Introducing the Wikidata Thesis Toolkit

Helen Williams, London School of Economics Library
Ruth Elder, University of York Library

The Wikidata Thesis Toolkit builds on the foundation of work initiated by Helen Williams (London School of Economics Library) in 2019 as a Covid "lockdown" project. This was prompted by a growing interest in Wikidata as a topic of conversation within the Metadata community, and led to her work on the development of a process to upload pre-existing theses metadata into Wikidata. Subsequent work by Ruth Elder (University of York Library) to develop a similar process flow at York added additional refinements, and the resulting collaboration between the two libraries resulted in the production of the Wikidata Thesis Toolkit.

Wikidata is described as a structured database operating as the central data store for all Wikimedia projects. It is a free and open knowledge base containing multilingual data that can be read, edited, and re-used by humans and machines, supporting global access to information.[1] The challenge was to take advantage of these beneficial attributes of Wikidata in order to develop a sustainable process to input theses metadata (already held in digital institutional repositories or datasets) into Wikidata. This enables the promotion of institutional original research to the widest possible audience through signposting back to the repository.

With competing demands and limited resources academic libraries need to be able to justify investing staff time and effort in developing Wikidata skills to promote resources in general, and institutional doctoral theses metadata specifically. The investment is justified by its support of library and institutional strategic priorities around enabling an "open as possible" approach to accessing research outputs. In linking the open content of digital repositories into Wikidata:

- scholarly content becomes more widely accessible and visible
- the role of the institution as a provider of open knowledge to local and global audiences is promoted
- the creation of unique identifiers for research outputs enables the institutional content (and the entities within it) to become part of the Linked Open Data ecosystem
- search engine results are populated with a fuller picture of globally available data because Google Knowledge Graphs, digital assistants, and Wikipedia infoboxes are all populated, in part, with information harvested from Wikidata.

In addition, many libraries will be looking to expand the work of their Metadata teams beyond traditional cataloging, and to develop staff skill sets to future-proof roles. Working with Wikidata is an accessible approach to introducing Linked Data work and expanding the range of staff digital skills, confidence, and experience.

Based on the experience of their learning journey, the authors collaborated to create the toolkit, which is created with the aim of reducing the development burden for other institutions looking to establish similar projects. It is designed as a guide, rather than a step by step handbook, and is presented with best intent, reflecting the skills, knowledge, and experience of the authors at the time of creation. The toolkit is available through a Wikidata project page, and has been shared with the wider academic library and Wikipedia communities. It supports the open research agenda by providing a method and workflow to help institutions promote their doctoral dissertations to the widest possible audience by increasing their visibility and accessibility. As the amount of doctoral research metadata in Wikidata grows the potential of surfacing unexpected connections and relationships increases, meaning the data can be explored in new ways, beyond institutional silos, to make sense of combined cultural heritage. In practical terms the toolkit includes the following:

[1] Wikidata, https://www.wikidata.org/wiki/Wikidata:Main_Page, accessed October 10 2023

- resources to introduce the basics of Wikidata and provide foundational knowledge (this takes 30-60 minutes and aims to be a low barrier starting point)
- links to further reading
- guidance on setting up an account
- practical/manual tasks to develop basic familiarity in editing Wikidata
- process overview
- steps to add metadata to Wikidata
    - data preparation
    - editing in OpenRefine
    - reconciling institutional names with Wikidata
    - creating Qids for individuals and thesis titles
- instructions to link theses to external identifiers
- guidance on creating links to theses from Wikipedia pages
- methods for using SPARQL to visualise data in ways not usually possible via an institutional repository.

The toolkit also provides guidance on measuring the impact of uploading thesis metadata to Wikidata. In the university environment the word 'impact' is often understood in its Research Excellence Framework context as 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia'.[2] In order to avoid confusion when promoting the Wikidata thesis project in a Higher Education setting it has been more appropriate within LSE's internal environment to refer instead to 'reach and engagement'.

The involvement of library colleagues in a Wikidata project provides immediate value through the development of digital skills, the opportunity to contribute to the broader network of Linked Open Data, and the promotion of unique and distinctive collections beyond the usual library silos. These benefits can be highlighted to institutional research and PhD teams, and alumni, to raise the profile of the library and demonstrate the value of metadata in ensuring that the library's unique content can be understood in the semantic web environment.

Senior leadership teams, however, will also want to see concrete evidence demonstrating the benefits of investing time and resources in this work. This evidence is not necessarily readily available. The objective of adding library metadata into Wikidata is to enable other information sources (such as search engines) to access Wikidata as a source of authoritative, referenced, structured data, indirectly enriching the wider data ecosystem. This means that measuring the direct outcomes of a Wikidata project, rather than recording basic statistics on the volume of data created or updated, can be challenging.

Many institutions are keen to benchmark with comparator organisations, so institutional rankings for doctoral theses in Wikidata can be useful, though it is important to note that these will change over time, and that for smaller institutions who simply do not have large numbers of PhD students, these rankings are not taking any account of theses produced per size of institution so they might be less valuable. This result from the Wikidata SPARQL query service provides a count of doctoral theses by institution,[3] while this dashboard includes a count of all theses types by institution and compares the completeness of the metadata available in Wikidata.[4]

LSE has investigated various measurements to review the reach of the project, including:

---

[2] UK Research and Innovation, *How Research England supports research excellence,* https://www.ukri.org/who-we-are/research-england/research-excellence/ref-impact/#contents-list accessed October 10 2023
[3] https://w.wiki/jwZ
[4] https://www.wikidata.org/wiki/Wikidata:WikiCite/Theses_by_institution

- Downloads from the institutional thesis repository, LSE Theses Online.

Increase in downloads

Total downloads 2021 **16% higher** than 2020

Total downloads 2022 **14.2% higher** than 2020

- Figures from Google Analytics

Google Analytics

Overall **250%** increase in users referred to LSETO from Wikipedia between 2019/20 and 2021/22

Contextually, in terms of total referrals into LSETO from Wikipedia:

2019 (pre project) 1%

2020 (project begins) 3%

2021 (project complete) **13% - still consistent Sept 2023**

- Twitter mentions

Twitter - etheses.lse.ac.uk

Nearly doubled in the time period reviewed

Paid resources could be used to analyse more extensively

38 mentions Feb - May 2020

74 for same time period 2021

- Extension of institutional names in Wikimedia

Increase in author/supervisor data in Wikimedia

2019: just 23% of LSE authors and supervisors existed in Wikidata

2023: 100% of LSE authors and supervisors are represented in Wikidata

Just 7% of authors and supervisors have a Wikipedia page - highlighting the quantity of unique data added which can now be used by search engines and Wikimedia editors

Wikidata is based on a structured format to describe entities, in that there are limited, prescribed ways of describing any item. For example, there are a finite number of ways of describing an author, or a dissertation, and it is not possible to add a freehand or self created description. In this way Wikidata is described as structured data.

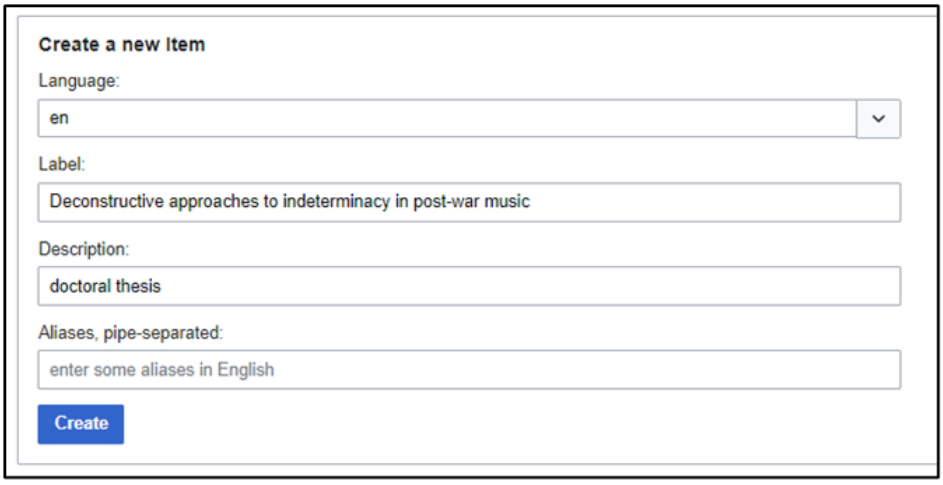
In addition to recording items Wikidata also expresses the relationships between those items. For example, a thesis entry can be linked to an author entry, the author entry can be associated with an awarding institution, and to a doctoral advisor. Each of these entities will have their own individual item entries and hence a Qid in Wikidata.

An accessible initial approach to this work for those who are unfamiliar with using Wikidata is by creating individual manual entries for theses and their authors. This helps to become orientated in Wikidata, make and correct mistakes, build confidence, and understand how the different elements work together, prior to moving onto bulk edits and uploads.

An outline of this process is set out below.

- Set up a Wikidata account and log in.
- Check to see if there is an existing entry for the specific thesis title by entering in search box on Wikidata front page.

- Select Create a new item (at the left of the screen).
- Enter thesis title and description as shown below.



- By selecting Create, this produces a minimal record of the item and a brief description.

At its most basic the structure of Wikidata is based on linking three pieces of information: an item, a property of the item, and a value associated with the property. In this example the item is the thesis titled "Deconstructive approaches to indeterminacy in post-war music." Each item or entity is given a Q number or Qid which is its unique identifier within Wikidata.

The item/property/value combination is referred to as a triple, and a triple constitutes a statement. Multiple statements are stored within a single item. The greater the number of statements within an item record, the more comprehensive, informative, accessible, and searchable the record (through linking to other relevant Wikidata entries).

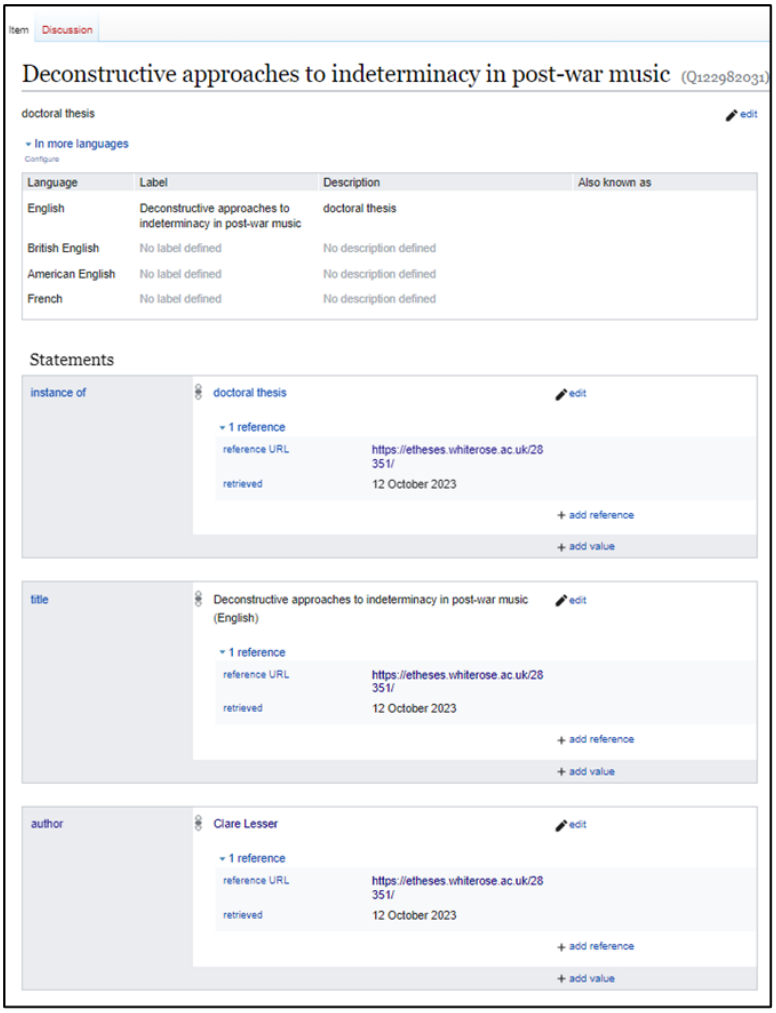
Additional statements combine to form a schema (as shown in the table below) and should be added to the basic record to make it more informative and to connect with other entities in Wikidata.

Item (Thesis title)	Deconstructive approaches to indeterminacy in post-war music	
Description	Doctoral thesis	
Statements **(Item + Property + Value)**		
Item	**Property**	**Value**
Deconstructive approaches to indeterminacy in post-war music	Instance of P31	Doctoral thesis Q187685
Deconstructive approaches to indeterminacy in post-war music	Title P1476	Deconstructive approaches to indeterminacy in post-war music
Deconstructive approaches to indeterminacy in post-war music	Author P50	Clare Lesser Q122982224

Deconstructive approaches to indeterminacy in post-war music	Dissertation submitted to P4101	University of York Q967165
Deconstructive approaches to indeterminacy in post-war music	Language of work or name P407	English Q1860
Deconstructive approaches to indeterminacy in post-war music	Publication date P577	2020
Deconstructive approaches to indeterminacy in post-war music	Full work available at URL P953	https://etheses.whiterose.ac.uk/28351
Deconstructive approaches to indeterminacy in post-war music	On focus list of Wikimedia project P5008	UniversityofYorkThesisProject Q1145883936

The addition of a reference to a statement is good practice, citing where information was sourced from and authenticating it. In this thesis example this would be in the form of referencing the thesis URL (linking to thesis entry in the digital repository) and confirming a retrieval date.

A search on Wikidata for the thesis title will now return the entry created above with a unique identifier of Q122982031. Further statements or identifiers can be added to the record at a later date and/or by any other Wikidata editor.



This structured method of storing data means that Wikidata can be queried using its built-in SPARQL query tool. SPARQL can appear complicated to beginners, and the authors recommend borrowing and editing queries written by other people to assist with learning. Multiple queries can be found on LSE's Wikidata Thesis Project page to support this.[5] By clicking on the links to any of the data visualisations on this page users can select to edit the SPARQL and substitute LSE's Qid or project Qid for that of another institution. These queries link institutional metadata on Wikidata with all the other data already existing in Wikidata and consequently make connections in ways that have not previously been easily discoverable or visible, such as:

- awards won by thesis authors and supervisors
- institutions holding their archives
- their employers (in list format, or visualised by location on a map)
- educational establishments attended
- authors and supervisors with a Wikipedia page
- relationships and chains between authors and supervisors
- image grids and Histropedia timelines
- thesis subject data
- thesis citation data.

It is important to note that an individual institution can be confident of the completeness of their institutional data, once it has been contributed to Wikidata, but the existing data these queries link with in Wikidata is not necessarily complete. This does not diminish the usefulness of the data for search engines, but it does need to be made clear to anyone making use of the data for research purposes.

For both authors Wikidata has proved an enjoyable, delightful, and satisfying challenge allowing the development of new skills, the contribution of unique institutional data to Wikidata, and the increased visibility and accessibility of this content. The Wikidata Thesis Toolkit is presented with the intention of helping other librarians to join this journey, alongside other sources of Wikidata guidance and support. These include:

- Wikidata discussion pages
- WikiEdu (https://wikiedu.org/)
- OpenRefine Community (https://openrefine.org/community)
- LD4 Wikidata Affinity Group (https://www.wikidata.org/wiki/Wikidata:WikiProject_LD4_Wikidata_Affinity_Group)

Support and guidance at the global level has been instrumental in helping both authors progress their projects, yet there is also an awareness that being able to discuss questions and challenges within a local community in the same time zone can be of value. To this end it is hoped that the Wikidata Thesis Toolkit will be a starting point to a growing community of practice among UK Higher Education and GLAM institutions (Galleries, Libraries, Archives and Museums) who are interested in developing Wikidata projects and sharing experiences with one another. To support this a new WikiProject, entitled WikiProject_UK GLAM Wikidata Projects, was launched at the conference.[6] This aims to be a space where those working with Wikidata in the UK GLAM sector can share workflows, reduce development burdens through shared communication, and spot opportunities for collaborative projects and exploring links between sets of data. This space will be a community-led initiative and by using Wikidata as its forum anyone can edit the project page, link to projects, and initiate discussion. Please do engage with the project page, link to useful resources, share projects or pre-project planning, and use the talk page to ask questions and discuss relevant topics.[7] It is hoped this space will grow, and provide details of other innovative and exciting Wikidata work across the UK, with theses and beyond.

---

[5] https://www.wikidata.org/wiki/Wikidata:WikiProject_LSEThesisProject
[6] https://www.wikidata.org/wiki/Wikidata:WikiProject_UKGLAMWikidataProjects
[7] https://www.wikidata.org/wiki/Wikidata_talk:WikiProject_UKGLAMWikidataProjects