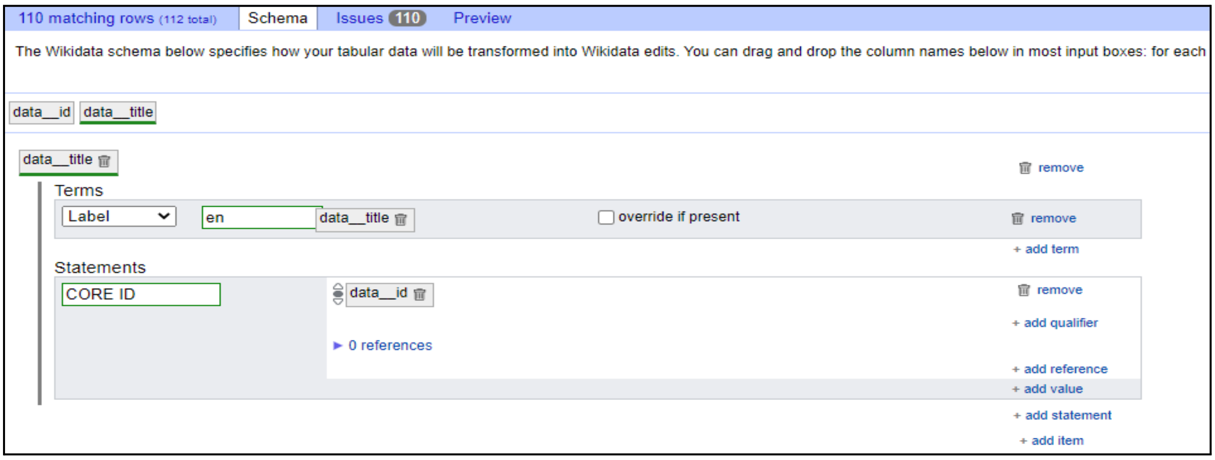
CORE

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.



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

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.

Wikidata project page

Setting up a project page will allow all project links and data to be collected in one location

- Example of project page https://www.wikidata.org/wiki/Wikidata:WikiProject_LSEThesisProject

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 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 experimenting with options to add digitised content to Wikidata.

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
Twitter: @HelsHRW

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

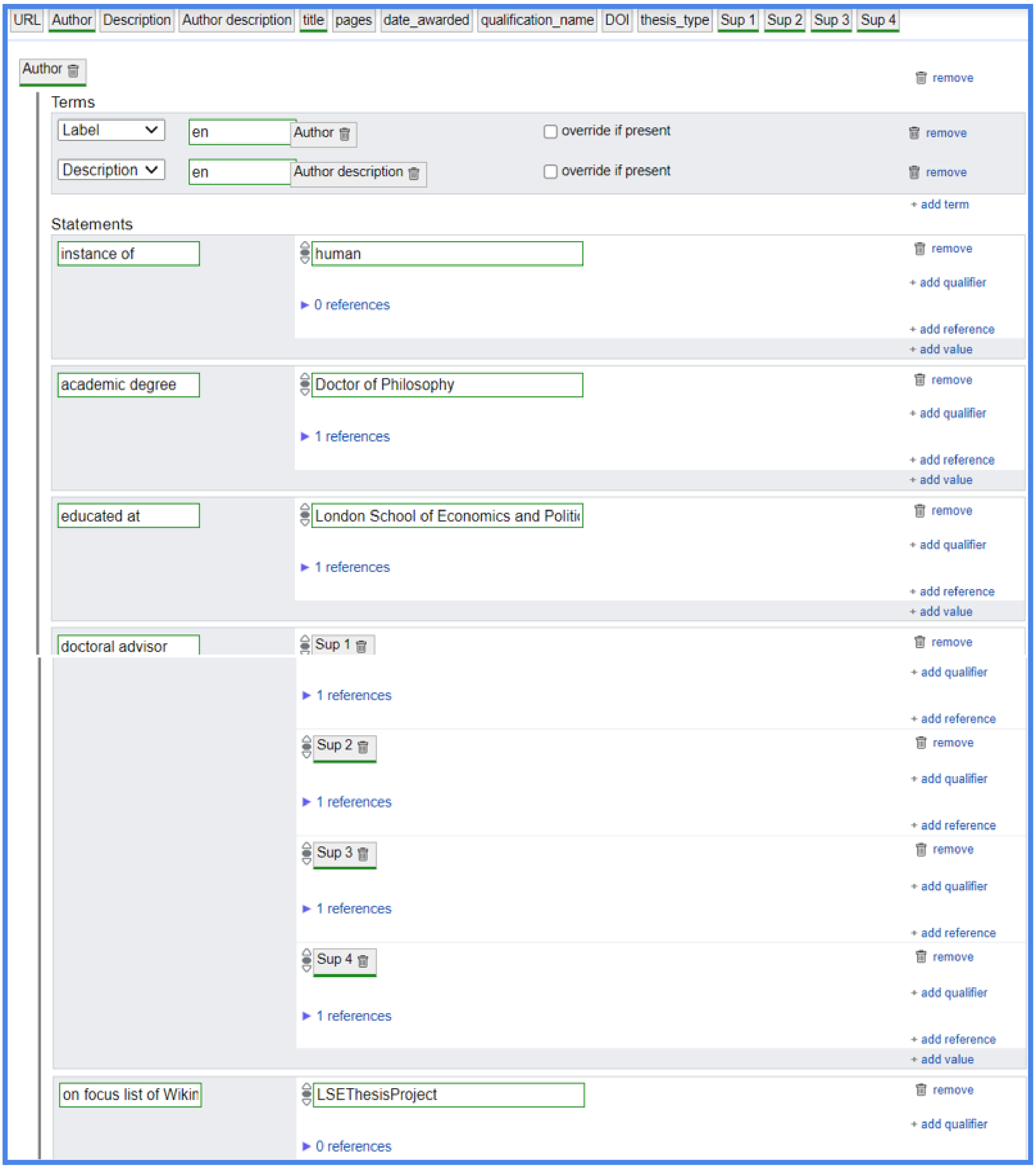
Williams, Helen K. R., Wikidata: what? why? how? Catalogue and Index (203). pp. 28-35. ISSN 2399-9667 (2020) http://eprints.lse.ac.uk/110987/1/Williams_wikidata_what_why_how_published.pdf, accessed November 30 2022

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

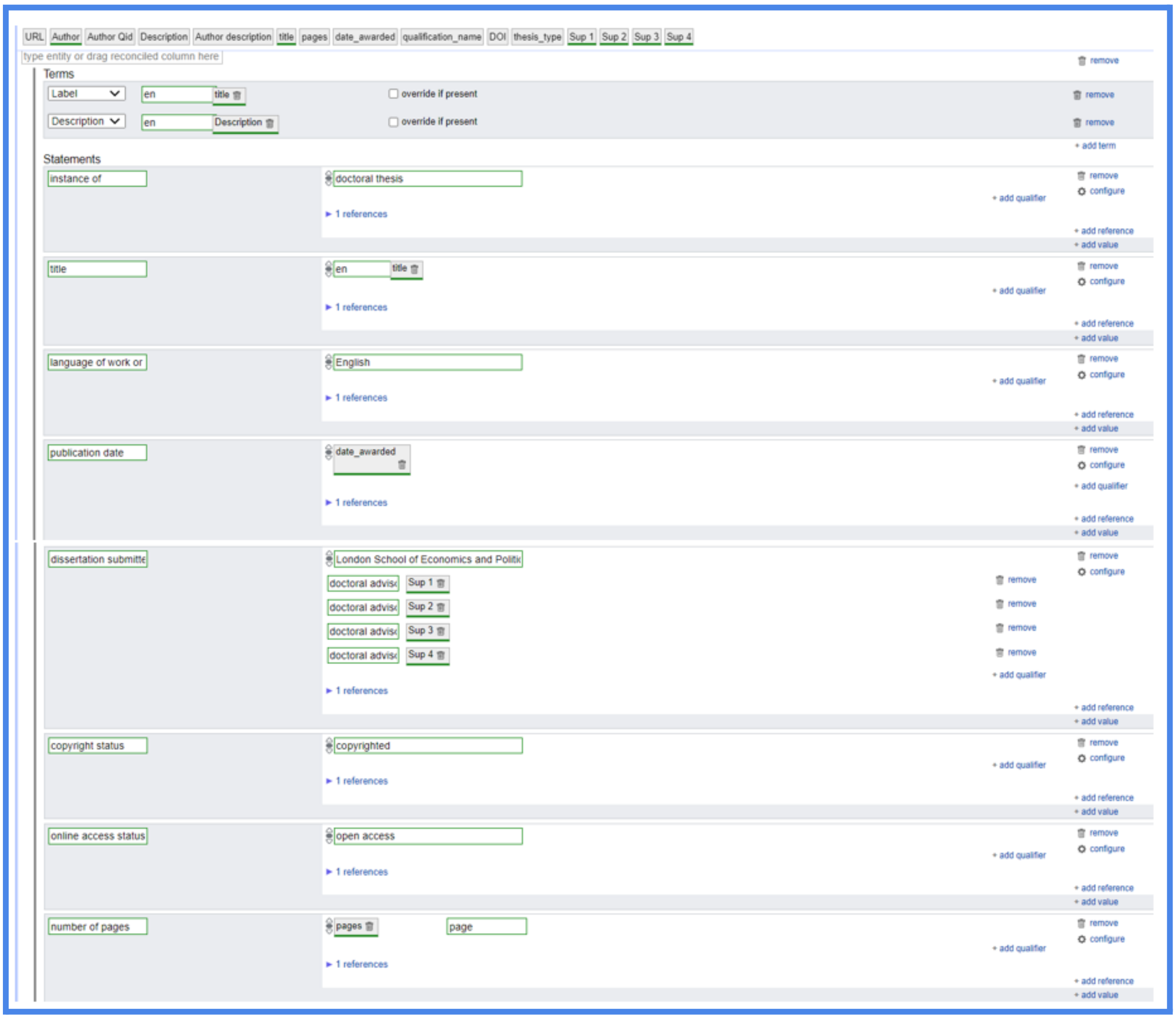
Terms

- Label — en — Author — □ override if present
- Description — en — Author description — □ override if present

Statements

- instance of — human — 0 references
- academic degree — Doctor of Philosophy — 1 references
- educated at — London School of Economics and Politi — 1 references
- doctoral advisor — Sup 1 — 1 references; Sup 2 — 1 references; Sup 3 — 1 references; Sup 4 — 1 references
- on focus list of Wikin — LSEThesisProject — 0 references

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

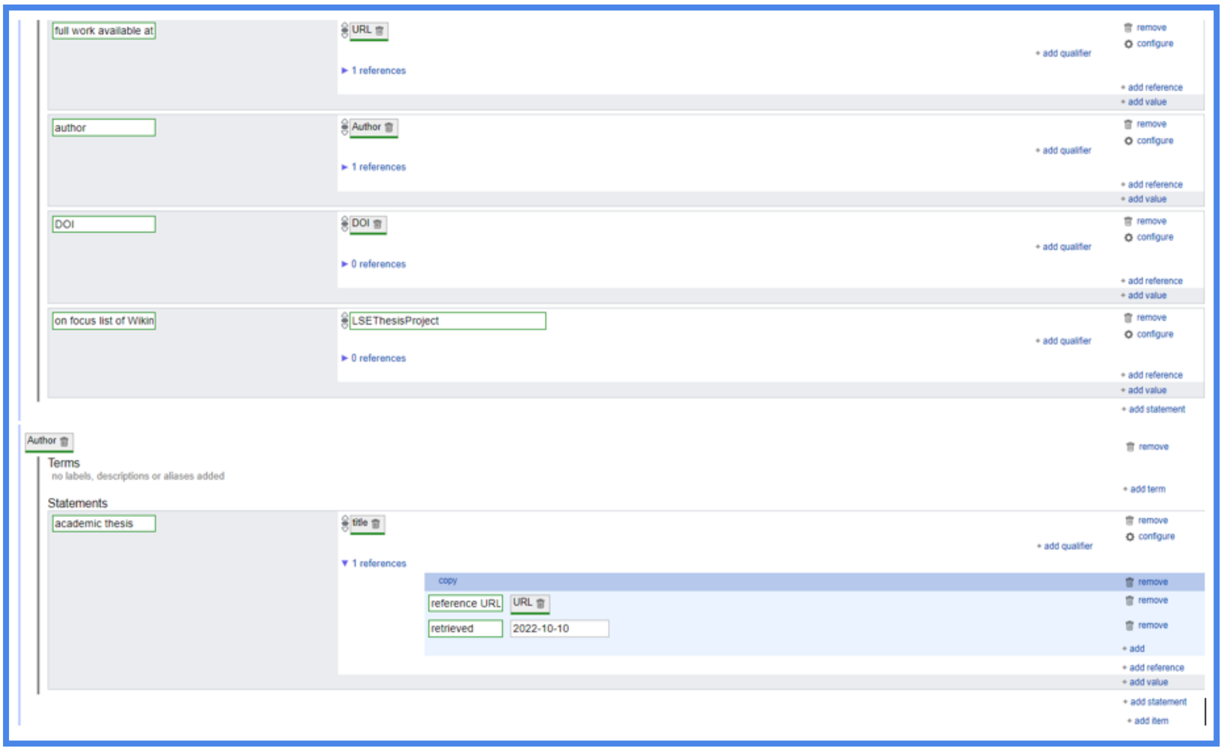
type entity or drag reconciled column here

Terms

- Label: en – title – □ override if present
- Description: en – Description – □ override if present

Statements

- instance of: doctoral thesis – ► 1 references
- title: en – title – ► 1 references
- language of work or: English – ► 1 references
- publication date: date_awarded – ► 1 references
- dissertation submitte: London School of Economics and Politic
  - doctoral advis: Sup 1
  - doctoral advis: Sup 2
  - doctoral advis: Sup 3
  - doctoral advis: Sup 4
  - ► 1 references
- copyright status: copyrighted – ► 1 references
- online access status: open access – ► 1 references
- number of pages: pages – page – ► 1 references



full work available at — URL — 1 references

author — Author — 1 references

DOI — DOI — 0 references

on focus list of Wikin — LSEThesisProject — 0 references

Author

Terms

no labels, descriptions or aliases added

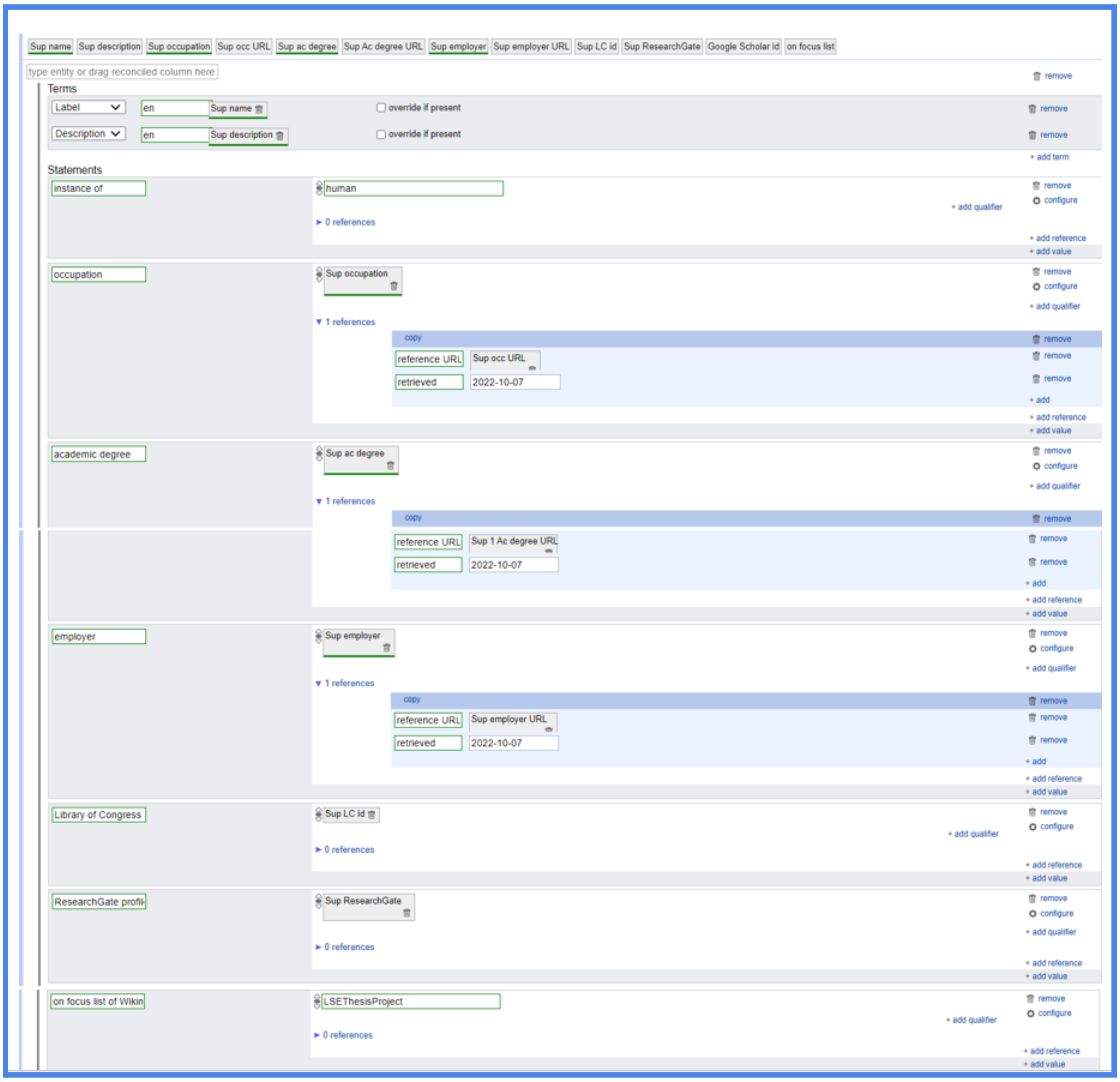
Statements

academic thesis — title — 1 references

reference URL — URL

retrieved — 2022-10-10

LSE supervisor schema



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	Lower case other than name
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)

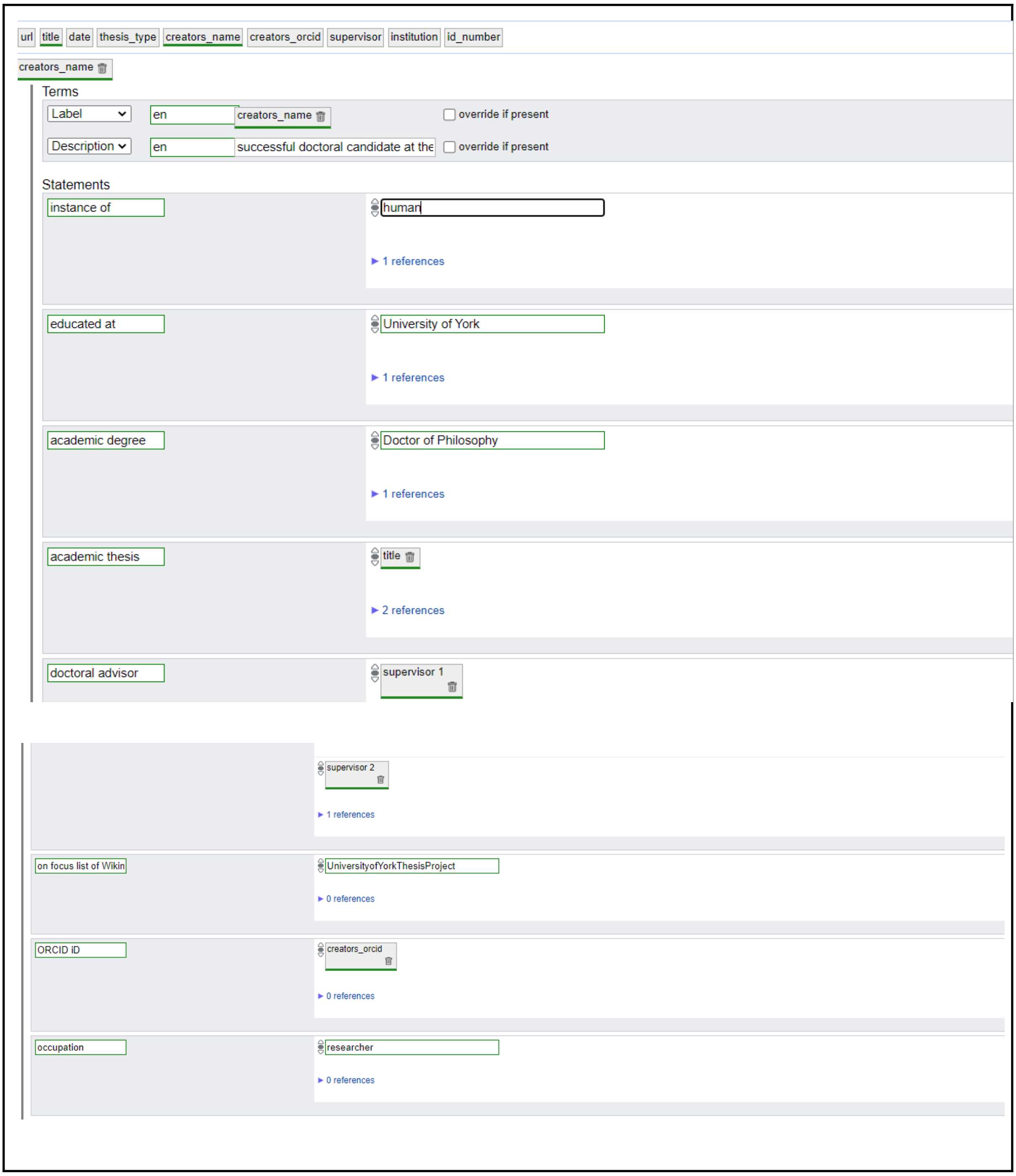
(Import text to OpenRefine)

{"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"},{"value":{"type":"wbentityvariable","columnName":"supervisor 2"},"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":"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)



url title date thesis_type creators_name creators_orcid supervisor institution id_number

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

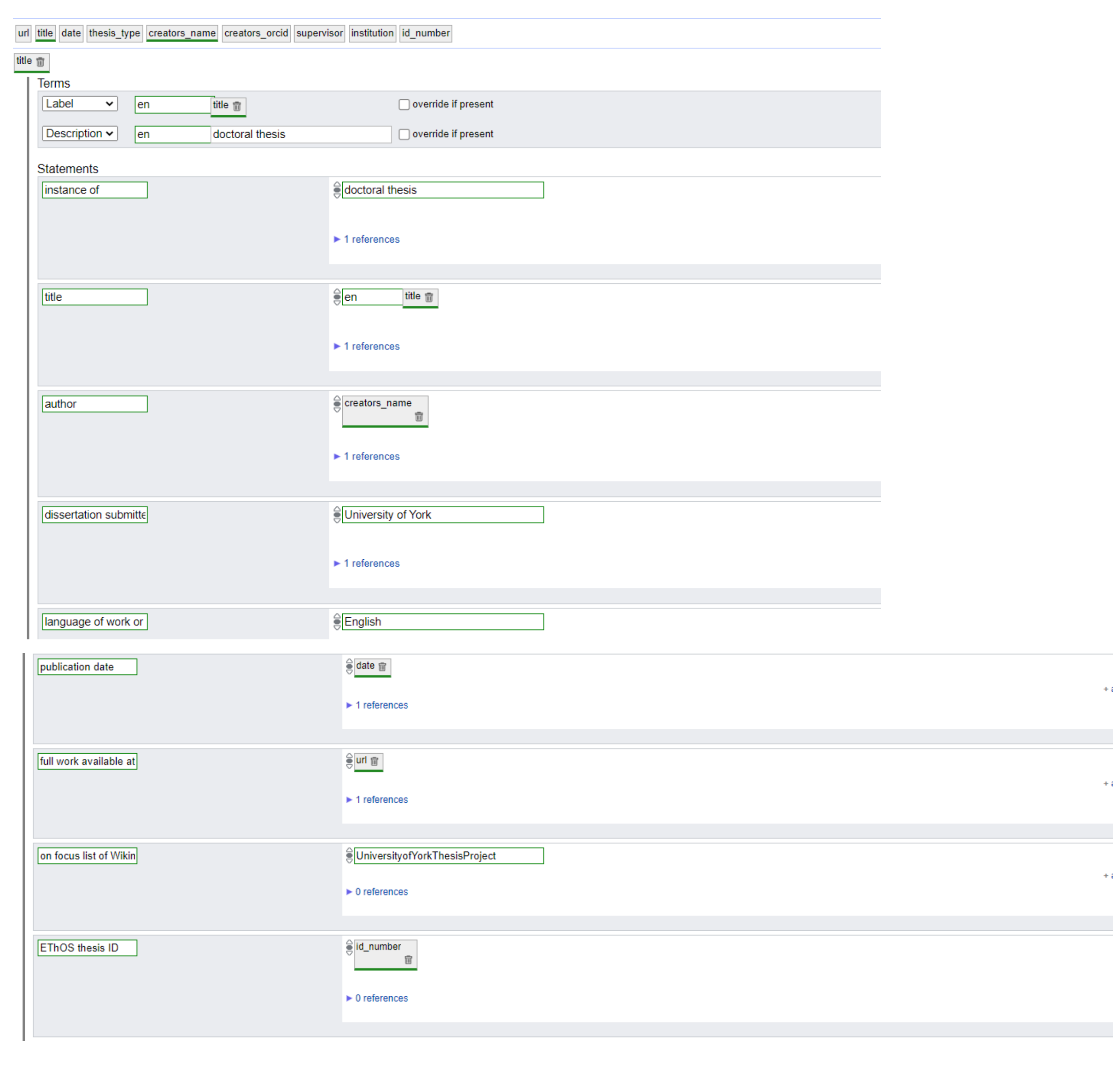
supervisor 2
▶ 1 references

on focus list of Wikin: UniversityofYorkThesisProject
▶ 0 references

ORCID iD: creators_orcid
▶ 0 references

occupation: researcher
▶ 0 references

University of York theses schema



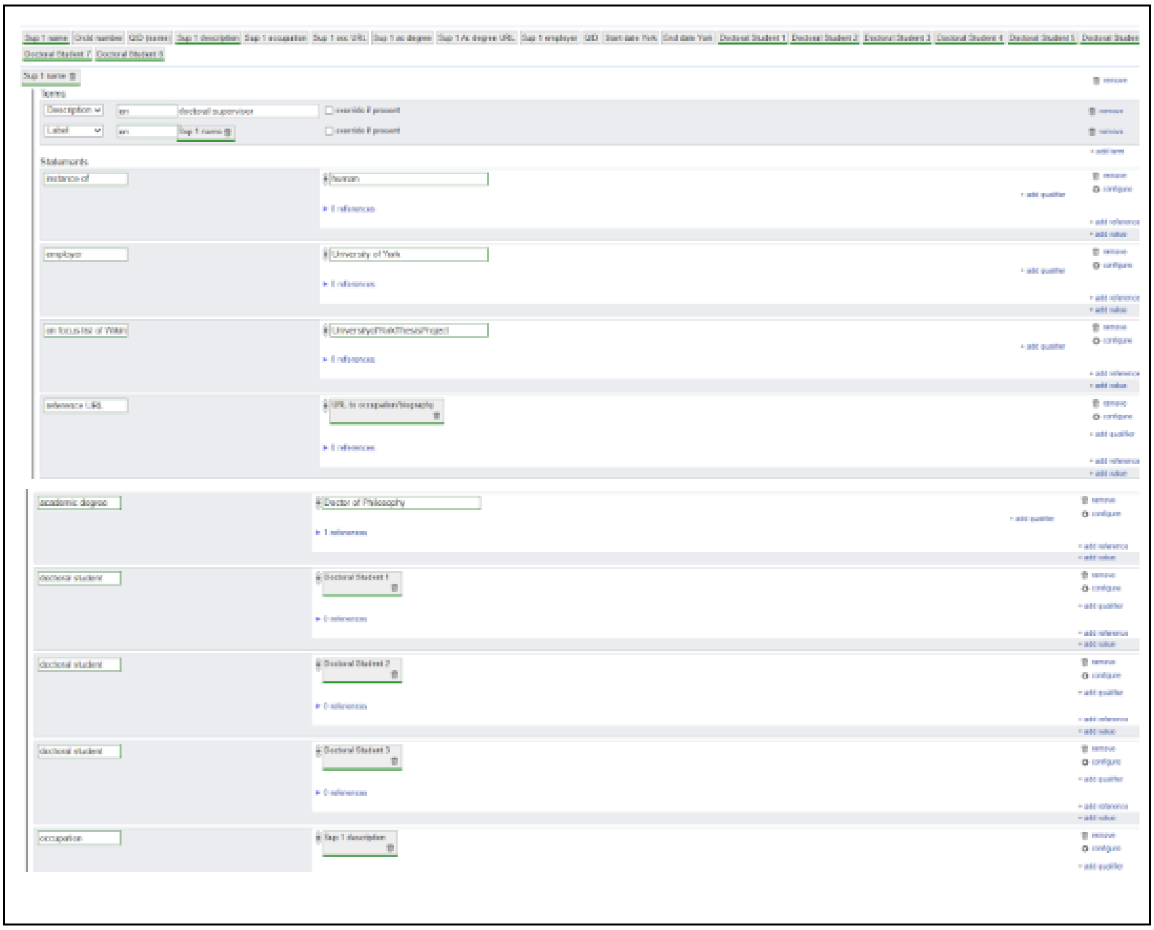
Terms

Label	en	title	□ override if present
Description	en	doctoral thesis	□ 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	url	► 1 references
on focus list of Wikin	UniversityofYorkThesisProject	► 0 references
EThOS thesis ID	id_number	► 0 references

University of York supervisor schema



Sup 1 name | Orcid number | QID (name) | Sup 1 description | Sup 1 occupation | Sup 1 occ URL | Sup 1 ac degree | Sup 1 Ac degree URL | Sup 1 employer | QID | Start date York | End date York | Doctoral Student 1 | Doctoral Student 2 | Doctoral Student 3 | Doctoral Student 4 | Doctoral Student 5 | Doctoral Studen
Doctoral Student 7 | Doctoral Student 8

Sup 1 name

Terms

Description: en — doctoral supervisor — □ override if present

Label: en — Sup 1 name — □ override if present

Statements

instance of — human — 0 references

employer — University of York — 0 references

on focus list of Wikim — UniversityofYorkThesisProject — 0 references

reference URL — URL to occupation/biography — 0 references

academic degree — Doctor of Philosophy — 1 references

doctoral student — Doctoral Student 1 — 0 references

doctoral student — Doctoral Student 2 — 0 references

doctoral student — Doctoral Student 3 — 0 references

occupation — Sup 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)	UniversityofYorkThesisProject
	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/XXXXX	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

(End of document)