LSE's Adventures in Wikidata-land: tears and triumphs down the rabbit hole

Helen Williams, Metadata Manager, LSE Library

In June 2021 C&I published *Wikidata: what? why? how?*[1] in which I outlined my experiences of learning about Wikidata, and detailed the practical processes developed to bulk-upload LSE theses metadata to Wikidata, matching it with external and Wikidata identifiers. In December 2021 I followed this up with a presentation to MDG's online conference, sharing the broader vision and environment of our work in the context of the digital shift, including some of the tears and triumphs we encountered in establishing a new area of work while operating remotely because of the pandemic, which sometimes felt like falling into a rabbit hole and ending up in a slightly different world as far as metadata was concerned.

Wikidata is a structured database operating as the central data store for all Wikimedia projects.[2] It is a "free and open knowledge base that can be read and edited by humans and machines"[3] and is multilingual, supporting global access to information. Google Knowledge Graphs, digital assistants, and Wikipedia infoboxes are all populated, in part, with information harvested from Wikidata, so its content has a significant impact on discovery. I am particularly interested in its ability to create links and show relationships between entities, and in the way in which that can create bridges between currently siloed domains and impact search engine results. If we can make our content more widely accessible and enable new connections and discoveries, this will have huge potential benefits to our libraries and institutions and to global research.

Most of us will have seen the digital shift accelerate both globally and locally since the start of the Covid-19 pandemic but will also have been aware that it was in evidence prior to that. Through the work of colleagues in the OCLC Research Library Partnership I had been noticing, during 2019, a growth in the number of experimental projects stepping beyond essential day-to-day work, a number of which included Wikidata.[4] As a result I was giving some early thought to the potential of Wikidata in our institutional context when at the beginning of 2020 the Metadata Team at LSE moved from being part of the Content and Discovery Group, to being part of the Digital Scholarship and Innovation Group. This gave a subtle change to our remit, broadening our focus from the management of scholarly content to include the development and exploration of new ways in which our metadata can support research, teaching and learning. This provided a natural opportunity to review and re-prioritise the work of the team and I was encouraged to take the risk of trying some new and different things. Initially I focused on looking to see where we could release time in order to create the space to investigate new areas of work, but when the first lockdown arrived colleagues in my team who had previously spent about 32% of their time dealing with metadata for print collections and 14% on service counters began working remotely; and suddenly the team had a significant amount of time released, sooner than I had anticipated or was quite ready for!

---

[1] Williams, Helen K. R. (2021) *Wikidata: what? why? how?* Catalogue and Index (203). pp. 28-35. ISSN 2399-9667 http://eprints.lse.ac.uk/110987/ (Accessed 14/12/2021)

[2] Wikidata (2019) www.wikidata.org (Accessed 13/12/2021)

[3] Wikidata (2019) www.wikidata.org (Accessed 13/12/2021)

[4] OCLC (2020), *Hanging Together: Wikimedia,* https://hangingtogether.org/?cat=202 (Accessed 13/12/2021)

The first step was to ensure that the Metadata team were gainfully employed in the digital direction that we were heading in, while I carved out time to learn the new skills needed to establish a new area of work in the team. Partnership was essential in achieving this. The digital shift is not something that individuals can navigate in isolation from their colleagues. Individuals and teams must acknowledge the need for partnership and actively seek to establish valuable relationships both within their own institution and externally. We already worked closely with our Research Support team to manage research outputs metadata, but now we sought to re-fashion that relationship as a collaborative partnership. Thus the Metadata team were able to develop new skills by becoming involved in the REF process, in Data Management Plans, and in LSE Press. That allowed me time to move on from reflecting on the potential of Wikidata to become engaged in practical learning and experimenting, as detailed in the June C&I article. This too required a sense of partnership in terms of my reliance on learning from the Wikimedia Community through online support pages, discussions, and contacts with a few UK experts who have very kindly answered questions and helped me troubleshoot various issues. I also gathered a few interested colleagues, namely our research support manager, digital library manager, and web editor, who have very kindly been a sounding board along the way as I have sought to examine the options and establish the direction of our work.

The Wikidata work also enabled me to establish some unexpected relationships, which I have often found to be the case when becoming involved in new areas of work. In 2020 an initiative called SHAPE[5] (Social Sciences, Humanities and the Arts for People and the Economy) was developed by the British Academy and various partner organisations. I happened to be in a meeting where SHAPE was mentioned, and I offered to create a Wikidata item to help increase the discoverability of the initiative via search engines. I realised that a Wikipedia page would also be helpful, and accordingly created that as well. This work was noted outside the Library and later led to a request to review LSE's own Wikipedia pages. We have been somewhat limited in our response to that request due to Wikipedia's Conflict of Interest Policies,[6] but we were able to establish a small group of colleagues from the Library and Communications teams to discuss these issues, and this has given a valuable opportunity to extend the conversation beyond the Library.

Having got our Wikidata project off the ground, it was time for me to develop this as a new area of work for the Metadata team by training two colleagues as 'early adopters'. Like me they did not have any existing familiarity with Wikimedia but they had begun to develop some experience with OpenRefine. The digital shift was impacting work across the library and one of my Digital Library colleagues who was also part of the Library's Training and Development Group set up what we call the Data Shapers Community of Practice. The group aims to help Library staff develop confidence with tools that will facilitate the manipulation of large datasets, so that we can streamline complex workflows. It is also intended to be a forum where we can learn and discuss together. OpenRefine was the first tool we worked with, and as our Wikidata processes rely heavily on OpenRefine this made the Wikidata project slightly less of a complete leap into the unknown than it might otherwise have been for my two colleagues.

Rather unusually, I began the training with instruction about the final step of the process. My colleague Gemma Read has kindly given me permission to share her experiences of this. She says:

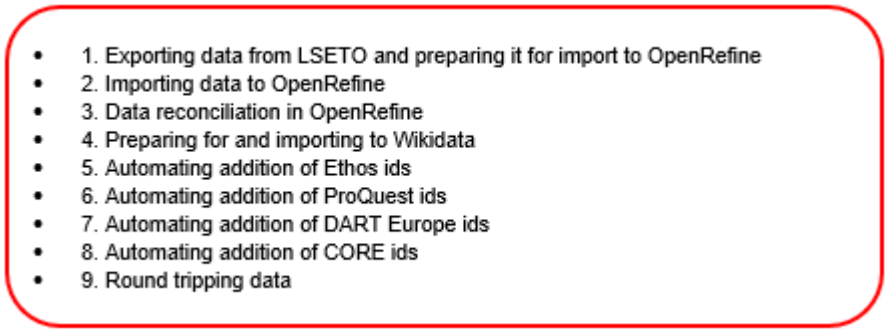
"*I first started work at the very end of the process - when all the necessary thesis data has been uploaded into Wikidata. I was shown how to create links between the new thesis records and author records which already existed in Wikidata, called roundtripping, along with adding a few further details. I also updated author entries in Wikipedia with their thesis information where required. I think it helped to see the final records in Wikidata ahead of being able to create them myself – it provided familiarity and a reference point for what I was working towards.*"

---

[5] SHAPE (2020) https://thisisshape.org.uk/ (Accessed 13/12/2021)

[6] Wikipedia (2021) *Conflict of interest* https://en.wikipedia.org/wiki/Wikipedia:Conflict_of_interest (Accessed 13/12/2021)

While my colleagues were working on the last step of the process, with data that I had already uploaded as part of my own learning, I developed detailed written instructions for that upload process. It involves nine stages, initially covering seven pages of documentation. I held individual training sessions in which the colleague concerned shared their screen and I talked them through each stage of the process in blocks of no more than an hour.



Training remotely on such a new and complex workflow was, unsurprisingly, not without its challenges. We were working without cameras, because of the high number of other windows open on our screens. This meant that it was impossible to read the visual cues that can indicate (in either direction) how well training is going. We found that having a Teams call open at the same time as trying to reconcile a dataset in OpenRefine overloaded most people's home broadband, so we had to disconnect the Teams call whenever we encountered that step in training and re-join once the data had reconciled. Not being physically together in the office made troubleshooting more of a challenge. In addition, OpenRefine runs on individual computers rather than on a central system, which meant that if colleagues had questions or problems they had to export a tar.gz file from OpenRefine and share it with me so that I could import it into OpenRefine on my laptop to troubleshoot. Finally, I was training my colleagues about a procedure that I had only just learnt myself. After 20 years of working in libraries I usually have a bit more expertise up my sleeve if I am training colleagues, but on this occasion I felt that I was barely one step ahead! This meant that we sometimes encountered problems that I had not seen before and could not immediately solve, with the result that we would have to pause the training while I worked out how to resolve the issue.

Despite all these challenges we successfully completed training and two of my colleagues have been working on the project in the latter part of 2021. My colleague Gemma helpfully added screenshots to my documentation, which extends it to 15 pages, but makes it easier to follow. She also noticed an 'extract and apply' function in OpenRefine whereby the user can extract various steps applied to one dataset and employ them on another dataset. Using this function has sped up the process a bit further and, returning to the theme of partnerships, is an example of the benefit of fostering a collaborative style of working in partnership within one's immediate team so that colleagues can be involved in improving processes. Quoting Gemma again:

"I have enjoyed learning something new, which felt completely out of my comfort zone at the start and very different to the work I have been doing in the Metadata team. I feel I have more confidence experimenting and troubleshooting in OpenRefine."

Both I and the team have found it rewarding to learn new skills, but it is important to assess whether this work is valuable to our institution. The mission of our Metadata team is to support LSE's strategy by facilitating discovery of LSE Library collections and LSE research for the LSE community and the wider world, and our vision for achieving this goal is to "create and manage comprehensive and authoritative metadata which adds value to LSE's outstanding collections, contributes to the global web of data, and facilitates wide use of the collections". In view of this goal, I was eager to carry out an interim analysis of the work, and this took place when we had added just over 1000 theses (about a quarter of the existing data) in order to investigate the impact of our work on the reach of, and engagement with, theses content.

Initially I used an existing query in the Wikidata SPARQL query service (my thanks to Martin Poulter for alerting me to this) to look at the amount of theses data we had in Wikidata in comparison to other institutions.[7]

We started somewhere between 287th and 467th place, in company with all the other institutions that had one single thesis in Wikidata (in our case the entry had been created by someone outside the institution). Six months into my experimental work we were 9th out of 467 and at the time of writing we are 5th in the table.

I also looked at data from LSE Theses Online (LSETO) between February and May 2021. Total downloads for that period were 14% higher than the same period in 2020. For comparison, the same four-month time period over the previous three years had seen an increase in downloads of 6.9 % in 2019 and decreases of 5% and 12% in the previous two years, so we were seeing a notable difference in usage. Those figures are for downloads across the whole of the theses repository, so I also wanted to analyse the data for specific titles which had been added to Wikidata. I analysed 80 titles looking at the downloads for those individual titles in the six months before and after addition to Wikidata. I found that, on average, downloads in the six months after the content was added to Wikidata were 47% higher than the preceding six months, which is a most encouraging uplift.

Google Analytics for LSETO was another source of data for analysing the impact of our work. I did not expect to see referrals to LSETO directly from Wikidata, because the purpose of putting the data in Wikidata is to enable other sources using that data to drive traffic. However, as part of the project, where a thesis author has a Wikipedia page we add their thesis title with a citation to LSETO, so I was expecting to see an increase in referrals from Wikipedia. Readers will need some background information to provide a context for the figures; our primary referral source has been (and remains) Google Scholar, and during the period of analysis that source accounted for approximately 40% of referrals to LSETO. The second referral source was Twitter at approximately 10%. Another 10 sources were referring between 1% and 6% of traffic each, and after that a long tail of approximately 300 sites were referring 0.x or 0.0x % of traffic. In the six months before the Wikidata work began LSETO received an average of 3.82% of its traffic from Wikipedia. In the following six months it increased to 9.31% (with the most recent week before my analysis ended being 13.61% of traffic). This moved Wikipedia from the 5th referral source in the six months before Wikidata work began, to the 3rd referral source in the six months since (still following Google Scholar and Twitter). Finally I had a closer look at Twitter itself, finding that between February and May 2020 there were 38 mentions of ‘etheses.lse.ac.uk’ on Twitter, increasing to 74 during the same period for 2021.

We found these interim figures to be very promising indications of increased reach of, and engagement with, the theses content, and we plan to carry out further analysis once the whole dataset has been loaded and sufficient time has elapsed to collect meaningful data. At the time of writing we have loaded about three quarters of the existing theses metadata to Wikidata.

---

[7] Wikidata query service (2021) *Count of doctoral theses in Wikidata by institution* https://w.wiki/jwZ (Accessed 14/12/2021)

Wikidata also facilitates visualising data in new ways through the SPARQL query service. An example is provided by this graph[8] showing the relationships between our theses authors and supervisors (I edited a SPARQL query written by the Massachusetts Institute of Technology in order to achieve this). We have recently organised two SPARQL sessions for our Data Shapers Community of Practice, with an external trainer, and are now hoping to produce some further visualisations with our data.

Building on our success with the Wikidata theses project I proposed expanding Wikidata work in the library, with the hope that it will not only extend the reach of content and data that is unique to LSE but will also continue to grow digital scholarship expertise in the team and, in turn, foster the skills required to progress future developments. Initially we have begun creating Wikidata items for content on LSE Press, and are now part-way through automating that process, planning training for the team, and investigating options to link our data with citation data via Wikidata. A further three proposals, as follows, are all significant pieces of work in terms of the time that would be required to scope out and implement them, and have been discussed with our Library Team Leaders:

- Special collections focus – under-exposed or under-represented content where search engine discoverability could be enhanced by inclusion in Wikidata
- Digital Library focus – increase discoverability of digitised content via Wikidata or investigate creating collections map of LSE Digital Library content on Wikidata
- Researcher focus – utilize potential of Wikidata as identifier hub to support management of names related to LSE, and enhance data for use by search engines.

The third proposal brings potential benefits in terms of promoting researchers and research, but as it involves data for living people[9] it would require some institutional-level discussion. This is, therefore, an idea we may return to when we have done some further work with digitized content and developed further expertise with Wikidata. Working with our own digitised content is under our control and the work can be contained within the library, rather than creating interdependencies with other parties, which feels important while we are still learning. This approach also enables us to increase the discoverability of content which can be accessed online, rather than pointing users to content that has to be viewed in the physical building.

In conclusion, it certainly feels as if we have thrown ourselves down the rabbit hole in terms of embracing the digital shift. Some of the challenges have indeed felt almost worthy of tears, but by using Wikidata, rather than embarking on a project where we had to develop our own system, we have also had the triumphs of seeing immediate results available on the Web at each stage of the project. We are excited about monitoring the impact of the theses work going forwards, and establishing new areas of work with the Press and digitised content, and I hope the Metadata team can use the confidence built through engaging in this aspect of the digital shift to pop down some other rabbit holes and explore how we can further develop the role of metadata in supporting research, teaching, and learning.

[8] Wikidata query service (2021) Relationship between authors and supervisors of doctoral dissertations submitted to LSE https://w.wiki/3Z$7 (Accessed 14/12/2021)

[9] Wikidata (2021) Living people https://www.wikidata.org/wiki/Wikidata:Living_people (Accessed 14/12/2021)