Strategic Inclusion of Library Data in Wikidata

Helping the National Library of Sweden take a foray into the Wikimedia projects

Strategic Inclusion of Library Data in Wikidata is a collaboration project between Wikimedia Sverige and the National Library of Sweden which was carried out between May 1, 2018 and July 31, 2019. The project will be continued, as another round of financing has been approved for the period August 1 2019 to August 31 2020.

This report outlines the main goals, activities and milestones of the project, especially in areas that might be of interest to the broader Wikidata/WikiCite community.

Introduction

The National Library of Sweden maintains a national database of library data, Libris, which serves a central hub of information about the collections of over 100 Swedish libraries – including public, college and special libraries. The main parts of Libris are the works database (books, journals etc.) and the authority database (authority posts for authors represented in the works database).

A large part of the works database is the Swedish National Bibliography – a dataset of information about all the books and other materials published in Sweden. Right now this dataset contains over 740,000 items, 650,000 of which are books. In 2011, the National Library released the Swedish National Bibliography, as well as the authority database, under an open license, CC0.

This licensing decision, although made before Wikidata launched, created exciting opportunities for working with and re-using the data on the Wikimedia projects. It was also one of the most important factors behind our project. Our goal became to upload parts of this data
to Wikidata. Our primary focus was on works that are used as references on Wikipedia in Swedish, and their authors.

At the time of starting the project, the National Library had initiated a transition of Libris to a new data framework, Bibframe 2.0 and implementing a more modern version of the public-facing interface of the catalog, which is still ongoing as of now. The introduction of Bibframe 2.0 and linked open data made it possible for the library to participate in the modern open data landscape – and we were given an opportunity to take part in it.

Interestingly, the new data model in Libris has a lot in common with the FRBR-based data model for books on Wikidata that was created by WikiProject Books. Most importantly, the work-edition distinction was going to be implemented in Libris, making it possible for users to identify multiple editions and translations of the same work.

From our communication with the National Library we know that this distinction is going to be implemented, and has been given high priority. At the time of writing, work-level items have not yet been created, meaning Wikidata cannot benefit from the data structure. However, in the future it might be possible to use Libris to automatically create work-level items and connect them to their editions. This would be extremely valuable for Wikidata in general and for WikiProject Books in particular, as the data model developed there could be tested in practice on a large scale.

 Goals

The project Strategic Inclusion of Library Data in Wikidata aimed at fulfilling the following goals:

1. Increase the use of Swedish bibliographical data on the Wikimedia projects.

2. Build a base for creating a work-level infrastructure before it is implemented in Libris.

3. Provide inspiration and expertise for the National Library staff involved in Linked Open Data work, such as by illustrating the value of their work with concrete examples and by together identifying future development possibilities for the free knowledge movement.

4. Show the importance of libraries to the general public by highlighting their work and their valuable collections.
5. Highlight how the Wikimedia projects can be used by libraries to prioritize their digitization projects.

**Activities**

**Preparations**

Before we started working with the data from Libris, we researched the current state of bibliographical data on Wikidata and on Swedish Wikipedia. The goal of this activity was to understand the available resources and generate ideas for our work. We wrote a series of reports aimed at library professionals so that they could get a picture of how bibliographical data is handled on the Wikimedia projects.

**Data uploads**

Three Libris-related properties exist on Wikidata. Because of the work to transition Libris to a new infrastructure, the way posts in Libris are assigned identifiers is not consistent.

- **P906 SELIBR ID**: identifier for posts in the Libris authority catalog – mostly humans, and to a smaller degree organizations and events. Since about 2018, this identifier is not assigned anymore.

- **P1182 LIBRIS Editions**: identifier for edition-level items in Libris. Since about 2018, this identifier is not assigned anymore.

- **P5587 LIBRIS URI**: universal identifier for all types of objects in the new Libris catalog. From about 2018, new Libris posts (both editions and authorities) have only this identifier. Older items have both a URI and a Selibr or Editions ID.

**The authority database**

Because of the aforementioned identifier transition, Wikidata already contained a large amount of items for humans with P906 (old authority ID), about 60,000, but very few with P5587 (the new URI).

We thus added Libris URI's to about 60,000 Wikidata items for humans. By connecting them to the new Libris, we made it easier to work with them in the future.
The editions database

While our overarching goal was to import a large amount of Libris data to Wikidata, we obviously did not want to aim at importing all the 700,000 items in the National Bibliography. First of all, we had to develop and test appropriate technical infrastructure. Secondly, the question of whether all of the bibliography should be uploaded was, and is, still open; we elaborate on this further in the final section of this report. We needed a small, well-defined test set, and so we settled for uploading about 500 posts representing some of the most cited books on Swedish Wikipedia. The reason for this selection was that we wanted to work with data that was interesting and relevant to the Swedish Wikimedia community, increasing the chances that it would be noticed and expanded upon. To identify those works, we used the Citations with identifiers in Wikipedia dataset.

The dataset provided us with interesting data; however, it turned out it was not sufficient to accurately represent the sources used on Swedish Wikipedia, as it only included works with identifiers (primarily ISBN). This put older, pre-ISBN books at a disadvantage, as they did not appear in the dataset. This was particularly unfortunate for Swedish Wikipedia, which has a strong tradition of using public domain literature hosted on Project Runeberg (a repository of digitized works predominantly from the Nordic countries, which is much more popular in Sweden than Wikisource). Because of that, we did a separate analysis of just links to Project Runeberg, which identified the most used sources from there.

Some of the results of our analyses can be seen here:

- **Most popular works with ISBN identifiers**
- **Most popular works collected** from the {{Cite book}} template, harvested from Swedish Wikipedia, regardless of having ISBN
- **Most popular websites used** in <ref> tags, harvested from Swedish Wikipedia
- **Most popular works included** in the special source citation template {{runeberg.org}}, harvested from Swedish Wikipedia

These analyses gave us a more complete picture of what sources are being used on Swedish Wikipedia. In the end, we chose to focus on the works with ISBN identifiers to upload from Libris to Wikidata.

Furthermore, we took advantage of the fact that Wikidata already has a property for identifying Project Runeberg books, P3155, and uploaded about 5,000 new items from the Runeberg catalog to Wikidata. This greatly increased the number of bibliographical Wikidata items.
potentially interesting to the Swedish Wikipedia community, helping us make our project relevant to local Wikipedians.

Events

We presented or discussed our project at the following events:

Hack for Heritage in Visby, Gotland. October 5–7 2018
Hack for Heritage is a Swedish hackathon aimed at GLAMs and similar actors who actively work with cultural heritage data from different sources. We had a short presentation about Wikimedia data and how it can be consumed.

WikiCite 2018, USA. November 27–29 2018
We had a presentation about the project and its place in the WikiCite landscape.

Libris användardag, Stockholm. December 5 2018
A one-day conference for library professionals working with Libris, organized by the National Library of Sweden. It did not focus entirely on Libris, but also included guests from other countries who presented more general viewpoints and projects centered around Linked Open Data (LOD) in libraries. We had a short presentation about our collaboration, where we focused on Wikidata as a platform for bibliographical data.

Library Data Hackathon, Stockholm. December 2018
A hackathon for National Library staff that we organized, with a focus on Wikidata, Libris and LOD.

Digikult 2019, Göteborg, Sweden. April 25 2019
A conference about digital cultural heritage for GLAM professionals. We had a presentation about open bibliographical data and Wikidata, where we emphasized how LOD creates value by linking between items from different sources, as well as highlighted the power of the volunteer community of Wikidata.

Libris Development Council, Stockholm. May 22 2019
A small meeting of representatives of different branches of the Swedish library sector. We presented our project, placing the library's work with Linked Open Data in a larger perspective
and highlighting the possibilities that participating in the open knowledge movement creates for the institution. We were also invited to share our experience of working with the developing new version of Libris as part of designing the work-level layer of the catalog.

**Reflections from the data uploading process**

The biggest issue we encountered when working with the data was the insufficient degree to which the entities in Libris were interconnected. The most palpable example is that not all authors have their own authority posts, meaning that many book entries represent their authors with strings rather than with authority ID's. Sometimes this was the case even if an authority post existed. This is not a problem unique to Libris; indeed, it has been acknowledged by the Wikidata community by making a frequent use of the Author name string (P2093) property. This makeshift solution, which has been applied over 21 million times, prevents the information from being lost and flags the items to volunteers who are able to improve them.

A much more difficult problem was the matching of publishers and publication places to Wikidata items. Unlike authors, in Libris they are represented with strings, not linked data. Since we were dealing with a comparatively small dataset, we were able to identify the most common publishers and publication places, and create mapping tables for them.

If we want to scale up and upload larger datasets, we will need a more sustainable solution to this problem. One way of approaching it is working from a data dump, rather than ingesting the Libris posts one-by-one via the API. That way, we will be able to generate reports about the most common publisher/publication place values and make sure they can be matched correctly to Wikidata. Indeed, using a local dump is more practical – both for us and for the data provider – when processing larger amounts of data, and it is what we did when uploading our couple hundred items. Still, we thought that the possibility to access live Libris data on a per-post basis was important enough to implement it as an option in our software, as it enables us to access new or recently modified data. In the future, we would like to develop a tool that enables any Wikipedia/Wikidata editor to create a Wikidata item based on a Libris post, in which case it will be important to convert as much Libris data to Wikidata statements as possible.

We have communicated this issue with the National Library, as it is something they could benefit from having in my mind when implementing a Linked Data interface in their catalog. In particular, the inconsistent way in which authors are represented, which could be solved by
creating stronger ties between the works database and the authority database, is something they are aware of.

**Plans for the future**

We are going to continue our collaboration with the National Library throughout 2019 and 2020. During that time, we are going to focus on:

- Uploading more bibliographical data to Wikidata. We are going to continue our work with the National Bibliography, identifying and uploading other interesting subsets of it. In addition, we are going to look into other sources of bibliographical data that are included in Libris but not released under a CC0 waiver due to being managed by entities other than the National Library. We are going to have a dialog with the National Library and other relevant entities about licensing additional data as CC0.

- Developing solutions to keep bibliographical data on Wikidata up-to-date. In particular, we are going to focus on tools to enable Wikidata contributors to add items based on Libris data when they are needed. We see this as an important contribution to the community which will enable the Swedish bibliographic commons to grow even after the conclusion of our project.

- Enabling, encouraging and researching the use of bibliographical data from Wikidata on Swedish Wikipedia. First of all, we would like to port the [[Cite Q]] template to Swedish Wikipedia, adjusting it to the citation formatting practices prevalent in the Swedish language. Furthermore, we are going to initiate a community discussion in order to gain insight into what the community needs and expects in this regard. The final decision about whether and how this template should be used will be made by the community.

- Communication and events. In order to increase the visibility of our project among Library and Information Science experts, Wikimedians and the general public, we are going to undertake active communication work. This will include presentations at conferences, workshops and training for volunteers who would like to work with bibliographical data, etc. We are going to address both Swedish and international target groups. We have invited staff from the National Library to Wikimania in Stockholm, and together we are organizing the Wikidata & Wikibase for National Libraries meeting as a related activity.
**WikiCite perspective**

We have been working on this project at the same time as the WikiCite initiative has been developing, providing insights and perspectives from around the world. This is very fortunate, as it gives us an opportunity to place our work in a larger context, as well as to learn from other's experiences and serve as a test case for building synergies between national libraries and Wikimedia organizations. Additionally, it gives us an opportunity to explore an area that is not part of the core focus of WikiCite, namely books. The WikiCite community has been concentrating primarily on academic papers and citations, culminating in uploads of millions of items of scientific articles – many of which are interconnected to demonstrate how they cite each other. This has been made possible by the availability of accessible databases. Books have received comparatively little attention, meaning that the data provided by the National Library can make a significant contribution to the project.

Something that has come up often when discussion our project with others is its place in the possible future scenarios for WikiCite. We have had an opportunity to review and think about all the three scenarios throughout working with the project. At first, we took a Wikipedia-focused perspective on the open data from the National Library. Our first goal was learning which books and other publications are most commonly used by Swedish Wikipedians, and we did create Wikidata items for a couple hundred of them – an attempt to materialize the "Database of all sources cited in Wikimedia projects" scenario, if on a smaller scale.

We then considered the National Bibliography dataset, which despite consisting of several hundreds of thousands posts is cohesive and manageable – and licenses as CC0. Such a limited and curated bibliographic corpus could be a subject of the second WikiCite scenario – "A platform to host, curate, and annotate bibliographic corpora". Finally, the ambitious "bibliographic commons" scenario has always been in our minds, since the data shared by the National Library is not enough to accurately represent even the richness of publications in Sweden. Indeed, we have also looked at other data sources, such as Project Runeberg – a digital library of older Swedish and other Nordic literature, akin to Project Gutenberg – and created thousands of items for the works in its catalog as well. Wikidata as a platform offers fantastic possibilities to mix and enrich data from different sources.

This brings us to a question that has been put to us several times, and which we have pondered: should – and could – the entirety of the Swedish National Bibliography be ingested onto Wikidata?
When it comes to notability, it is an interesting question, and the answer depends on the direction in which we want WikiCite to go. The National Bibliography contains everything from scientific literature to novels and children's books, as it is meant to encompass the entirety of materials published in Sweden. The vast majority of them, if uploaded to Wikidata, might never be relevant to anyone; a one-off brochure published back in the 1970's by an author with an indistinguishable name that neither we nor the library can match to an authority file is unlikely to attract the attention of editors or contribute to the citation networks. On the other hand, the very fact that they are included in the catalog, and assigned an ID number, means they fulfill Wikidata's notability criteria; they are "clearly identifiable conceptual or material entities" whose existence can be proven by linking to Libris. The same can be said about every author in the Libris authority database; as a community, we have agreed that an identifier like VIAF or ISNI is enough for a person to be Wikidata-notable.

There are currently about 60,000 Wikidata items that are a version, edition or translation – as well as an unknown number of badly modelled items that are meant to represent editions. This means that just by uploading the entirety of the Swedish National Bibliography, we could raise this number tenfold – a tremendous step forward for the WikiCite movement that would nevertheless require careful consideration before taking. We would need to take a close look at the data model developed by WikiProject Books and confirm that it is appropriate for our needs, as by implementing it in such a large number of items – within a single institutional upload – we would be de facto accepting it as a standard.

Then there is the question of the work-edition distinction. There has been, to our knowledge, no large-scale organized upload of work-edition networks. Indeed, only about 21 thousands edition items – about a third – have an edition or translation of statement. There is a lot to do in this area, and it is something we would like to investigate after the National Library has implemented a usable work layer in Libris.

To summarize, our work with bibliographical data from the National Library of Sweden gave us valuable insight into the reality of working with bibliographical data on Wikidata. We are looking forward to continuing this work, both in terms of uploading larger amounts of data and, more importantly – reflecting upon and discussing how to best make it work for all the actors involved: the Wikimedia community, other users of our platforms and the data providers who have so generously shared their resources.