Wikidata: What? Why? How?

Helen Williams, Metadata Manager, LSE Library

During 2019 I noticed that Wikidata was a growing topic of conversation in the world of metadata. Its power to 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. Wikidata takes the collaborative creation and management of metadata beyond the Library to a global landscape, and this issue of C&I, focused on collaboration, seemed to offer a timely opportunity to write about our Wikidata progress, even though we are still learning.

What is Wikidata?

Wikidata is a structured database operating as the central data store for all Wikimedia projects.[1] It is a 'free and open knowledge base that can be read and edited by humans and machines'[2] and is multilingual, supporting global access to information. Google Knowledge Graphs, digital assistants such as Alexa and Siri, and Wikipedia Infoboxes are all populated, in part, with information harvested from Wikidata so its content has a real impact on discovery.

Why is the Library interested in it?

Sharing library metadata with Wikidata means that unique identifiers can be minted for library content and the entities within it, pulling it into the Linked Open Data ecosystem and connecting it with unique identifiers from external knowledge systems. This starts to create bridges between currently siloed domains, which in turn impacts search engine results, making the content more widely accessible and enabling new connections and discoveries which can support global research.

How did I begin working with Wikidata?

Working with Wikidata has involved a steep, but very rewarding, learning curve. Some institutions are able to draw on the expertise of a Wikimedian in Residence, which is not our situation, so I very much started from scratch, reading articles and webpages and watching presentations and videos online. Content produced by Jason Evans,[3] Martin Poulter,[4, 5] Nick Sheppard,[6] The University of Edinburgh and Wikimedia UK,[7] and Harrison Pim[8] was particularly useful, as were discussions within the OCLC Research Library Partnership Metadata managers group, some of which are summarised on the Hanging Together blog.[9] To familiarise myself with practical work on Wikidata I started with Dan Scott's blog post about creating and editing libraries in Wikidata[10] and edited the existing item for the British Library of Political and Economic Science,[11] adding some new properties and identifiers. I also created some new items, such as LSE Digital Library.[12] which I could link using the 'has part'[13] and 'part of'[14] properties. I next looked at mapping different content types to Wikidata and created data models for LSE Digital Library collections, blogs, OA journals and online exhibitions, as well as for corporate bodies related to the Archives. I also created some corresponding Wikidata items for each model.[15, 16, 17, 18, 19] This provided some ideas for potential avenues of work, which I summarised by area of focus, including organisational, community, research, theses (inspired by Martin Poulter's work at Oxford University[20]), open access, digital and archival. A few interested colleagues kindly reviewed and discussed these options with me.

Having learnt the basics, I summarised for non-metadata colleagues how Wikidata could be of value to the Library, and why it was worth the investment of our time. Alongside this I proposed that the theses data in LSE Theses Online (LSETO) would be a valuable dataset on which to begin experimental work. This is a boundaried dataset of about 4000 records, the incorporation of which on to Wikidata would offer value to early career researchers and alumni by promoting their work, and would allow us to expose DOIs in Wikidata as these were added to the repository. Our theses deposit agreement allows us to share the metadata[21] and this project would not expose any data which had not already gone through internal processes for being made available in the public domain. As Wikidata is outside the control of the Library there are potential concerns about our data being edited by external users. However as we are using Wikidata to extend the reach of metadata that is already held in an internal library system we are not exposing ourselves to any risk of data loss; and if edits in Wikidata created additional links with related entities and sources this would be positive. Trusting that the vast majority of the Wikimedia community will be acting in good faith, and as our metadata is not controversial, I am not expecting problematic edits. We will, of course, need to keep this under review, address any issues arising and note any lessons learned so that we can take them into account for any future projects.

The work was approved and having contacted the Wikimedia Foundation to discuss our planned data donation,[22] I embarked on teaching myself how to export and manipulate data for bulk import into Wikidata, how to reconcile LSE names with existing Wikidata identifiers and how to link LSE data to identifiers in external datasets. The process I have developed is outlined below.

Exporting data from LSE Theses Online

For the purposes of learning and experimenting I wanted to work with relatively small sets of data, so I have been exporting data from LSETO one year at a time, by multiline CSV, deleting unnecessary data and concatenating various columns where required.

Preparing LSETO data for import to Wikidata

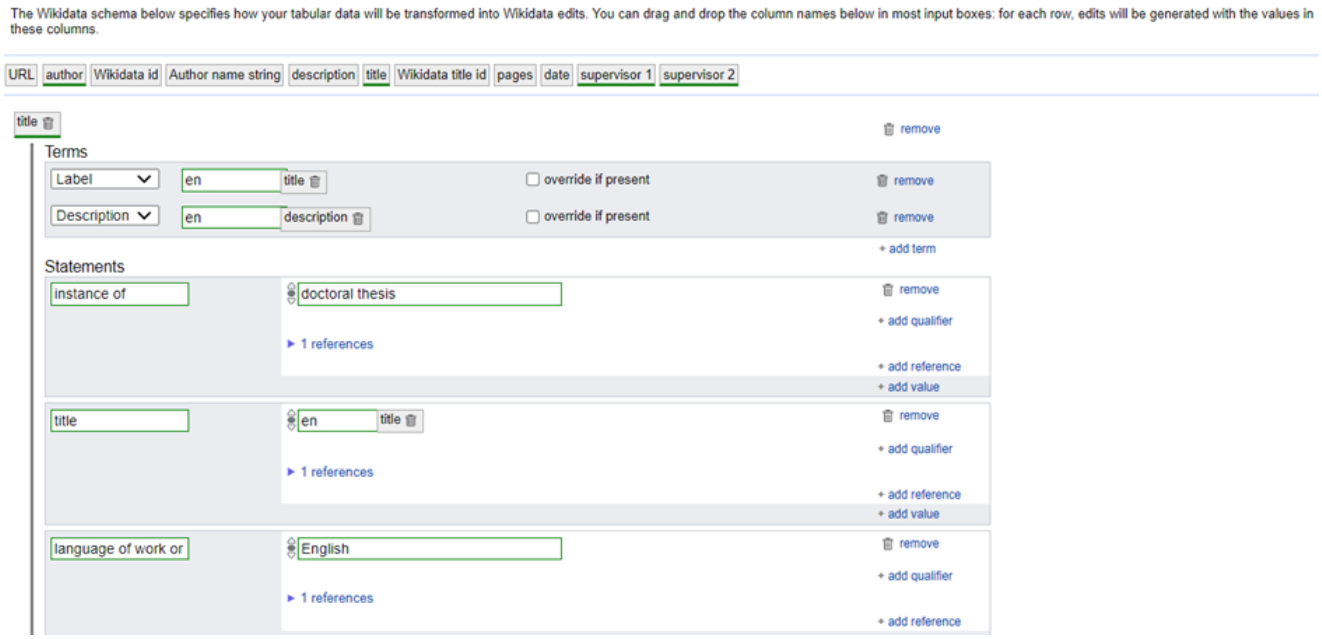
I next wanted to format the data for upload to Wikidata using QuickStatements. I followed a YouTube tutorial from Sara Thomas[23] to learn about this, and I am very grateful for subsequent guidance from Simon Cobb who helped me to identify small formatting mistakes I was making which were preventing data upload. The table below represents our data model.

A	B	C	D	E	F	G	H	I
CREATE								
LAST	Len	"Title"						
LAST	Den	"doctoral thesis by author"						
LAST	P31	Q187685			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P1476	en:"Title"			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P2093/P50	"Author"/Qid	P1932	"Author"	S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P4101	Q174570	P184	supervi-sor Qid	S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P407	Q1860			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P577	+YEAR-00-00T00:00:00Z/9			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P1104	pages			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P6216	Q50423863			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P953	"url"			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11
LAST	P6954	Q232932			S854	"url"	S813	+2021-mm-ddT00:00:00Z/11

- Each new record is identified by the presence of CREATE in column A.
- Each new statement is identified by the presence of LAST in column A.
- Column B identifies the statements to be created.
- Column C provides the data which those statements will contain.
- Len means Label in English, for which we are using thesis title.
- Text strings need to be surrounded with double quotation marks.
- Den means Description in English.
- The following properties are used to populate the Wikidata item for each of our theses:
    - P31 (instance of) Q187685 (doctoral thesis)
    - P1476 (title) – text string of thesis title, prefaced by en: to represent language
    - P2093 (author name string)
        * If the author is already represented in Wikidata P2093 is replaced with P50 (Wikidata item for author) and their Qid
        * If the name is formatted differently in LSETO the qualifier P1932 (stated as) is added in column D and the text string in column E
    - P1401 (dissertation submitted to) Q174570 (London School of Economics and Political Science)
        * If the doctoral supervisor for the thesis has a Qid the qualifier P184 (doctoral supervisor) is added in column E and the Qid in column F
    - P407 (language) Q1860 (English)
    - P577 (date) +YEAR-mm-ddT00:00:00Z/9 (for syntax see Wikidata[24])
    - P1104 (pages) – number of pages
    - P6216 (copyright status) Q50423863 (copyrighted)
    - P953 (full work available at URL) – LSETO URL
    - P6954 (access status) Q232932 (open access)
    - P365 (DOI) – not included in table above but added if already in LSETO
- S854, in column F, represents the reference URL required for each statement and column G is populated with the LSETO URL to generate that reference.
- Each reference URL requires an S813 (date retrieved) statement.

Once the metadata is formatted, as above, multiple records can be pasted into QuickStatements for upload to Wikidata.

I subsequently learnt that data can also be uploaded to Wikidata directly from tabular data in OpenRefine[25] and after some trial and error have been able to establish an upload process using this method too. This has required using the custom facet and transform functions in OpenRefine to manipulate the data into the appropriate format, including separating out author names with Qids requiring a P50 statement, and those requiring a P2093 author name string statement. I was then able to create a Wikidata schema in OpenRefine to transform the tabular data into Wikidata statements. The following 2 images show the initial canvas from which the schema can be created, and then a subsection of the schema I have created.



This schema can be exported from OpenRefine in JSON and then imported again for use on subsequent theses datasets in OpenRefine. While I have been working on this one of my colleagues is working on efficiencies in the processing and manipulation of LSETO data into our QuickStatements model. Having refined both processes we will then be able to assess which is the most effective method for upload.

Data reconciliation

OpenRefine has been a vital part of the project for two aspects of data reconciliation and I am indebted to Owen Stephens for his YouTube tutorials on this.[26] We are using P2093 (author name string) to enter author name (for example, https://www.wikidata.org/wiki/Q106600336), but this does not provide any additional author details or further discovery paths. If, however, the author is already represented in Wikidata we can use P50 (Wikidata item for author) which creates a hyperlink, allowing the discovery journey to continue more easily. Taking https://www.wikidata.org/wiki/Q105822883 as an example it can be seen that the use of P50 facilitates the discovery of additional information about the author, where this already exists in Wikidata, including various identifiers in other datasets. There are a number of authors and supervisors in our data who are not already represented in Wikidata and, although it may be valuable to create Wikidata items to represent them, limited time and resources means our initial focus is on exposing the bibliographic data. The creation of new items for living persons would also require additional ethical considerations[27].

In instances where both author and supervisor have a Qid this data can be 'round tripped'. For example, from the item on the left, clicking on 'Maurice Bloch' links to his Wikidata item where P185 (doctoral student) has been added with the identifier for 'David Lan'. On following his hyperlink users discover P1026 (academic thesis) with a hyperlink back to the item for the thesis title. For each thesis containing P50 one of my colleagues adds P1026 (academic thesis) to the author's Qid page and checks for the existence of/adds P69 (educated at) and, if both author and supervisor have Qids, links these via addition of P184 (doctoral advisor) and P185 (doctoral student).

It would be very time consuming to search Wikidata manually for each name to see if it is already represented, but the OpenRefine reconciliation function can automatically check a spreadsheet of names against Wikidata and return the relevant Qids. Some matches are made automatically (though should be checked for false positives), while others have potential matches for investigation.

The second aspect of data reconciliation which OpenRefine has facilitated is linking the newly added LSE theses content to identifiers in external datasets. I have been able to obtain thesis titles and identifiers to support automating the process for the following datasets:

Dataset	Wikidata Property	Process
EThOS	P4536	Extract LSE data from Ethos dataset[28] Edit Ethos URL to match Wikidata formatter URL[29]
DART-Europe	P8184	Contacted UCL Library Services for LSE dataset containing theses titles and corresponding identifiers
ProQuest	P6572	Extract LSE data via advanced search on ProQuest Dissertations and Theses[30]
CORE	P6409	Extract LSE data via CORE API[31] Convert JSON results to CSV

Once the titles and corresponding identifiers are in a spreadsheet a project can be created for each dataset in OpenRefine. The titles can then be reconciled with Wikidata and the facets filter used to select matched titles before creating a Wikidata schema in OpenRefine which uses label, language and title to identify the correct Wikidata item, adds statements containing the corresponding identifiers and uploads edits directly to Wikidata. The history tab of these Wikidata items shows that the original upload was from QuickStatements, and the identifiers added via OpenRefine.[32]

Following the example of Martin Poulter's SPARQL query to show Oxford theses in Wikidata[33] I created a similar query for LSE.[34] Where the author has a Wikipedia page we add the thesis title to the main body of text, with an external link to LSETO, supporting this with an inline citation using the {{cite thesis}} template, and adding the thesis to an existing infobox where applicable.[35]

At the time of writing I have added 589 theses to Wikidata, so there is still much to do, not only in terms of adding further content and refining the process, but also to consider how we might visualise the data and to investigate what kind of impact the work has had on thesis downloads from the repository. It will be evident that I am still learning and that the work is still in progress, but I see this as an exciting piece of work for the Library, with the potential to make our content more widely available and to provide an opportunity for Library staff to learn new skills. I hope that this project will go on to provide a helpful foundation from which to explore the value of Wikimedia engagement both to the Library and to LSE.

Thanks to my Library colleagues Andy Jack, Helen Porter and Neil Stewart who have been valuable 'sounding boards' for this work, and to my Metadata colleagues Ryan Kermode and Gemma Read who have been willing to get involved with this experimental area of work for our team.

References

[1]Wikidata (2019), *Welcome to Wikidata,* www.wikidata.org (Accessed 11/5/2021)
[2]Wikidata (2019), *Welcome to Wikidata,* www.wikidata.org (Accessed 11/5/2021)
[3]Evans, Jason (2020), *Creating and enriching Linked Data with Wikidata,* CILIP Metadata & Discovery Group conference 2020, https://www.youtube.com/watch?v=RApW2x4KJfw (Accessed 19/10/2020)
[4]Poulter, Martin (2019), *Wikidata: GLAM/Oxford,* https://www.wikidata.org/wiki/Wikidata:GLAM/Oxford (Accessed 19/10/2020)
[5]Poulter, Martin and Nick Sheppard (2020), *Wikimedia and universities: contributing to the global commons in the Age of Disinformation.* Insights 33 (1): 14, http://doi.org/10.1629/uksg.509 (Accessed 19/10/2020)
[6]Sheppard, Nick (2019), *Wikimedia in universities.* Leeds University Library blog, 1 Feb 2019, https://leedsunilibrary.wordpress.com/2019/02/01/wikimedia-in-universities/ (Accessed 30/4/2021)
[7]Wikimedia UK and the University of Edinburgh (2020), *Wikimedia in education,* https://open.ed.ac.uk/wikimedia-in-education/ (accessed 19/10/2020)

[7]Wikimedia UK and the University of Edinburgh (2020), *Wikimedia in education,* https://open.ed.ac.uk/wikimedia-in-education/ (accessed 19/10/2020)
[8]Pim, Harrison (2020), *Concepts, Wikidata etc.* Wikidata and Cultural Heritage Collections, Science Museum Heritage Connector webinar, 19 Jun 2020,https://www.youtube.com/watch?v=kZGcnL8gI6Y (Accessed 19/10/2020)
[9]OCLC (2020), *Hanging Together: Wikimedia,* https://hangingtogether.org/?cat=202 (Accessed 30/4/2021)
[10]Dan Scott (2018), *Creating and editing libraries in Wikidata,* https://coffeecode.net/creating-and-editing-libraries-in-wikidata.html (Accessed 11/5/2021)
[11]Wikidata (2021), *British Library of Political and Economic Science,* https://www.wikidata.org/wiki/Q2371017 (Accessed 11/5/2021)
[12]Wikidata (2021), *LSE Digital Library,* https://www.wikidata.org/wiki/Q96354844 (Accessed 11/5/2021)
[13]Wikidata (2021), *has part,* https://www.wikidata.org/wiki/Property:P527 (Accessed 11/5/2021)
[14]Wikidata (2021), *part of,* https://www.wikidata.org/wiki/Property:P361 (Accessed 11/5/2021)
[15]Wikidata (2021), *Oral evidence on the suffragette and suffragist movements,* https://www.wikidata.org/wiki/Q100380678 (Accessed 12/5/2021)
[16]Wikidata (2021), *LSE Impact Blog*, https://www.wikidata.org/wiki/Q77710848 (Accessed 12/5/2021)
[17]Wikidata (2021), *Journal of Illicit Economies and Development*, https://www.wikidata.org/wiki/Q96715673 (Accessed 12/5/2021)
[18]Wikidata (2021), *Making Modern Women*, https://www.wikidata.org/wiki/Q105556001 (Accessed 12/5/2021)
[19]Wikidata (2021), *Society for the Ministry of Women in the Church*, https://www.wikidata.org/wiki/Q82749481 (Accessed 12/5/2021)
[20]Poulter, Martin (2017), *A step forward in the sharing of open data about theses.* Bodleian Digital Library, 19 Jul 2017, http://blogs.bodleian.ox.ac.uk/digital/2017/07/19/a-step-forward-in-the-sharing-of-open-data-about-theses/ (Accessed 11/5/2021)
[21]LSE (2021), *LSE Theses Online-FAQs,* https://etheses.lse.ac.uk/faq.html (Accessed 11/5/2021)
[22]Wikidata (2021) *Wikidata: data donation,* https://www.wikidata.org/wiki/Wikidata:Data_donation (Accessed 11/5/2021)
[23]Thomas, Sara (2020), *How to use QuickStatements - a tool to bulk upload data onto Wikidata,* https://www.youtube.com/watch?v=QI7gC91eWss (Accessed 11/5/2021)
[24]Wikidata (2021) *Help: QuickStatements* https://www.wikidata.org/wiki/Help:QuickStatements (Accessed 11/5/2021)
[25]OpenRefine (2021), *Wikidata,* https://docs.openrefine.org/manual/wikidata (Accessed 11/5/2021)
[26]Stephens, Owen (2017), *Reconcilliation [sic] in OpenRefine ,* https://www.youtube.com/watch?v=q8ffvdeyuNQ&list=PL_0jeq3PjvtADzbovAgHNzOFvOlyF6uL1(Accessed 11/5/2021)
[27]Wikidata (2021), *Wikidata: Living people, https://www.wikidata.org/wiki/Wikidata:Living_people* (Accessed 11/5/2021)
[28]EThOS (2020), *EThOS* dataset, https://doi.org/10.23636/1188 (Accessed 06/11/2020)
[29]Wikidata (2021), EThOS thesis ID, https://www.wikidata.org/wiki/Property:P4536 (Accessed 11/5/2021)
[30]ProQuest (2021), *ProQuest Dissertations & Theses Global,* https://search.proquest.com/pqdtglobal (Accessed 11/5/2021)
[31]CORE (2021), *CORE API v2,* https://core.ac.uk/searchAssets/docs/#!/articles/searchArticlesBatch (Accessed 11/5/2021)
[32]Wikidata (2021), *Joining the ERM: Core executive decision-making in the UK, 1979-1990,* https://www.wikidata.org/wiki/Q106600328 (Accessed 11/5/2021)
[33]Poulter, Martin (2017), *A step forward in the sharing of open data about theses.* Bodleian Digital Library, 19 Jul 2017, http://blogs.bodleian.ox.ac.uk/digital/2017/07/19/a-step-forward-in-the-sharing-of-open-data-about-theses/ (Accessed 11/5/2021)
[34]Wikidata (2021), *LSE Theses in Wikidata,* https://w.wiki/vAK (Accessed 11/5/2021)
[35]Wikipedia (2021), *Wikipedia: list of infoboxes,* https://en.wikipedia.org/wiki/Wikipedia:List_of_infoboxes (Accessed 11/5/2021)

Images: CC0 (1.0), Wikidata, 3dman: Pixabay , OpenRefine,