Expanding what is possible around GLAMs on the Wikimedia projects

A White Paper as Guidance for Future Work

developed as part of the FindingGLAMs project

WIKIMEDIA SVERIGE

Table of Contents

Background

Many GLAMs collaborate with the Wikimedia community to share their collections. Among them is the Metropolitan Museum of Art with several hundred thousands images on Wikimedia Commons. Embroidered map sampler (1783). CC0, via Wikimedia Commons.

FindingGLAMs

FindingGLAMs[1] was a project by Wikimedia Sverige in collaboration with UNESCO and the Wikimedia Foundation, with generous funding from the Swedish Postcode Lottery Foundation. It ran from August 13, 2018 to February 29, 2020.

The overarching goal of the project was to improve the representation of the world's cultural heritage institutions – Galleries, Libraries, Archives and Museums (GLAMs) – and their collections on the Wikimedia platforms. We initiated several types of activities within the project to achieve that goal; one of them was adding structured data about GLAMs to Wikidata to make it fast and easy to find them, which gave the project its name.

We embarked upon this project in hopes of laying the foundation for long-term and sustainable collaboration between cultural heritage institutions and the Wikimedia movement (the global community of contributors to Wikipedia, Wikidata and other projects hosted by the Wikimedia Foundation). That is why a large part of FindingGLAMs was done together with our GLAM partners and their collections, with an eye on identifying problems, challenges and possibilities for improvement. In this white paper, we present what this work

[1] https://meta.wikimedia.org/wiki/FindingGLAMs

has taught us about the technical infrastructure for content partnerships in a number of case studies.

These case studies and the experience we have gained in the process will allow us to reduce the time and learning curve needed to share content from institutional partners in the future and greatly enhance the amount of high quality knowledge that can be made available to the general public through the Wikimedia platforms.

For a more detailed overview of the other activities included in the FindingGLAMs project, apart from the GLAM content uploads, see Case Study 1: FindingGLAMs.

Structured Data on Commons are opening up new opportunities

Structured Data on Commons (SDC)[2] is a project that provides the technical infrastructure to complement the wikitext[3], templates[4] and categories[5] on Wikimedia Commons with structured data, which is machine-readable, queryable, straightforward to add and edit, and multilingual (currently up to 300 languages). This will make media files much easier to describe, discover, understand and analyze. This project is being developed by the Wikimedia Foundation and the first functionalities were enabled to the public in January 2019. The team working on this project was also developing case studies and coordinated the work with the FindingGLAMs project.

Case Study 6 describes Structured Data on Commons in the context of a GLAM collection and explores this technology in more detail.

Standing on the shoulders of giants

Wikimedians have collaborated with GLAM institutions for a long time. Synergies between the two sectors feel natural, unavoidable even. The Wikimedia community works towards making the world's knowledge available for everyone, and it is the world's libraries, museums, galleries and archives that for centuries have been safeguarding and sharing knowledge. At the same time, it is in the interest of GLAMs to reach wider and more diverse audiences. Wikipedia being one of the world's most popular websites, it is a superb platform to target. Some GLAMs are famous worldwide, but for the majority of them, attracting audiences outside their home countries or regions is a challenge. For them, the access to a free, multilingual platform with a worldwide readership is invaluable.

Some GLAMs, educational institutions and similar organizations employ **Wikimedians in Residence** – experienced members of the Wikimedia Community who serve as liaisons between the organization and the open knowledge movement.[6] By supporting the organizations in sharing their resources under open licenses and educating them about the

[2] https://commons.wikimedia.org/wiki/Commons:Structured_data
[3] https://en.wikipedia.org/wiki/Help:Wikitext
[4] https://en.wikipedia.org/wiki/Help:Template
[5] https://commons.wikimedia.org/wiki/Commons:Categories
[6] https://outreach.wikimedia.org/wiki/Wikipedian_in_Residence

Wikimedia platforms, Wikimedians in Residence build mutual understanding and lay the foundation for sustainable, long-term collaboration between the organizations and the Wikimedia movement. The first assignment of this type took place at the British Museum in 2010; since then, over 150 organizations around the world have welcomed Wikimedians in Residence, including the US National Archives, the National Library of Israel, the Bodleian Libraries and the National Museum in Warsaw. While most of these have been short-term assignments, there are also some long-term collaborations; UNESCO has had a Wikimedian in Residence since 2015.

The global **GLAM-Wiki** initiative connects Wikimedians working together to support cultural heritage institutions who want to work with Wikimedia to produce open-access, freely-reusable content for the public.[7] One of their tasks is documenting the GLAM collaborations taking place around the globe to provide examples of what can be done with the Wikimedia platforms and to inspire both Wikimedians and GLAMs to work with free materials.[8]

The **Metropolitan Museum of Art** in the United States is an example of a successful partnership between a high-profile cultural institution and Wikimedia. In 2017, the museum launched their Open Access Policy, putting over 406,000 images of public-domain artworks under the CC0 license.[9] This is the most generous of the Creative Commons licenses, enabling anyone to freely copy, modify and distribute the materials, including for commercial use.[10] Those images were then uploaded to Wikimedia Commons by a Wikimedian in Residence.[11]

A year later, it was clear that the resources shared by the museum were very popular.[12] The nearly 4,000 images included in Wikipedia articles were reaching 10 million viewers per month. What was particularly interesting is that the most popular artworks reached a much larger audience through Wikipedia than through the museum's own website; a third of that audience came from Wikipedia articles in languages other in English. One of the world's most famous museums still benefited tremendously from sharing their materials on open platforms.

On the other end of the GLAM-Wiki spectrum, there are initiatives by the Wikimedia community to better describe the world's cultural heritage. A notable example is the **Sum of All Paintings** project, aiming at creating a Wikidata item for every notable painting in the world.[13] This covers every painting in the collection of a notable institution or created by a notable artist: an ambitious scope. The project coordinates the efforts of Wikimedians across the globe who collect information about open-access data on artworks and work together to upload it to Wikidata and improve it. Institutions that have released such data include the Finnish National Gallery,[14] the Tate Gallery[15] and the Minneapolis Institute of Art.[16]

[7] https://outreach.wikimedia.org/wiki/GLAM
[8] https://outreach.wikimedia.org/wiki/GLAM/Case_studies/Archived
[9] https://www.metmuseum.org/about-the-met/policies-and-documents/open-access
[10] https://creativecommons.org/publicdomain/zero/1.0/deed.en
[11] https://commons.wikimedia.org/wiki/Category:Images_from_Metropolitan_Museum_of_Art
[12] https://www.metmuseum.org/blogs/now-at-the-met/2018/open-access-at-the-met-year-one
[13] https://www.wikidata.org/wiki/Wikidata:WikiProject_sum_of_all_paintings
[14] https://github.com/hugovk/finnishnationalgallery

Other Wikimedians are collaborating to improve the information about cultural heritage institutions and increase its re-use on Wikipedia. The **Sum of All GLAMs** project was carried out by the Wiki Movement Brazil User Group[17] in collaboration with OpenGLAM CH,[18] with the goal of laying the foundations for an international knowledge base for heritage institutions.[19] An important part of the project was examining and improving the data modelling standards for cultural heritage institutions. Another was the implementation of Wikidata-driven infoboxes on Wikipedia in various languages, making it possible to quickly create articles with essential information from Wikidata. The **WikiProject Heritage Institutions** is an informal initiative on Wikidata dedicated to the coordination of activities regarding heritage institutions, including the Sum of All GLAMs and FindingGLAMs projects.[20]

The Sum of All GLAMs project had similar goals as Wikimedia Sverige's FindingGLAMs, but a different focus; for example, we chose not to work on implementing Wikidata-driven infoboxes, and put more emphasis on rapid data ingestion through dataset uploads and crowdsourcing rather than on data cleansing and modelling issues. Our similar ambitions, as well as the fact that the timelines of our projects partially overlapped, lead to many interesting discussions and exchange of experiences. If anything, two projects focused on GLAMs in Wikidata running in parallel increased the visibility of this subject matter among the Wikimedia community, shining the light on the many problems to tackle and the different ways for volunteers to get engaged.

Why Case Studies?

Wikimedia Sverige has a decade of experience of content partnerships. What we have learned from uploading materials to Wikimedia Commons and Wikidata on a large scale is that while every GLAM and every collection is unique and idiosyncratic, they have a lot in common. By studying real life projects, we hope to highlight the challenges and problems that Wikimedians – both independent volunteers and affiliates – face when working with the different types of content that cultural heritage institutions collect and share. We also hope that our experiences will be interesting and helpful for cultural heritage institutions engaged in, or considering getting engaged in, collaboration with the open knowledge movement.

Method for choosing Case Studies

Our intention in selecting the case studies was to represent the breadth of material that Wikimedians encounter in their work with content shared by GLAM institutions. We were particularly interested in exploring media and data types that are less straightforward to work with thana "typical" GLAM collaboration e.g uploading a number of digitized artworks to Wikimedia Commons.

[15] https://github.com/tategallery/collection
[16] https://github.com/artsmia/collection
[17] https://meta.wikimedia.org/wiki/Wiki_Movement_Brazil_User_Group
[18] https://glam.opendata.ch/
[19] https://www.societybyte.swiss/2019/07/04/an-international-knowledge-base-for-all-heritage-institutions-part-1/
[20] https://www.wikidata.org/wiki/Wikidata:WikiProject_Heritage_institutions

The purpose of this white paper is not merely to describe our projects, but, most importantly, to highlight the problems and challenges that need to be solved in order to strengthen the technosocial infrastructure of content partnerships. That is why we decided to engage in projects that were new to us, with an emphasis on new and imperfect technologies, such as Structured Data on Commons and Lexicographical data on Wikidata. We believe that using those technologies to solve real life problems can provide valuable guidance for their development.

Finally, the choice of case studies was heavily influenced by our GLAM partners and the material they could provide while the FindingGLAMs project was underway. We actively approached a number of partner organizations we have previously worked with and asked if they had specific content they wanted to share, offering our support in the cases when they did. This in itself was valuable as it renewed contact and allowed us to deepen our relationship with a number of important cultural heritage institutions in Sweden.

Case Study 1: FindingGLAMs – Finding the Galleries, Libraries, Archives and Museums

Key facts

Time: August 2018 – February 2020

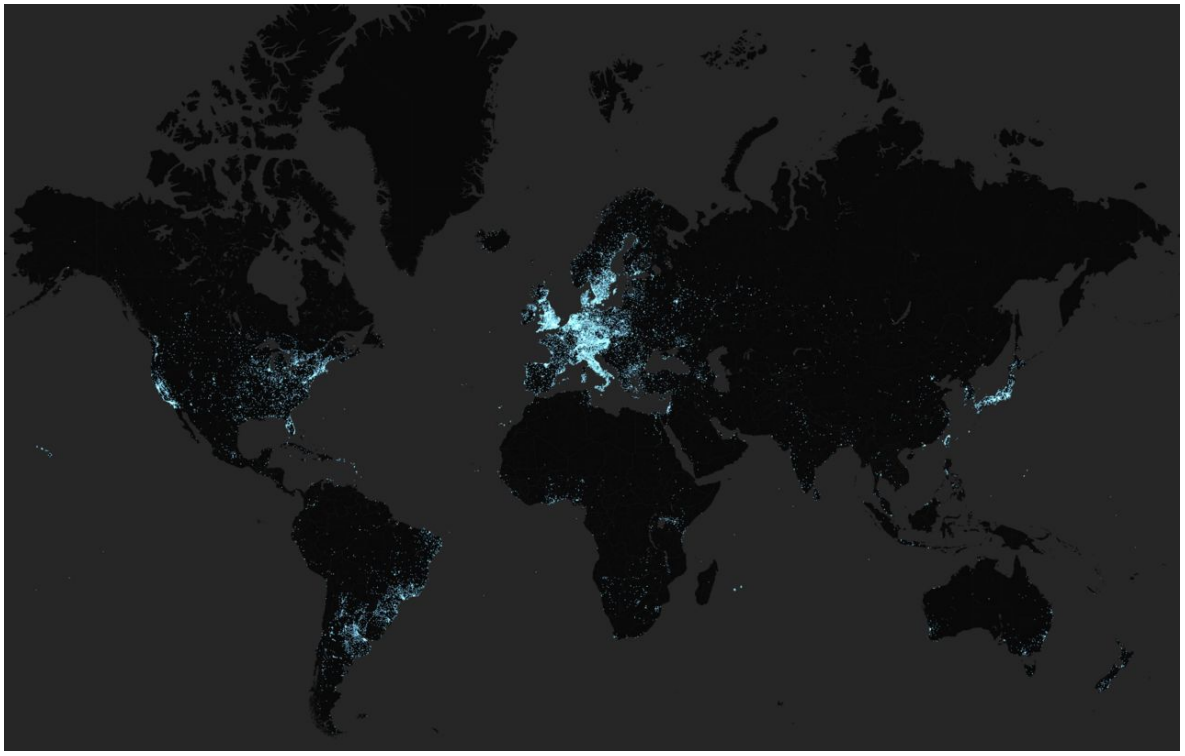
Organizations involved: UNESCO, Archives Portal Europe

Wikimedia/free knowledge communities involved: Wikimedia Sverige, Wikimedia Foundation

Keywords: GLAM, Wikidata

Key conclusions

- Wikidata is a suitable platform to develop a global database of cultural heritage institutions, as it is free, structured, multilingual and open to edit for everyone.
- By surveying the availability of GLAM datasets, we have brought the community's attention to the wealth of pre-existing data, which – even if it cannot be imported to Wikidata due to copyright restrictions – can be used e.g. as sources in Wikipedia articles.
- Due to its scope and flexibility, Wikidata has a steep learning curve. In order to recruit new users more easily, the development of user-friendly tools is crucial.
- Communication and documentation are important tools in reaching out to Wikidata editors. In our project, they helped us make clear why work on GLAM data is important, and how to start contributing.



GLAMs on Wikidata (October 2019). John Cummings, CC-BY-SA-4.0, via Wikimedia Commons.

See the updated map in full resolution on Carto.

Background

Having an overview of the cultural heritage institutions of the world is important to help them gain protection through visibility. In many natural and man made disasters, cultural heritage institutions have been damaged, lack of basic information (e.g. their location) has made response more difficult. This project will work towards helping solve this problem. Making this information visible and explorable will let more people learn about cultural heritage from many different cultures, it will create new insights and new knowledge.

The Wikimedia platforms are used and edited around the world – as of February 2020, Wikipedia has 299 active language versions. On Wikimedia Commons, photos, artworks and other media sourced from and related to cultural heritage institutions are collected. Wikidata, a free and open database, has a flexible structure that is well suited for storing data about everything under the sun, which can easily be edited, queried, analyzed and re-used. All of this created by volunteers and available to everyone, for free. In short, the Wikimedia platforms are the obvious choice for creating a truly global, multilingual, accessible and free database of cultural heritage institutions.

Wikidata contains over 150,000 Wikidata items representing the world's cultural heritage institutions (galleries, museums, libraries and archives). While it's hard to say how this corresponds to the total number of GLAMs around the world, it is clear that much remains to be done. For example, IFLA estimates that there are over 2.5 million libraries in the world;[21]

[21] https://librarymap.ifla.org/

there are 97,188 on Wikidata.[22] At the country level, it becomes obvious not only how much data is missing, but also that the degree of coverage differs greatly across nations: Kyrgyzstan, Turkmenistan and Madagascar are represented with 7, 9 and 10 GLAMs respectively, despite each country boasting millions of inhabitants.[23] Even without official data about the number of cultural heritage institutions in those countries, it is easy to see that much work remains to be done if Wikidata is to show even a moderately accurate picture of Asia and Africa.

Problem

The focus of this case study was adding and improving data about the world's cultural heritage institutions to Wikidata, as well as researching and evaluating different ways of doing that. We were particularly interested in investigating how to recruit new Wikidata contributors from the GLAM sector, as they have both professional expertise in the topic and an inherent interest in sharing their knowledge with the general public.

Implementation

Dataset index: surveying what we know

Many datasets of cultural heritage institutions already exist. They differ greatly in format, scope, completeness, verifiability and copyright terms. Some of them can be uploaded directly to Wikidata – as long as they are in some machine-readable, structured format and covered by a Wikidata-compatible license, such as CC0 or Public Domain. But even if they cannot be imported en masse, they are still valuable to Wikimedians: they give us an insight into areas of the world where Wikidata lacks data, and can be used as reference. That's why throughout the whole project, we have been compiling an index of GLAM datasets.[24] It combines our own research with the findings of Wikimedians around the world, who have contributed with their local knowledge and language skills.

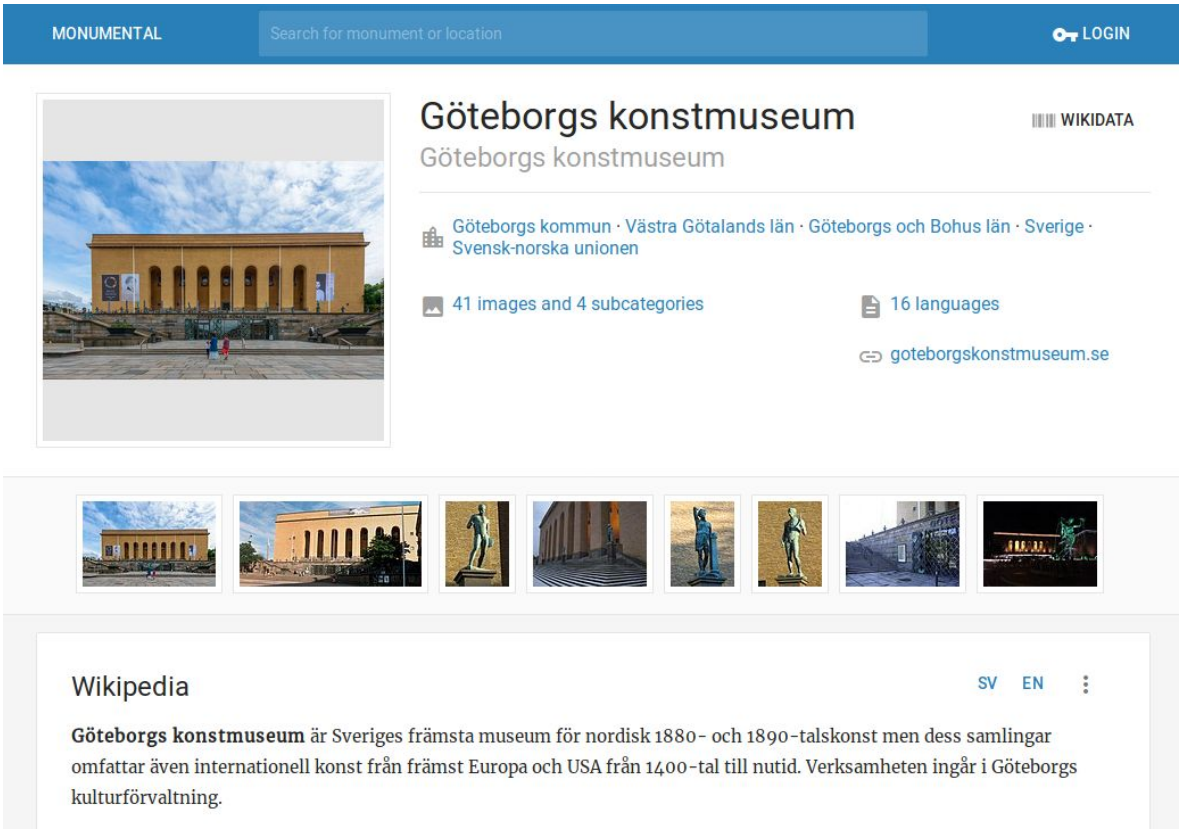
Data uploads: every little bit helps

We imported a small part of the identified datasets to Wikidata. The selection of datasets to upload was fully determined by their copyright status; while it is common knowledge that Wikidata editors do not always agree on what data is kosher to upload, we stayed on the safe side and only uploaded datasets that were clearly licensed CC0 or public domain.

OpenRefine was a crucial tool in the upload process, due to its flexibility; our data came from different sources and in different formats, but could all be explored and normalized in OpenRefine. The robust reconciliation capabilities were key to matching the data to existing Wikidata items and avoiding creating duplicates.

[22] https://w.wiki/JgN
[23] https://meta.wikimedia.org/wiki/FindingGLAMs/GLAM_statistics
[24] https://meta.wikimedia.org/wiki/FindingGLAMs/GLAM_datasets

FindingGLAMs Campaign: developing user-friendly tools for newcomers



The prototype of Monumental, showing information about a cultural heritage institution. CC-BY-SA-4.0, via Wikimedia Commons.

While uploading datasets is an efficient way to quickly increase the number and quality of GLAMs on Wikidata using reliable sources, we cannot only rely on it if we want to achieve our goal: putting all the world's GLAMs on Wikidata. Many of the published datasets are copyrighted and thus cannot be copied en masse. More importantly, such datasets do not exist for every country and every cultural sector in the world. This is where the power of Wikidata – its community – really comes into play. Every day, Wikidatans manually create, review, update and enrich thousands of articles. Just like Wikipedia, Wikidata exists thanks to millions of individual contributions.

Increasing the number of editors interested in cultural heritage is thus one way of coming closer to our ambitious goal. That's why we had an idea to make it easier for newcomers – especially GLAM staff – to take their first steps on Wikidata. We know from our experience in education that Wikidata has a steeper learning curve than e.g. Wikipedia. Newcomers have to internalize a lot of information before they feel comfortable editing: how the data in Wikidata is structured, how to filter and find data using the Wikidata Query Service[25] (which requires learning at least the basics of SPARQL), and most importantly – how the particular

[25] https://query.wikidata.org/

types of items they are interested in is normally modelled. In particular the last element is often not documented well, if at all; active editors rely on their experience and knowledge of unwritten rules. Even something as simple as finding the right place to ask for help might be difficult.

As Wikidata can be edited remotely via an API, we decided to develop a user-friendly web application specifically for browsing, displaying and editing data about cultural heritage institutions. The application would enable users to learn the basics of Wikidata editing without having to interact with its actual interface. Most importantly, it would provide a GLAM-specific editing form, presenting the user with some fields commonly used in GLAM items, such as administrative location, geographical coordinates, social media handles, etc. On Wikidata, one has to know which properties to choose and where to find them; in our application, they users would be guided along the way.

We hired a contractor to develop the application and the first usable version, named Monumental, was ready in May 2019.[26] This version prioritized libraries in Sweden, and we tested it with a number of Swedish librarians. Our goal was to have the application finished in late 2019 or early 2020, and to use it in a global campaign aimed at GLAM staff, especially in regions with little GLAM coverage on Wikidata.

Due to reasons beyond our control, the development of the tool could not be finished in time, and we had to revise our plans. Instead of a campaign aimed at GLAM staff, we decided to run a competition for the Wikidata community.

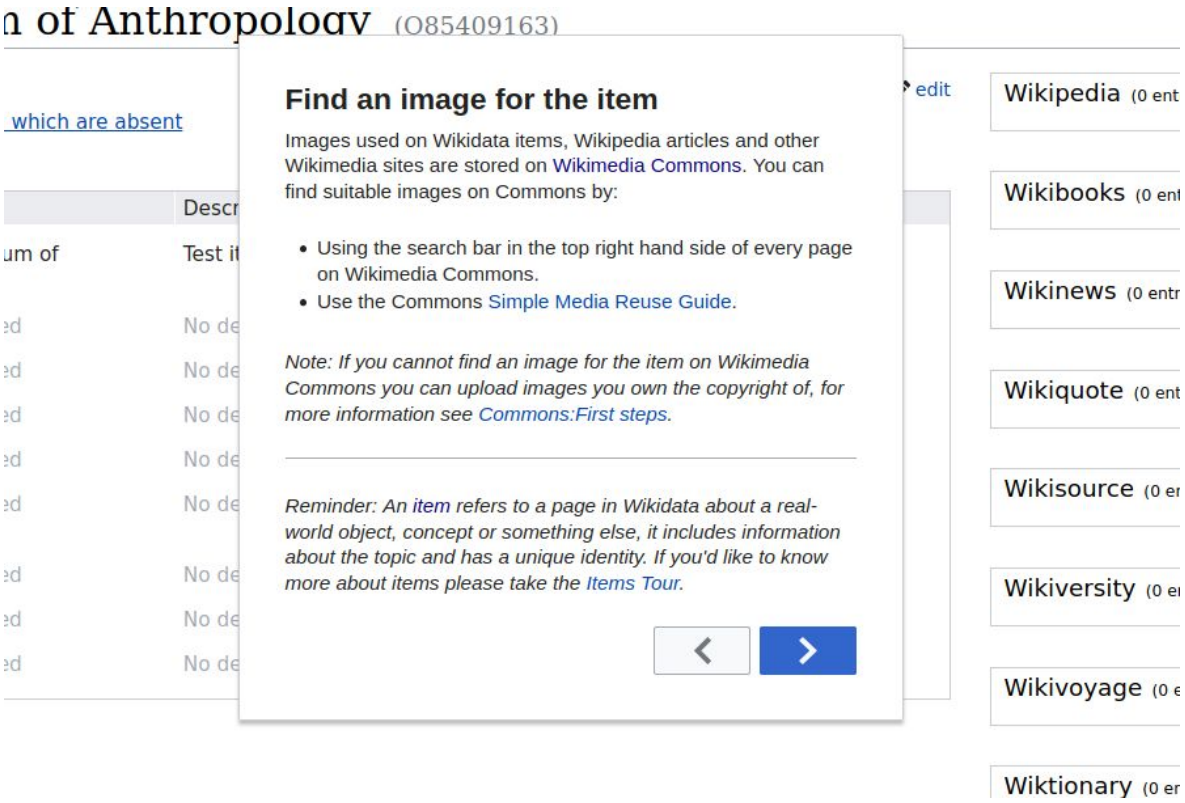
FindingGLAMs Challenge: editing together

Competitions can be a good tool to highlight particular areas in need of improvement and engaging the community in focused work. We have previously run competitions such as the WikiGap Challenge and the UNESCO Challenge, so we know that gamification works. As our original plan to recruit new editors using a newly developed tool fell through, we decided to instead run an activity aimed at the international Wikimedia community: the FindingGLAMs Challenge.[27]

The goal of the Challenge was simply to add as much information to Wikidata items of GLAM institutions as possible. Our intention was to engage community members in editing data in this area, hoping to increase their awareness of how much data is missing and that every single contribution counts. We also wanted the Challenge to be a memorable finale of the whole FindingGLAMs project, emphasizing that without international collaboration, we would not have been able to achieve what we did.

[26] https://tools.wmflabs.org/monumental-glam/#/
[27] https://meta.wikimedia.org/wiki/FindingGLAMs_Challenge

Wikidata Tours: taking the first steps



Wikidata Tours, an interactive tutorial introducing new users to Wikidata. CC-BY-SA-4.0, via Wikimedia Commons.

Documentation is a crucial element of every platform for user participation. This is especially true in the case of Wikidata, which, as mentioned previously, can feel intimidating to newcomers. However, despite the complexity of the platform, it is possible to start making valuable contributions after learning the very basics.

The Wikidata Tours[28] distill the absolute basics of editing into a series of short, interactive tutorials aimed at empowering beginners to take their first steps. They were first created back in 2014, but were not very detailed, offering only a brief introduction to statements and items – the building blocks of Wikidata. Furthermore, the software they were built on contained bugs, making it impossible for volunteers to improve the tours or create new ones.

We collaborated with the developers to find those bugs and have them fixed, and then created six new tours, focusing on small concrete tasks that editors of GLAM items might want to do, such as adding the geographical coordinates and location of an institution, or linking to a relevant photo on Wikimedia Commons. The tours can be edited and improved by volunteers, and they can also be translated into other languages.

[28] https://www.wikidata.org/wiki/Wikidata:Tours

The tutorials are linked from the main page of Wikidata, making them easily accessible to newcomers. Thanks to our contribution, many people's first experience of editing Wikidata has hopefully been made more positive.

Communication and building awareness

The FindingGLAMs idea cannot, by definition, be realized by a single working group or organization. It requires that people from many different countries collaborate, contributing with their local knowledge and language skills, helping data owners share their resources with the community and engaging volunteers. That is why we put a lot of effort into building awareness of our project, in hopes of sparking a fire that will burn long after the project has formally ended. We used our contact network to speak directly to representatives of other Wikimedia affiliates, as they have the resources to spread the FindingGLAMs message locally.

We participated with posters and talks in several conferences, of which Wikimania 2019 in Stockholm, Sweden, deserves a special mention, as it gave us a unique opportunity to address some of the world's most enthusiastic and knowledgeable Wikimedians face to face. We were also given the chance to present our project to Ambassadors to UNESCO and other delegation staff. This gave us the opportunity to highlight the importance of the Wikimedia projects for the world's cultural heritage institutions and encourage them to share open data from their own countries.

Outcome

Dataset indexing and uploads

In our dataset index, we collected information about 67 datasets from 44 countries. The project revealed significant discrepancies in access to GLAM datasets around the world; for example we only found one dataset covering Africa, and it was created by researchers aiming specifically to improve the very poor situation on the continent in this respect.[29] Europe was the continent with the largest number of datasets, which aligns with our experience in working with and educating about open data.

The vast majority of the datasets do not have a license compatible with Wikidata. That's why we only could process and upload a small number of datasets. Despite this limitation, we managed to add a significant amount of data to Wikidata by editing over 38,000 items, most of which were created from scratch. Possibly the most interesting of the datasets was the US Public Libraries Survey[30], comprising data about 9,000 public library systems and 17,000 individual library outlets – a good picture of the American library landscape. The majority of those had not existed on Wikidata previously, despite the US being privileged as a first-world country with many Wikimedians and data sources.

[29] https://osf.io/entpf/
[30] https://www.imls.gov/research-evaluation/data-collection/public-libraries-survey

More GLAM data under an open license

One of the successful outcomes of the project was facilitating the release of a dataset of European archival institutions under an open license. The dataset, owned and developed by Archives Portal Europe[31], was under copyright when we included it in our dataset index. We found the data very interesting and valuable for our project, as many of the described institutions did not have Wikidata items. Since the goal of Archives Portal Europe is to provide information about archival collections under the CC0 license, we thought they might be willing to also release their institution directory as open data. We reached out to them directly, informing them about Wikidata and our work. The response was positive, and a couple months later most the data was made free – after the relevant data providers had expressed their agreement.

What was particularly interesting about this case is that Archives Portal Europe themselves did not have the power to release the data under an open license. The data about the institutions was provided by their regional partners, and it was up to each and every one of them to decide whether an open license could be used. That is why the process took a long time – about half a year from the first contact. Nevertheless, the positive outcome shows that active work towards making data open does pay off. Thanks to Archives Portal Europe serving as a hub for European archives, institutions from several countries have been informed about the value of Wikidata and Linked Open Data. Hopefully this will build a foundation for future collaboration and provide a model for coordinating organizations to act as an internal champion for license change.

FindingGLAMs Challenge

The week-long FindingGLAMs Challenge was advertised in social media channels with an international audience, such as Wikimedia- and GLAM-focused Facebook groups and Twitter. We also contacted the representatives of other Wikimedia organizations directly so that they could pass on the information to their local communities using their prefered channels and languages.

The interest in the Challenge exceeded our expectations; 90 participants signed up, of which 53 did at least one edit to a GLAM item on Wikidata. 9 participants did at least 500 edits each; 21 participants, that is nearly half of the active participants, did at least 100 edits each. In total a staggering 19,200 improvements were made in a single week of the Challenge.[32]

The structure developed for the Challenge can be easily reused in the coming years and the activity can hence be re-organized for a very low cost. Furthermore, our experience and expectation is that the engagement is only likely to grow when repeated.

Future

The FindingGLAMs project had a very limited timeline and resources considering its ambitious scope. We knew, obviously, that putting all of the world's millions of GLAMs on the

[31] https://www.archivesportaleurope.net/directory
[32] https://outreachdashboard.wmflabs.org/courses/Wikimedia_Sverige/FindingGLAMs_Challenge/home

map over a year and a half would be impossible. Even though the project has formally ended, it was our intention that it would become the first step to future work – done not only by us, but also by other Wikimedia organizations and volunteers around the world.

The dataset index can continue to be maintained by the community. More importantly, it can serve as a starting point for Wikimedians to find trustworthy information about cultural heritage institutions and for Wikimedia affiliates to reach out to data owners to share their resources under open licenses. The awareness we have built around the project – and, more generally, around the important role the Wikimedia platforms play in engaging, bringing together and educating about cultural heritage institutions – will hopefully lead to more Wikimedians looking out for free GLAM data and supporting each other in including it on Wikidata, Wikipedia and Wikimedia Commons.

The data we uploaded to Wikidata will continue to increase in value every time it is viewed, queried and edited. Just like every other item on Wikidata, it can be improved by anyone. Errors can be corrected, outdated information can be updated, labels and descriptions can be translated to additional languages, GLAMs can be photographed and Wikipedia articles written and linked. The FindingGLAMs Challenge showed that there's a lot of interest in improving data about GLAMs on Wikidata – people clearly care about their local cultural heritage institutions.

While it is regrettable that we did not get an opportunity to launch the full version of the Monumental software, the early tests and the discussion with the community members were very promising. They made it clear that a tool like Monumental is necessary if we want to recruit new editors. Wikidata has a learning curve; many of its most active users have a background in databases, computing or at least a long experience with other Wikimedia platforms. This is not something that can be expected from new editors if we want to increase their number – and we absolutely do if the goal of finding all of the world's GLAMs is to be achieved. We hope that our experience with Monumental will spur research into, and development of, editing tools not primarily aimed at existing Wikidata editors.

Case Study 2: SANG – Sharing Audio files and Note sheets Globally

Key facts

Time: April – May 2019

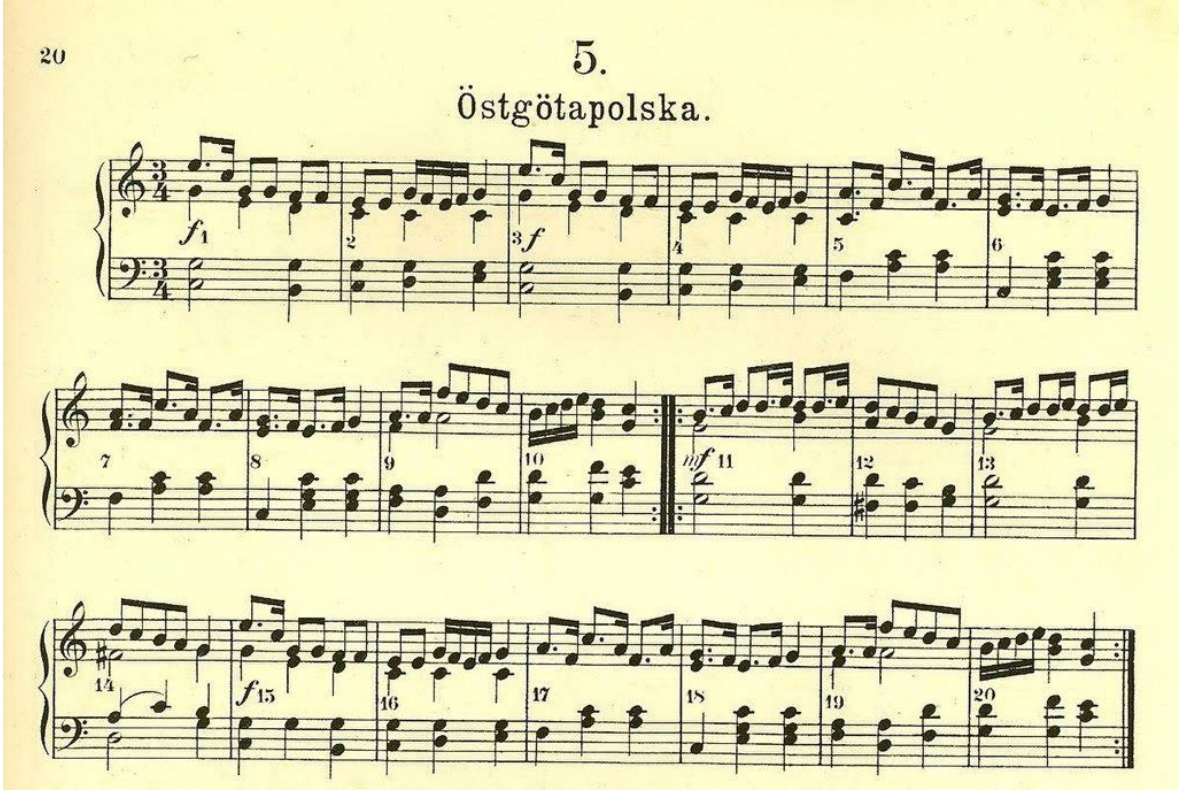
Organizations involved: The Swedish Performing Arts Agency (*Musikverket*)

Wikimedia/free knowledge communities involved: Wikimedia Sverige

Keywords: music, Wikimedia Commons, sheet music, Optical Music Recognition

Key conclusions

- We can to a certain degree meet the needs of GLAM institutions who want to share multimodal material on Wikimedia Commons.
- Audio files are underrepresented on Wikimedia Commons in terms of number and infrastructure, making it hard to find good examples to follow. More good examples are essential to encourage editor activity.
- There is potential to make working with sheet music on Wikimedia Commons more rewarding, but it needs both documentation and good examples.
- There is a treasure trove of unique heritage in GLAMs in the form of audio files that should be included onto the Wikimedia platforms to enrich the articles on Wikipedia.
- By connecting different types of material from GLAM institutions through structured data added value for both the partner organizations and the end users can be created.



The song *Östgötapolska*. Abr. Hirschs förlag 2237, Public Domain, via Wikimedia Commons.

Background

The Swedish Performing Arts Agency (*Musikverket*) is a government agency promoting music and related art genres (theatre, opera, dance, etc.) in Sweden. They have a specialized library and archive documenting the history of music in Sweden.

We had worked with Musikverket previously, uploading digitized archival photographs to Wikimedia Commons and organizing Wikipedia writing workshops for the staff. Because of that, we already had established contact with staff members who were knowledgeable about the Wikimedia projects and the practicalities of preparing material – most importantly for this project, the copyright requirements. We reached out to Musikverket with information about FindingGLAMs, indicating we would like to work with material that is more complex than the photographs we had worked with previously, especially something involving several different types of media.

The provided material consisted of audio files (MP3), images (TIFF) and digitized sheet music (PDF). The theme was Swedish music history and the selection of material was done entirely by Musikverket staff. They selected resources that highlighted the variety of material the organization worked with, with an emphasis on unique collections, such as the joik collection.

Problem

GLAM institutions often have different types of media (audio, visual, text) related to the same subject. In a physical setting, such as a museum exhibition, ties between the different media types can be created in order to place the topic in a wider context, to deepen the viewer's understanding and to increase the emotional impact of the material.

On Wikimedia Commons, visual material, such as photographs and artworks, is the main focus. It is reasonable to claim that the majority of the users – both directly on Commons, but also on the other Wikimedia platforms, such as Wikipedia – also focus on the static visual content, and are not as familiar with other media formats as audio or video. It is a challenge to examine how other types of media can be organized on Wikimedia Commons in order to offer users a more interesting experience while at the same time creating a more fair representation of GLAM collections. Indirectly, this may lead to the user becoming more interested in what the particular GLAM has to offer.

That's why in this project, we focused on examining how multimodal material from a GLAM collection related to the same topic can best be organized on Wikimedia Commons. Our central idea was to both maximize the benefit for the user and create a richer, more nuanced representation of the collection.

Implementation

The upload consisted of three different media formats: audio, images and sheet music.

In the audio upload, a major problem that we encountered was that due to the comparatively low amount of audio files on Wikimedia Commons, the existing infrastructure felt severely underdeveloped in comparison with the infrastructure for images. For example, there is only one information template for audio files[33], while there are many templates derived from *Artwork* or *Photograph* templates. This means that Commons editors have not felt the need to create more specific templates for audio, which might reflect a low level of interest in this type of media. Consequently, editors who want to edit and create audio-related informational templates do not have many examples or documentation to learn from. In order to present information about the audio files, we created the template Musikverket-audio[34] which can be used with files provided by the GLAM institution.

The reason why so many information templates exist is to make it easier both for the editors and the readers. A template designed specifically for a certain collection or media genre provides fields that are especially useful for this subset of files, as well as display some information without the editor having to input it themselves. For instance, the BASA-image template[35], designed for use with files shared by the Bulgarian Archives State Agency, automatically adds the institution's informational banner, as well as prompts the editor to input some accession details specific to this institution. The editor does not risk forgetting to include this information, neither do they have to worry about making it look good to the

[33] https://commons.wikimedia.org/wiki/Template:Musical_work
[34] https://commons.wikimedia.org/wiki/Template:Musikverket-audio
[35] https://commons.wikimedia.org/wiki/Template:BASA-image

reader; the template takes care of the formatting, ensuring a consistent look across the collection.[36] Thus, informational templates are a helpful tool that both empowers editors and makes the data more legible.

Another aspect that heavily influenced our work was the quality of the metadata we received from the GLAM. The metadata contained the basic information about each soundtrack, such as the creation date, title and where and by whom it was performed. However, the information was not very well structured. For example, there were no links to authority posts for the involved people, despite the fact that Musikverket maintains its own authority database. We were informed that the audio database is not connected to the authority database, which made it difficult for us to place the files in relevant categories, as those had to be identified manually. In addition the metadata lacked structured information about the type of the music, such as classical music, piano music, etc. Again, this made it impossible to identify relevant categories automatically, as the only way to do that was by reading the title manually and making assumptions; something that requires good knowledge about music and introduces the risk of making errors due to the uploader's bias. As a Wikimedia organization, we believe our role is to serve as a neutral middleperson between the GLAM and the Wikimedia platforms; if we make guesses on e.g. what music genre a file belongs to that are not based on the metadata (or other information provided by the GLAM), it can make the community believe that the information originated from the GLAM and is thus correct and reliable.

The audio files were uploaded using a Python script developed specially for this collection[37], which in turn leans heavily on a reusable library[38] which we had created previously. The library contains functions for building description templates for Wikimedia Commons files and uploading the files together with the descriptions. We have used this library multiple times for Commons uploads. The specially created script is custom for this collection, and could only be re-used for a collection with exactly the same metadata format, i.e. coming from the same institution/database. Building tools that are universal and can be used directly with any institution's metadata with as little adjustment as possible is a big challenge that affects content partnerships – both the speed at which data can be processed and materials uploaded, and the easiness and accessibility of the process. The reason why we used custom scripts for this upload is that we had previously worked with metadata from Musikverket's database and thus could perform the upload by modifying older code. Otherwise we would have used the more flexible tools available today, such as OpenRefine for processing the metadata and Pattypan for uploading the files to Wikimedia Commons.

The image upload consisted of a dozen files which, according to the GLAM staff who selected them, were "related to the audio files in some way". Those were not accompanied by machine-readable metadata at all; all information we had were the filename (e.g. Etikett) and the inventory number of the related audio file. Because of that, it was impossible to categorize the files, as we simply did not know what was shown in them.

[36] https://commons.wikimedia.org/wiki/File:BASA-RS-59-1-296-Nikola_Obretenov,_1890.jpg
[37] https://github.com/Vesihiisi/musikverket_batches
[38] https://github.com/lokal-profil/BatchUploadTools

The image files were uploaded manually, due to the small number of files as well as the poor metadata accompanying them, which made it inefficient to create a special script to process such a small amount of data.

The advantage of there only being a handful of image files is that it was easy to link them to the corresponding audio files. We did it by adding <gallery> tags in the *Related files* section of the infobox template, which is a common way of linking files that have strong ties to each other (see for example how it can be used to direct the user's attention to the other side of a photograph[39]). That was done in both the audio and the image file. That way, when looking at an image file, the user can click on the play button and listen to the music track. And the other way round, on the audio file's page, the user can easily see the related image files. The scanned sheet music files were processed in the same way, i.e. linked to the music files just like the other image files. However, due to their nature they presented an interesting conceptual challenge, which is why they deserve separate treatment.

Sheet music[40] is a way of representing sounds using special notation, much like text is a way of representing language using the alphabet. Musicians can read the notation and reproduce the music based on it. Since the purpose of the notation is to convey the same information to different people, it must be standardized and objective – again, just like text. There should exist a way for computers to interpret the notation and play music based on it – and that such a solution should exist on Wikimedia Commons, so that users can listen to all the openly licensed sheet music that has been uploaded there.[41] Ideally, it would be implemented in a user-friendly way, such as a button next to the sheet music image. We did not find such a solution so we investigated whether it would be possible to create it. This required exploring the technical aspects of the stages necessary to make it possible.

First of all, the sheet music image would have to be converted to meaningful musical characters, just like a scanned text has to be OCR'ed[42] in order to be read out loud by a computer. We found an open source tool that did exactly that.[43] The tool can produce output in the MusicXML format, which in turn can be converted[44] into LilyPond[45] – a format that to some degree is already used on Wikimedia Commons and Wikidata.

We then studied the material on implementing LilyPond and musical scores on Wikimedia Commons and Wikidata. We found that intensive discussions have been held on the topic of musical notation files on Commons, with arguments for against different formats.[46] The advantage of LilyPond is that it is supported on MediaWiki through the Score extension; musical code in the LilyPond format can be displayed as a PNG image and played as an

---

[39] https://commons.wikimedia.org/wiki/File:MC,_Radio,_information,_education_-_UNESCO_-_PHOTO0000003514_0001.tiff

[40] https://en.wikipedia.org/wiki/Sheet_music

[41] https://commons.wikimedia.org/wiki/Category:Sheet_music

[42] https://en.wikipedia.org/wiki/Optical_character_recognition

[43] https://github.com/Audiveris/audiveris

[44] http://lilypond.org/doc/v2.18/Documentation/usage/invoking-musicxml2ly

[45] https://en.wikipedia.org/wiki/LilyPond

[46] https://commons.wikimedia.org/wiki/Commons:Village_pump/Proposals/Archive/2018/11#RfC:_Musical_notation_files

audio file, which is close to what we imagined as a good solution.[47] This extension has documentation on Commons,[48] and a template exists to make it easier for the user to enter LilyPond data.[49] At the point of writing, the category *Images with LilyPond source code* contains around 100 files.[50]

These findings indicate that processing sheet music files would encounter several challenges. First of all, it is unclear how accurate music transcription software is, especially when dealing with less than perfect quality scans (as in our case). To make a comparison with OCR software, OCR'ed text e.g. on Wikisource undergoes proofreading and usually requires significant correction. It is reasonable to assume the same would have to be done with musical notation. The crucial difference is that the pool of potential proofreaders of text is comparatively large, as this task can be done by laypeople; no special knowledge other than a familiarity with the language is required. In order to both detect and correct mistakes in musical notation, one needs to be able to read and interpret the special characters, which is a special skill.

Furthermore, between the discussions on Wikimedia Commons and the small number of files using LilyPond code, it is our impression that there is no well established process on how to implement an idea such as ours. In fact, we did not succeed in locating LilyPond files on Commons that used the Score extension to actually play the sound. A typical file that we examined[51] contained both a visual representation of the music (a picture of musical notes) and a LilyPond source code, but no functionality to hear the sounds.

Outcome

The project resulted in 145 audio files[52] and 33 image files[53] uploaded to Wikimedia Commons. Due to the previously mentioned difficulties in categorizing the files automatically, we asked the Swedish Wikipedia community for help. We posted information about the upload[54] on the page *Månadens uppdrag* (the month's tasks), where Wikipedians can share tasks they would like to have help with. We specifically asked for help with categorizing the newly uploaded materials[55] on Wikimedia Commons.

At the time of posting the request, there were 118 files needing categorization. At the time of writing, this number has decreased to 100. Objectively this is a small decrease; based on our experience, the volunteer involvement has been smaller than it tends to be with more "typical" uploads, such as photographs or artworks. On the other hand, it might be that the

[47] https://www.mediawiki.org/wiki/Extension:Score
[48] https://commons.wikimedia.org/wiki/Help:GNU_LilyPond
[49] https://commons.wikimedia.org/wiki/Template:LilyPond_source
[50] https://commons.wikimedia.org/wiki/Category:Images_with_LilyPond_source_code
[51] https://commons.wikimedia.org/wiki/File:Augustan_Club_Waltz.pdf
[52] https://commons.wikimedia.org/wiki/Category:Media_contributed_by_the_Swedish_Performing_Arts_Agency:_2019-04
[53] https://commons.wikimedia.org/wiki/Category:Media_contributed_by_the_Swedish_Performing_Arts_Agency:_2019-05
[54] https://sv.wikipedia.org/w/index.php?title=Wikipedia:M%C3%A5nadens_uppdrag&diff=prev&oldid=45681199
[55] https://commons.wikimedia.org/wiki/Category:Media_contributed_by_the_Swedish_Performing_Arts_Agency:_needing_categorisation

material in this particular upload is extremely niche and by its nature only interesting to a small number of editors. This is reflected by the poor usage of the files; none of the audio files are used in Wikipedia articles[56] while two of the sheet music files are used to illustrate relevant articles in Swedish Wikipedia[57].

Apart from adding the categories, a small number of files have been edited by adding information, such as descriptions.[58] This provided added value, as the information could not have been added based solely on the metadata provided by the GLAM institution.

Future

This project demonstrates that there are significant barriers to fully use the potential of multimodal resources on the Wikimedia platforms. The following observations should be kept in mind both when working with audio material from GLAM institutions and developing new tools for GLAM uploads.

Firstly, and most importantly, a lot could be done on Wikimedia Commons to make it more suitable for audio and other non-visual content, from the point of view of contributors and users alike. After all, Wikimedia Commons has an educational mission: all the content should be “instructional or informative”.[59] As the majority of Commons users are not musicologists, we argue that contextless musical sounds are hardly informative. When one looks at a typical audio file uploaded as part of our project,[60] the screen is dominated by the file’s metadata, and the actual playing interface is hard to locate. The “call to action” – the “play” button – is small and unobtrusive. This makes it obvious that the interface was designed for visual files, which take a central position on the screen.

On the one hand, the priority given to images is not surprising. Wikipedia is a text based encyclopedia, and images can enhance the understanding of pretty much all of the articles, regardless of topic while audio arguably is not as applicable across the spectrum. On the other hand, today Wikimedia Commons is much more than that: a repository of free media, available for anyone to re-use and re-mix. Today’s Wikimedia Commons holds videos, music, 3D models and geographic data files. This variety should be reflected in the direction of its development. What **is** a good interface for music files? Maybe it should include a possibility for users and uploaders to curate collections of audio files played one after another, like a music tape, that could contextualize songs with a common theme. Or a tool to

---

[56] https://tools.wmflabs.org/glamtools/glamorous.php?doit=1&category=Audio_files_from_the_Swedish_Performing_Arts_Agency&use_globalusage=1&show_details=1&projects[wikipedia]=1&projects[wikimedia]=1&projects[wikisource]=1&projects[wikibooks]=1&projects[wikiquote]=1&projects[wiktionary]=1&projects[wikinews]=1&projects[wikivoyage]=1&projects[wikispecies]=1&projects[mediawiki]=1&projects[wikidata]=1&projects[wikiversity]=1

[57] https://tools.wmflabs.org/glamtools/glamorous.php?doit=1&category=Media_contributed_by_the_Swedish_Performing_Arts_Agency%3A_2019-05&use_globalusage=1&show_details=1&projects[wikipedia]=1&projects[wikimedia]=1&projects[wikisource]=1&projects[wikibooks]=1&projects[wikiquote]=1&projects[wiktionary]=1&projects[wikinews]=1&projects[wikivoyage]=1&projects[wikispecies]=1&projects[mediawiki]=1&projects[wikidata]=1&projects[wikiversity]=1

[58] For example, here volunteers added both relevant categories and information: https://commons.wikimedia.org/w/index.php?title=File%3AEtikett_-_SMV_-_78F_285A.tif&type=revision&diff=361703388&oldid=349568053

[59] https://commons.wikimedia.org/wiki/Commons:Project_scope

[60] https://commons.wikimedia.org/wiki/File:Till_trollkarlen_Ana_Vuolla_-_SMV_-_SVA_CYL_0272.wav

build slideshows with related images and texts that the user could look at while the music is playing; looking at a static page of metadata for several minutes might feel unsatisfying for anyone apart from the most hardened sound fans.

A large part of this case study was researching the available technologies for encoding and decoding music. Above, we identified software such as Audiveris, LilyPond etc. that has been developed to deal with different aspects of this problem. This indicates that open source developers have been actively working to make dealing with music on the computer easier. At the same time, understanding what possibilities there are and the functions and limitations of the different tools requires both research and specialist knowledge.

An ambitious idea would be implementing all these functionalities directly in Wikimedia Commons. A Commons user could upload a digitized sheet music, for example shared by a museum, and have it automatically converted to sounds using built-in Optical Music Recognition. Any errors generated in the automatic transcription process could be corrected by Wikimedians using a proofreading tool, akin to how Wikisource works as a platform for transcribing and proofreading text. Such a proofreading tool could either be available directly in Wikimedia Commons, or as an expansion of Wikisource. Furthermore, once the sheet music is converted to machine code, it could be connected to the Wikidata items of the music pieces. Applying the power of structured data to music would enable new research applications, empowering Wikimedia users to answer questions such as what types of sounds are most common in the works produced in a certain century, etc.

Furthermore, an uploaded sound file could be converted to musical notes automatically, using the same technology, but the other way round – this would be especially appreciated by people interested in music but affected by hearing loss, making the Wikimedia platforms more accessible. Also music educators could have use for such a solution, being able to generate sheet music from public domain recordings to share with their students.

An interesting aspect of this project for further development opportunities was working with poor metadata. On the one hand, rich and detailed metadata makes it easier for users to find files and benefit from them educationally. On the other hand, poorly described files can be improved by Wikimedians: more specific categories can be added and descriptions of what the files are depicting can be made more detailed and easier to understand. Experts in any field can be found contributing to the Wikimedia platforms. The improvements they make benefit primarily other users of the Wikimedia platforms, but they could also provide added value to the GLAMs who have shared the files in the first place. This is called data roundtripping[61] and we explore it in depth in Case Study 7 on the basis of research and pilot projects done in collaboration with several Swedish cultural heritage institutions.

The potential of using Structured Data on Commons (SDC) with audio files is something that we did not research at the time of the upload, but it might deserve further exploration as well. The work done on SDC so far, as well as the editor engagement, have been focused on image files, for obvious reasons (them being both prevalent on Wikimedia Commons and accessible to the vast majority of editors). As development is progressing and community standards around SDC for images are taking shape, investigating non-visual media is a

[61] https://meta.wikimedia.org/wiki/Wikimedia_Commons_Data_Roundtripping

logical next step. It might even increase the currently low interest in other media types and indirectly cause a surge of activity around them.

Finally, something that this project made very clear to us, is that active participation of domain experts in content partnerships is crucial. While we have a lot of experience with image uploads, we encountered difficulties in our work with the audio files. That was due to both insufficient metadata and our unfamiliarity with how music is described and categorized, and what technical possibilities exist for those who work with it. We believe that our role in content partnerships is to serve as a neutral middleman between the Wikimedia platforms and the GLAM. If we make decisions on e.g. categorizing the files based on our best guess rather than on the metadata, we risk introducing bias due to our lack of domain knowledge. This is something that Wikimedians who upload GLAM content should be aware of; sometimes it cannot be avoided, in which case it is important to know one's own biases and limitations. The best solution to this problem is to encourage active participation of GLAM staff in the upload process – which requires documentation, training and empowerment to help them gain confidence as Wikimedians, and developing the technosocial infrastructure of the Wikimedia platforms to be more user-friendly, robust and understandable.

Case Study 3: WORD – Wikimedia Organizes lexical Resources Digitally

Key facts

Time: Spring 2019

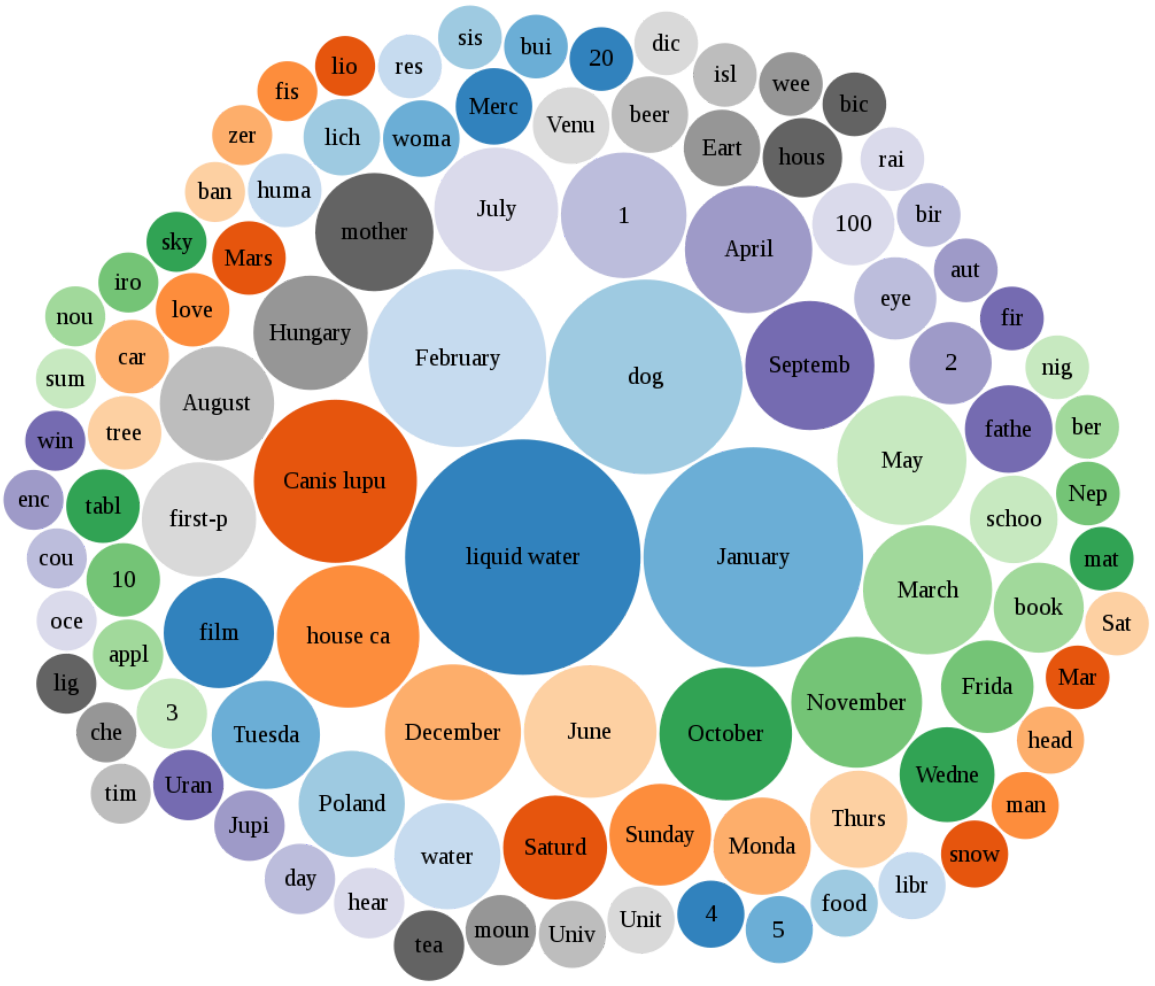
Organizations involved: Swedish Institute for Language and Folklore, European Commission

Wikimedia/free knowledge communities involved: Wikimedia Sverige

Keywords: Wikidata, lexicographical data, Wiktionary

Key conclusions

- There is a burgeoning interest from partners in lexicographical data on Wikidata.
- Many lexicographical data editors use external tools developed by volunteers, the emergence of which is caused by the lack of efficiency of the native interface.
- The existing popular tools for large-scale editing lack support for lexicographical data, which makes it harder for experienced editors to apply their expertise.
- Synergies between Wiktionary and Wikidata are possible and would benefit the communities by using volunteer resources more efficiently, but require that the possibility of linking between lexemes and Wiktionary entries is implemented. However, the different licensing terms of the two platforms are an obstacle to data re-use.
- The potential of the lexicographical layer of Wikidata as a multilingual dictionary is severely underutilized due to the lack of agreement on how to link between equivalent lexemes in different languages. A more straightforward technical solution could make this more user-friendly.
- The value added from developing a multilingual thesaurus, glossary and dictionary for the GLAM sector has been highlighted by multiple actors. This needs further investigation for a fruitful implementation.



The 100 most translated concepts using lexemes in Wikidata (February 2020). So9q, CC0, via Wikimedia Commons.

Background

The Swedish Institute for Language and Folklore (*Institutet för språk och folkminnen*) is a government agency that works with research and preservation of languages used in Sweden. In spring 2019, we had a number of meetings with them to investigate the possibilities of sharing one of their openly licensed lexicographic resources on the Wikimedia platforms. At that time, the lexicographical namespace on Wikidata had been available for about a year and had dozens of active users working on many languages. It was established enough for us to research its suitability for this task, but not as mature as at the time of writing (January 2020).

Staff from Wikimedia Sverige participated in a number of discussions and workshops organized by the European Commission around the next steps of financing the digitization of

Europe's cultural heritage.[62] The value of creating dedicated language resources for the cultural heritage sector was frequently highlighted. The hope is that efforts in the area would support cross-European efforts to digitize materials. This is considered of high importance to the European Commission and many industry experts.

The material

Among other things, the Swedish Institute for Language and Folklore publishes material, such as glossaries, for translators and interpreters between Swedish and other languages.[63] Many of them are based on the so-called *Basic glossary for interpreters* (*Basordlista för tolkar*), a list of vocabulary in the fields of civics, medicine and law that the Institute considers essential for interpreters who assist clients in the Swedish public sector. This document, covering over 6,500 words and terms, has been released by the Institute under the CC0 license, and shared with Wikimedia Sverige.

Every entry in the document consists at least of the following:

1. the lemma, e.g. *ackordsättning*,
2. the definition in Swedish, e.g. *fastställande av betalning i förhållande till mängden utfört arbete*, and
3. the domain tag, e.g. *arbetsmarknad-och-pension*, classifying the entry as belonging to the area “labor market and retirement”.

Some of the lemmas are proper names, e.g. of social institutions. Some of the entries contain additional information such as synonyms. The following thematic areas are covered: medicine and psychiatry (2,498 entries), law (1,535 entries), labor market and retirement (917 entries), social insurance (844 entries) and miscellaneous (622 entries).

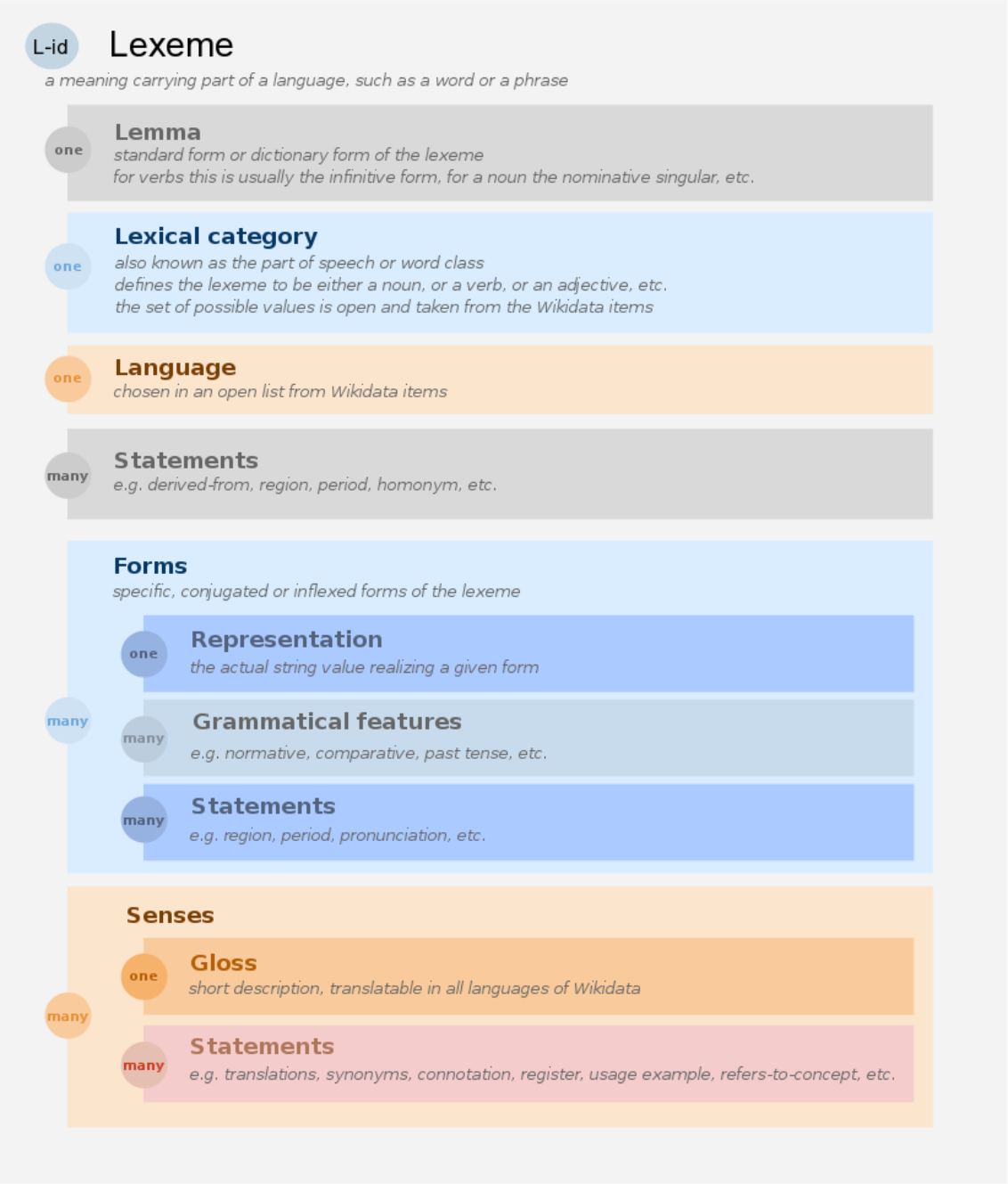
Lexicographical data on Wikidata

Lexicographical data[64] was introduced to Wikidata in May 2018. Since then, over 234,000 lexemes have been created, representing 350 languages. 12 languages have over 1,000 lexemes each. The largest languages are Russian (101,000 lexemes), English (38,000 lexemes), Hebrew (28,000 lexemes), Basque (18,000 lexemes) and Swedish (11,000 lexemes). As a point of comparison, there are over 78,000,000 items in the main Wikidata namespace.

In February 2020, the English Wiktionary had 512,042 gloss entries (basic forms, i.e. not inflected forms; in Wiktionary, inflected forms can have their own pages) in English.[65] It usually is the Wiktionary in a particular language (i.e. not the English Wiktionary) that boasts the largest number of entries in that language; Russian Wiktionary has 432,650 Russian entries[66] and Swedish Wiktionary has 78,194 Swedish entries[67]. This gives an indication of

[62] https://ec.europa.eu/digital-single-market/en/digital-cultural-heritage
[63] https://www.isof.se/om-oss/publikationer.html
[64] https://www.wikidata.org/wiki/Wikidata:Lexicographical_data
[65] https://en.wiktionary.org/wiki/Wiktionary:Statistics
[66] https://ru.wiktionary.org/wiki/Викисловарь:Статистика
[67] https://sv.wiktionary.org/wiki/Wiktionary:Om/Statistik

how much work remains for Wikidata editors to at least match the quantity of data in Wiktionaries.



Visualization of the Lexeme data model. Lea Lacroix, CC-BY-SA-4.0, via Wikimedia Commons.

Lexemes are the basic unit of lexicographical data on Wikidata, akin to how items are the basic unit in the main namespace.[68] Each lexeme has a unique identifier starting with the

[68] https://www.wikidata.org/wiki/Wikidata:Glossary

letter *L*, to distinguish them from main namespace items, whose identifiers begin with the letter *Q*.

A lexeme contains at least the following information:

- Language, e.g. English
- Lexical category, e.g. verb.
- Lemma, e.g. *distinguish*.

This data is necessary to input to create a minimum viable lexeme item. A lexeme can also contain more detailed information, including but not limited to:

- Grammatical gender, e.g. neuter or common for Swedish; neuter, feminine or masculine for Bulgarian.
- Forms, e.g. *achieve*, *achieves*, *achieved*.
    - A form can be tagged with its grammatical properties, e.g. achieves with simple present and third-person singular.
    - A form can be accompanied by a link to a recorded pronunciation on Wikimedia Commons.
- Senses (definitions) – one or more.
    - On the sense level, the lexeme can be linked to other lexemes, indicating semantic relationships, such as synonyms and antonyms.
    - The style of the sense can be included, e.g. to indicate that *ass* is a pejorative synonym of *buttocks*.
- Usage examples.
    - If the example comes from authentic material, such as a book or newspaper, it can – and should – be accompanied by a reference.
    - If the lexeme has several senses, the usage example can be linked to one or several of them.
- The component stems of compound words can be indicated, e.g. *foot* and *ball* for *football*.
- Links to entries in external services, such as external databases and dictionaries, can be supplied. None have been implemented for Swedish so far, but links in other languages include DanNet 2.2 (P6140) and Uralonet (P5902).

A more detailed description of the data model can be found in the official documentation.[69]

[69] https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/Documentation

Wikidata as a multilingual dictionary

Due to the international scope of Wikidata and its multilingual community, we find the potential to use the lexicographical namespace to build a translation dictionary very interesting. Wikidata – both the underlying software and the community – already has a heavy emphasis on internationality and multilinguality, as the labels and descriptions can be entered and displayed in any language.[70] This is one of the aspects of Wikidata that make it unique and powerful.

Translations constitute a major part of Wiktionary, but are not easy to handle for either editors or readers. Every language version of Wiktionary is independent, meaning that in order for a Swedish entry to include a translation to another language, the translation has to be added manually – even if the corresponding foreign language entry already contains a translation to Swedish. Such an edit in one language version of Wiktionary is not reflected in other language versions. New editors often find this architecture difficult to understand, as it requires duplication of work, putting a strain on limited volunteer resources.

Wikidata, with its centralized architecture, could be providing a solution to this problem. In practice, it does not do this in an efficient or user-friendly way. The lexeme namespace, which is under current development, at this stage does not yet seem to be designed with interlanguage links in mind. Most importantly, there is no standardized way to connect lexemes in different languages that share a meaning to each other. One way to do this is by using the property *item for this sense (*P5137) to link one lexeme's sense to a corresponding Wikidata item. For example, the linguistic concept *island* is linked to the encyclopedic concept *island*. This makes it easy to link multiple lexemes in different languages to the same encyclopedic item, indicating they are all expressions of the same concept. However, it is a rather blunt tool, best suitable for concrete nouns, such as *cat*, *knife* or *Sweden* (of course, translators know that great care must be taken in assuming perfect one-to-one relationships even between such words). It is less good at expressing subtle stylistic or dialectal differences, and even worse if applied to non-concrete vocabulary such as adverbs or conjunctions. For example, in Swedish the words *ja* and *jo* can both be translated as *yes*, but are used in different grammatical contexts and express different levels of agreement; a simple one-to-one relationship fails to represent the difference between them and could be actively misleading the user.

Another useful property is *translation* (P5972), which is used to link to a "word in another language that corresponds exactly to this meaning of the lexeme". This seems to be a fitting solution, but in practice it is not very useful. If the word in question has equivalents in 50 languages, then an editor would have to add all of them as values of this property. This is not practical or realistic; as a result, the property is used only 1,537 times – not a lot considering there are over 234,000 lexemes.

In conclusion, the available architectural solutions are unsatisfactory from the point of view of creating a multilingual dictionary, both for the editor and the reader and will need to be developed more to address these issues.

[70] https://www.wikidata.org/wiki/Help:Multilingual

Problem

Our goal was to investigate the viability of including the monolingual lexicographical data provided by the Institute for Language and Folklore on Wikidata, with a long-term perspective of also including multilingual data in a structured way to provide maximum benefit for the users. An important aspect of the case study is comparing the lexicographical layer in Wikidata with another Wikimedia platform, Wiktionary, which has an established position as a free dictionary platform and is presumably more familiar to the general public.

The identification and connection between material in different institutions spread across the world is currently very hard, partly due to language barriers. The structured lexicographical layer has the potential to simplify the process, while also improving automatic translation tools etc.

Implementation


```
   {
        "uppslagsord": "arbetarskydd",
        "definition": [
            "verksamhet för att förhindra och begränsa olycksfall och
yrkessjukdomar i arbetet"
        ],
        "doman": "arbetsmarknad-och-pension"
    },
    {
        "uppslagsord": "arbetsavtal",
        "definition": [
            "avtal om villkor för viss arbetsprestation gällande
anställning el. uppdrag"
        ],
        "doman": "arbetsmarknad-och-pension"
    }
```


Example of the data structure in the data shared by the Swedish Institute for Language and Folklore.

Using this background, we examined the existing possibilities for working with the dataset provided by the Swedish Institute for Language and Folklore on Wikidata.

At the time we carried out the case study, the majority of the 6,000 terms in the glossary could not be found on Wikidata. Many of them are specialized terms, such as *akrofobi* (acrophobia) or *etsskada* (a type of tooth damage), a vocabulary class that the editor community might not have prioritized, instead focusing on building up the core, everyday vocabulary. The interpreter glossary could thus make a valuable contribution to Wikidata.

The glossary contains definitions of all the words, which raises its value for the project. Definitions are a neglected area on Wikidata, partially due to the previously mentioned copyright ambiguities of Wiktionary. In the absence of importable definitions, Wikidata

editors have had to resort to writing their own, which takes time and requires lexicographical skills. Languages with comparatively few speakers and, consequently, few editors, such as Swedish, are at a disadvantage here.

The main challenge we encountered when working with the material provided by the Institute for Language and Folklore was that it lacked information about lexical categories. Those are necessary when creating lexemes on Wikidata; it is not technically possible to create a lexeme without including this information. Tagging the entries in the glossary with their lexical categories is a necessary prerequisite to importing it to Wikidata.

Some of the entries in the glossary are “phrases” in which a word is accompanied by its abbreviation, such as *barnavårdscentral, BVC* or *svenska som andra språk, sva*. Since a lemma can consist of several words and include punctuation (such as a phrase or a proverb), it is not possible to filter out this type of entries automatically. In this particular case, due to the small size of the dataset, it is possible to find them via manual examination and decide how to handle them on a case by case basis.

Outcome

Wikidata and Wiktionary

We found that the relationship between lexicographical data on Wikidata and Wiktionary, the free dictionary, is a topic of much discussion among the editor community. Wiktionary is an established platform with years of history. Many languages, including Swedish, have an active community which, despite not being as large as Wikipedia’s, have managed to build sizable and useful dictionaries. The Swedish Wiktionary contains over 77,000 entries in Swedish, as well as a large number of entries in other languages – mostly German, English, Finnish and French – with Swedish translations.

The main difference between Wikidata and Wiktionary is that Wikidata is a structured database, while Wiktionary stores all information about a word or a phrase in a single text field, not unlike a Wikipedia article. The structure of Wikidata enables analysis and complex searches, such as “find all Swedish verbs with usage examples sourced to books published in the 19th century” or “find all Swedish entries without a definition”. This is not possible in Wiktionary, at least not without downloading the contents and developing specialized analysis tools, something that in reality only researchers would be willing to do.

As a response to the “free” form of Wiktionary, its users have developed templates to display the content in a more structured and user-friendly way, akin to that of a commercial dictionary. Since all the content is free-text, editors are at a risk of making mistakes or accidentally violating community standards. It is important to note that since each Wiktionary is independent and governed by its community, the templates, guidelines and written and unwritten rules are different in each language version; an experienced Swedish Wiktionarian will face a considerable learning threshold trying to get involved in the English or German Wiktionary.

While data from the main Wikidata namespace can be displayed on Wikipedia, such a connection does not exist between the lexicographical namespace and Wiktionary. The potential for synergies between the two projects has been discussed by both communities, so it is not unrealistic to expect that such links will be enabled in the future. Information about the current technical situation regarding linking between the project can be found on the following pages: https://www.wikidata.org/wiki/Wikidata:Wiktionary and https://www.wikidata.org/wiki/Wikidata:Lexicographical_data/FAQ.

It is crucial to note that, for copyright reasons, it is not possible to import all the content of a language version of Wiktionary to Wikidata. Wiktionary is licensed under Creative Commons Attribution-ShareAlike (just like Wikipedia), a license not compatible with the CC0 license on Wikidata. The Wikimedia Foundation has published a preliminary perspective on the legal aspects of copyright for lexicographical data, with a US focus, concluding that definitions are creative enough to be copyrighted.[71] The resources that Wiktionary editors have spent years creating and thus of limited use for Wikidata editors.

Lexicographical infrastructure on Wikidata

The infrastructure for lexicographical data on Wikidata is not as mature as that for regular data. Because of that, some of the processes are not intuitive or comfortable.

One of the most tangible shortcomings is that searching for lexemes with the Wikidata search engine is not user-friendly. Firstly, it requires one to input a prefix to include the lexeme namespace, making it non-transparent to new users; secondly, it does not support filtering the results by language or part of speech, features one might expect from full-fledged dictionary software.

As a consequence, most of the active editors in the lexicographical namespace were established Wikidata editors before they started editing lexemes. They were already familiar with the platform and knew how to search for documentation. Many have been tracking the implementation and development of lexicographical data from the beginning. The learning curve for new editors without a background in Wikidata is steep, even if they have a background in linguistics or lexicography. Creating new lexemes using the native Wikidata interface is slow and requires a certain level of knowledge.

Because of these shortcomings the Wikidata community has developed a number of own tools for viewing and editing lexemes, e.g. Wikidata Lexeme Forms,[72] which enables users to quickly input several forms of a lexeme in pre-defined languages, without having to add the form tags by hand, like in the native editor.

Another area where the available tools are not satisfactory are large-scale automatic uploads. There are several toolkits for editing outside of Wikidata's interface and uploading data to the main namespace, which are used by both volunteers and organizations, including Wikimedia Sverige. These include Pywikibot[73] (a Python library for interacting with the

[71] https://meta.wikimedia.org/wiki/Wikilegal/Lexicographical_Data
[72] https://www.wikidata.org/wiki/Wikidata:Wikidata_Lexeme_Forms
[73] https://www.mediawiki.org/wiki/Manual:Pywikibot

Wikimedia projects), OpenRefine[74] (an application for cleaning, enriching and reconciling data) and QuickStatements[75] (a web-based tool to edit large numbers of Wikidata items in a batch). None of these tools, which are well established among main namespace editors, support lexicographical data. As a result, the many proficient users of these tools cannot apply their skills and experience to the lexicographical namespace.

Some users have developed their own tools to automate performing specific tasks on a large scale. These include LexData[76] and Lexicator[77], both developed in Python and available as open source. LexData is a small, language-independent framework designed to serve as a flexible base for scripts for editing lexemes. Lexicator is a tool to process grammatical data from – currently only Russian – Wiktionary and upload it to Wikidata.

The tools described above were created in order to do large scale mass uploads of lexicographical data. LexData is a generalized version of AitalvivemBot, built to work with French and Occitan lexemes from the Lo Congres database.[78] [79] [80] Lexicator was developed to import lexemes and their forms from Russian Wiktionary to Wikidata. Because of this upload, in November 2019 Russian was the largest language in the lexicographical namespace. The upload led to a discussion among the Wikidata community about how the previously described licensing differences between Wiktionary and Wikidata should be interpreted.[81] The conclusion was while the content of Wiktionary was nominally licensed CC BY, simple information, such as the individual lemmas and their forms, was so uncreative as to not being copyrightable in the first place, and thus should be possible to copy to Wikidata. The definitions, on the other hand, written by Wiktionary editors, do meet the threshold of originality and thus cannot be copied en masse into Wikidata without breaking the license terms. This is the same interpretation that the Wikimedia Foundation has reached.[82]

Future

Despite the fact that the *Basic glossary for interpreters* is in Swedish only, it is still very interesting to look at from the perspective of a multilingual dictionary. As mentioned previously, the Institute for Language and Folklore uses it as a base for their translation resources between Swedish and other languages. If those resources were released under an open license, it would give us an opportunity to do practical work with multilingual translation on Wikidata. That was a major motivation for us to examine Wikidata's lexicographical namespace from a multilingual perspective.

Our conclusion in this area is that there are no standards or best practices when it comes to linking between lexemes in different languages. To a considerable degree, this might be

[74] http://openrefine.org/
[75] https://www.wikidata.org/wiki/Help:QuickStatements
[76] https://nudin.github.io/LexData
[77] https://github.com/nyurik/lexicator
[78] https://github.com/aitalvivem/
[79] https://www.wikidata.org/wiki/Wikidata:Requests_for_permissions/Bot/AitalvivemBot
[80] https://locongres.org/
[81] https://www.wikidata.org/w/index.php?title=Wikidata_talk:Lexicographical_data&oldid=1052797489#Reminder:_your_input_needed_about_integration_of_Lexemes_in_Wiktionaries
[82] https://meta.wikimedia.org/wiki/Wikilegal/Lexicographical_Data

caused by insufficient technical infrastructure; as there is no obvious way to do this task, editors have had to discuss and develop their own solutions. This causes problems for anyone who wishes to query and re-use the multilingual data, e.g. researchers examining the relationships between lexemes in different languages. It also raises the barrier of entry for new editors, increasing the risk of them becoming frustrated with the experience and not being as active as they would like – a significant problem for the young lexicographical data community.

Editors should be able to link between equivalent lexemes in different languages in an easy and straightforward way. An inexperienced user might think that translation property is the right tool to use for this purpose, but they are not aware that they have to manually add it to all equivalent lexemes in other languages – including the lexeme referred to as the translation. This task could be done automatically.

The technical possibilities of large-scale editing are also something that should be actively developed. Right now, there are several options, with different scopes and capabilities. What they have in common is that they are developed independently by volunteers to fulfill their specific goals and interests. While those tools might be very good at performing their tasks, it does not change the fact that the lexicographical namespace does not have an equivalent of the powerful tools used by main namespace editors, such as QuickStatements, OpenRefine or Pywikibot. This widens the gap between the lexicographical namespace and the rest of Wikidata, making it harder for experienced Wikidata editors to apply their skills to lexicographical data. This gap could be eliminated by either implementing lexeme support in any of the mainstream tools – OpenRefine is a good candidate, as it is a robust, flexible and popular application used by both Wikidata editors and other data specialists – or developing a specialized tool that is similarly flexible and user-friendly, not requiring programming knowledge and well-documented so that the barrier of entry can be lowered and non-Wikidata editors with an interest in lexicography can join the project.

The lexicographical namespace of Wikidata is young and under development. While the Wikidata community has done a tremendous job creating the content from scratch, we believe that systematic, large-scale uploads of existing data, as well as collaboration with domain expert partners, are the key to making lexicographical data on Wikidata as exhaustive and relevant as the main namespace. By basing our examination on authentic material, we brought into the spotlight those aspects of Wikidata’s lexicographical namespace that are important from the point of view of partnerships and large scale uploads.

To develop dedicated language resources for the cultural heritage sector is likely to receive funding from the European Commission in the coming years. Wikidata has the potential to be a very suitable platform to combine these resources if it receives some further development.

Case Study 4: DOCS – Documents Obtained, Compiled and used for Sourcing

Key facts

Time: Fall 2019

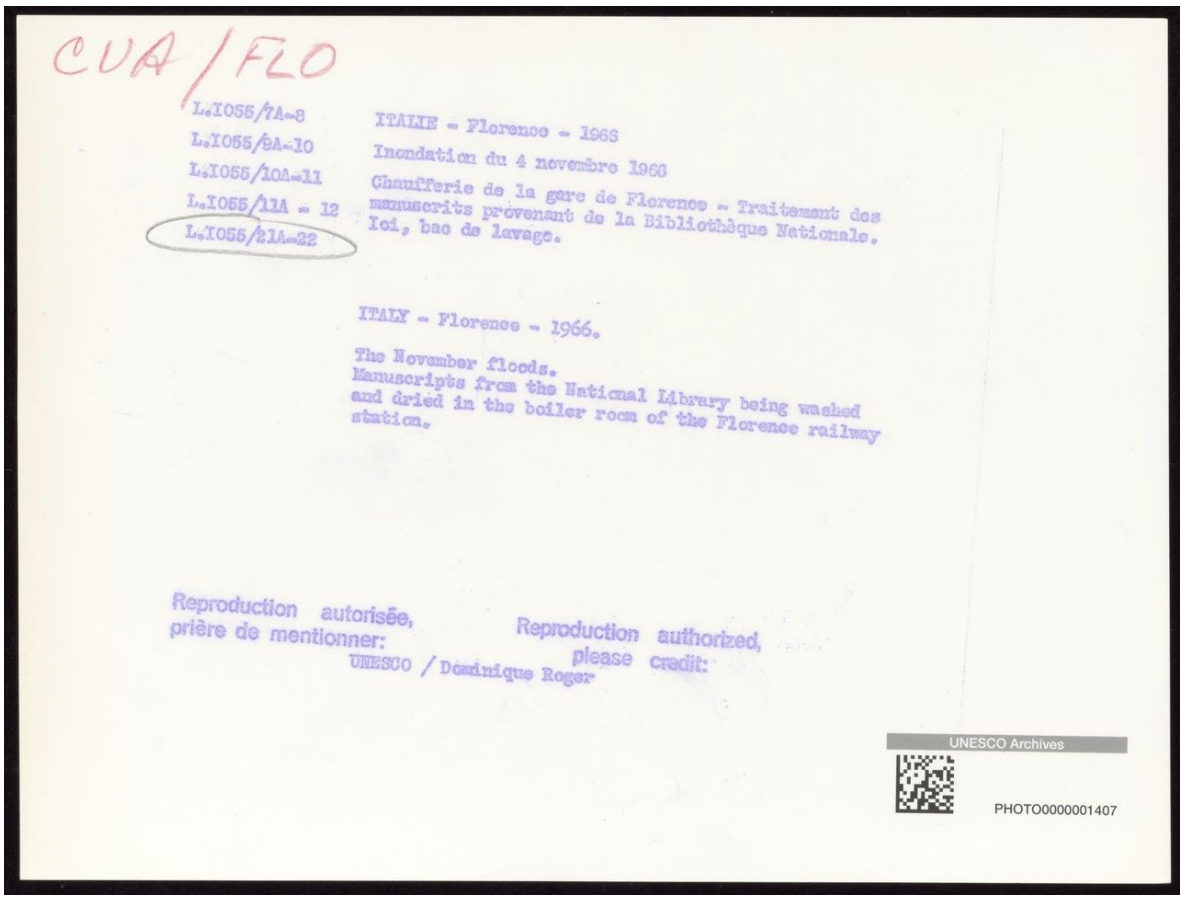
Organizations involved: Nordic Museum, UNESCO

Wikimedia/free knowledge communities involved: Wikimedia Sverige, Swedish Wikisource community

Keywords: Wikimedia Commons, Wikidata, Wikisource, digitalization, literature, copyright, archives

Key conclusions

- Uploading a GLAM collection to Wikidata and Wikimedia Commons can involve several different, independently developed tools, which are not naturally connected to each other. The uploader has to learn each of them in order to develop a functional working process.
- The process requires experience and knowledge of the Wikimedia tool ecosystem, including tools maintained by volunteers, which can make it inaccessible to new editors and GLAM staff.
- OpenRefine and Pattypan are powerful GUI-based tools that can be used in tandem in a Wikidata + Wikimedia Commons process, lowering the entry barrier for new uploaders.
- There is a very large potential value to create a clear path on how material uploaded can be (semi-)automatically uploaded to multiple specialized Wikimedia platforms and connected with each other. This path and the tools needed are still to be developed.



The backside of a photograph from the UNESCO Archives. UNESCO / Dominique Roger, CC-BY-SA-3.0-IGO, via Wikimedia Commons.

Background

A digitized ethnographic library reaches Wikimedians

The Nordic Museum (*Nordiska museet*) is a Swedish museum, located in Stockholm, dedicated to the cultural history of Sweden. Wikimedia Sverige has previously collaborated with the museum on projects ranging from uploading digitized artworks[83] to Wikimedia Commons to supporting high school students who edit Wikipedia using materials from the museum's library[84]. The museum staff who work with digital media have a keen interest in and familiarity with the open knowledge movement and our organization.

Fataburen is a journal that has been published by the Nordic Museum, for over 120 years. The museum had undertaken a digitization project to make it more accessible to the public, and had cataloged and published the articles in PDF format on the open publication platform DiVa.[85] Furthermore, high resolution individual page scans in TIFF format are stored in an internal database that is not available to the public. Since the digitization project covered the whole history of the journal, from its beginnings to the most recent volumes, a not

[83] https://commons.wikimedia.org/wiki/Category:Images_from_Nordiska_museet
[84] https://sv.wikipedia.org/wiki/Wikipedia:Projekt_GLAM/Nordiska_museet/Kulturhistoria_som_gymnasiearbete
[85] http://nordiskamuseet.diva-portal.org/

insignificant part of the articles were in the public domain due to more than 70 years having passed since the death of the authors.

In order to make the collection more known and accessible, the museum reached out to Wikimedia Sverige to help share Fataburen on the Wikimedia platforms. We judged the material to be highly valuable to the open knowledge movement, as the journal is a classic in Swedish cultural research, with many notable contributors over the years, and treats topics that are interesting to the general public (like folklore, life in the countryside, etc.) in an accessible way. In other words, we anticipated that the material could be put into good use by Wikipedians.

As we had access to both the scans and the metadata, it made sense to both upload the (public domain) scans to Wikimedia Commons and (all) the data to Wikidata. Furthermore, we envisioned that the collection could be of interest to the Swedish Wikisource community, as each article is short enough for one or two volunteers to proofread, and it would make them easier to use as sources in Wikipedia articles. That way, the museum would gain higher visibility for its publications, and at the same time the Swedish Wikimedia community would get access to valuable source materials.

Archival photos – archival documents

As part of our collaboration with the UNESCO Archives, we worked with their collection of photographs documenting the organization's history. The Archives manage a collection of about 170,000 visual resources, 5,000 of which have been deemed to be particularly important and included in the Digitizing Our Shared UNESCO History project.[86] We were tasked with uploading a smaller, curated selection of 100 photographs to Wikimedia Commons.

What makes this project interesting in the context of processing documents is that the photographs' metadata is not yet digitised and only exists on the backs of the photographs. The descriptions provide relevant information, the omitting of which would significantly lower their educational value. That's why we were provided with two files of each photo, one of the front and one of the back.

For the information written on the back side of a photo to be truly valuable, it has to be converted to text. That way, it is easier to read, find, analyze and share. OCR technology was developed for this task, but it is not perfect: the results of the automatic recognition process have to be validated by a human, as they can contain errors. This is particularly true when dealing with text that has developed bad contrast and blurriness due to age, as is the case with much archival material – including the UNESCO photos. Manually validating thousands of files does not require specialized knowledge, but can get tedious, and might not be considered an efficient use of staff resources by GLAM institutions.

That's why the UNESCO Archives have implemented a crowdsourcing project to transcribe the photograph captions which are then validated.[87] A volunteer does not have to invest a lot of time into transcribing a photograph that contains a paragraph or two of text. In exchange,

[86] https://digital.archives.unesco.org/en/collection/photos/
[87] https://heritagehelpers.co.uk/projects/view/details/project/unesco_tagging_photos

they get a sense of achievement from helping out one of the world's most renowned organizations. As of February 2020, 214 volunteers have entered the data from 48% of the 5,000 participating photographs and 37% has been validated.

It was the output of the crowdsourced transcription process that was shared with us together with the photograph files. We did not have to process the texts ourselves in order to upload it as metadata accompanying the images on Wikimedia Commons. Even though we had access to the transcribed captions, we still found the actual back sides of the photographs important enough to be uploaded as well. Firstly they provide an educational value, serving as an example of how archival photographs can be described. Secondly, information that is not conveyed by the raw text, such as the graphic layout or even the color of the ink might be of interest to researchers; we cannot imagine all possible uses people can have for them, so we should not limit them by withholding material that we have the power to share. Thirdly, the viewer can refer to the back sides to confirm that the transcribed text is indeed true to the source, which is important to researchers, journalists and others who need to be sure the resources found on Wikimedia Commons are reliable. That's why both sides of every photo were uploaded and linked to each other by including a thumbnail of the other side in the file's information box.[88]

Problem

The following steps were identified as necessary for the completion of the Nordic Museums project:

- All bibliographic metadata of the about 2,000 articles is uploaded to Wikidata. This was identified as necessary because structured data is much more powerful than unstructured data, making it possible to link to authors, topics, etc., as well as enabling complex queries and thus increasing the discoverability of the data. In particular, it would enable us to find articles written by authors who have been dead long enough for their works to have entered the public domain.
- Public domain articles in the collection are identified. This was identified as necessary because the museum had digitized all the articles, regardless of their copyright status, and did not provide information about the copyright status in their system. We had to identify the articles that could be uploaded to Wikimedia Commons.
- The public domain articles are uploaded to Wikimedia Commons. This step would make the articles available to Wikimedians.
- The public domain articles on Wikimedia Commons are connected to the corresponding Wikidata items. This step was identified as necessary to enable those using Wikidata to find interesting articles to access the scanned files and read them.

88 https://commons.wikimedia.org/wiki/File:Culture,_Triangle_campaign_-_UNESCO_-_PHOTO0000002100_0001.tiff

- A single article is published and proofread on Swedish Wikisource, to act as an example and to hopefully initiate community engagement with the collection. This step was identified necessary to serve as a proof of concept and basis for reflection on how it might be possible to use Wikisource to improve the educational benefits of digitized literature from GLAM collections.

The following problems were encountered in the course of pursuing these goals:

The author information in the metadata was only available as strings

This caused difficulties on two levels. First of all, it reduced the value of the data on Wikidata, as it made it impossible to link to the authors' items and utilize the power of Wikidata as a platform for linked structured data. It did not, however, prevent us from uploading the data – the Wikidata property author name string (P2093)[89] for precisely this reason, making it possible to add author information without having identified the corresponding Wikidata item. Indeed, since its creation back in 2015, the property has been used over 30 million times, primarily by editors importing large datasets of bibliographic metadata, indicating that we were far from alone with this problem. Volunteers have developed tools to work with author name strings on Wikidata, such as the Author Disambiguator.[90]

Using author name strings for all the authors, however, while fast and convenient, would have caused significant problems in our work on Wikimedia Commons, if not prevented it in its entirety. In order to upload the articles to Wikimedia Commons, we had to be sure they were in the public domain; this status depends on the author's death date, as described above. Neither the copyright status of the articles nor the authors' death dates were provided in the museum's database, meaning that we had to find this information on our own. Since many of the contributors to Fataburen were notable in Wikipedia/Wikidata terms, we expected to find they already had Wikidata items with the basic biographical information we needed.

To solve this problem we employed Mix'n'match, a volunteer-developed tool for matching strings (e.g. person names) to Wikidata items.[91] The tool makes it possible for editors to collaborate on matching a dataset. Once we published a Mix'n'match catalog with the authors' names, it made it quick and easy for both WMSE's and the museum's staff to match the names to Wikidata items, and even attracted the attention of two volunteer editors.[92] Once we were happy with the matching – that is, when all the names that could be linked to existing Wikidata items, a number of new items for authors that we found notable were created, and over 60% of the entries in the catalog where matched – the matching results were downloaded using the export function in Mix'n'Match. The authors who were matched were also the most prolific contributors to the journal, ensuring a high number of articles with linked authors. The remaining names would be added using the author name string property.

[89] https://www.wikidata.org/wiki/Property:P2093
[90] https://tools.wmflabs.org/author-disambiguator/
[91] https://meta.wikimedia.org/wiki/Mix%27n%27match
[92] https://tools.wmflabs.org/mix-n-match/#/catalog/2776

The mapping of authors' names to Wikidata was useful not only for the needs of this project. A couple of weeks later, it was ingested into KulturNav[93], a shared authority platform used by a number of cultural heritage institutions across Sweden and Norway. Since a KulturNav property exists on Wikidata (P1248)[94], the KulturNav URIs could then be added to the authors' items. The mapping we did can now be re-used by the Nordic Museum to enrich their collection, or indeed by any other GLAM institution using KulturNav.

Public domain files uploaded to Wikimedia Commons must be in the public domain in both the country of origin and the United States

Copyright law can get complicated on a global scale; Wikimedia Commons can only host free content, the definition of which is not the same in every country. It is an international project, with users and content from all over the world. However, the servers Wikimedia Commons (and the other Wikimedia platforms) is hosted on are located under U.S. jurisdiction. Because of that, a work must be covered by a valid free license or have entered the public domain in both the country where it was first published and the United States.[95]

In 2019, in order for a work to be in the public domain in the United States, it had to be published before 1924. In Sweden, works enter the public domain 70 years after the author's death. It is thus possible for a work to be in the public domain in Sweden, but not in the United States.

This problem was solved by only uploading articles that were in the public domain in both countries. It was a fully acceptable solution, as it left us with over 200 files to process, and the museum staff involved in the project were familiar with that particular Commons policy. Had we been working with a GLAM without previous experience of the Wikimedia projects, this issue might have taken some time to explain; it is also not inconceivable that it might have influenced their decision to contribute to Wikimedia Commons.

We had access to the scanned material in two formats. On the museum's open publishing platform DiVa, the articles are published as PDF files. Some of these PDF files contain the cover image of the issue in which the article was originally published, which was problematic: the copyright status of the cover image would have to be investigated separately in order to determine whether it could be uploaded. Had we decided to preventively exclude the cover images, we still would have had to process the PDF files to remove the offending page.

Furthermore, while PDF is an acceptable format on Wikimedia Commons and Wikisource, DjVu is often considered more suitable, due to it being an open standard from the start and the resulting availability of free and open tools to work with it.[96] We made the decision to use this open standard to make our contribution more accessible to the community.

The alternative to those PDF files were the individual page scans, in TIFF format, stored in the museum's internal database, not available to the public. They had a higher resolution

[93] https://kulturnav.org/5a10f039-2b24-4248-adff-f948e8eb8ca7
[94] https://www.wikidata.org/wiki/Property:P1248
[95] https://commons.wikimedia.org/wiki/Commons:Licensing#Interaction_of_US_and_non-US_copyright_law
[96] https://en.wikisource.org/wiki/Wikisource:DjVu_vs._PDF

than the PDF files, which was important to us, as it would make OCR-ing and proofreading the articles in Wikisource easier. This also gave us the freedom to collate them into DjVu files. This was what we decided to do – even though it involved more work than downloading the readily available PDF files, as we had to write custom scripts to download and convert the files. The museum staff agreed that sharing the materials in the highest possible quality was important.

TIFF is a lossless format and thus preferable to JPG in cases where high fidelity is important. Scanned text is such a case, as any aberrations will make it harder to process for OCR software and human eyes alike. That was our principal motivation for choosing to engage directly with the TIFF files. Another was that uploading the highest quality resources available is considered a good custom in the Wikimedia Commons community.[97] Not only are high-resolution files better suited for print and modern displays, they also prove the GLAM's dedication to sharing their material with the public in the same format that is available internally.

Implementation

The project was carried out in several steps, each step using a suitable tool.

Metadata processing

OpenRefine[98], an open-source application for data cleanup and reconciliation, was used to process the bibliographic metadata and upload it to Wikidata, creating a new item for each of the about 2,000 journal articles. OpenRefine is a popular tool among Wikidata editors, even though it is unaffiliated with Wikidata and was not originally built with Wikidata in mind. We used it to execute all the steps of the Wikidata process, from examining the raw metadata to creating the Wikidata items.

File processing and upload

Downloading the TIFF files from the museum’s internal database and collating them into DjVu files required us to write a dedicated script.[99] While highly specific for this particular task, we imagine the script could be partially reusable, as the database uses the same API as KulturNav.[100] Operations such as downloading documents from a list would be done in the same way in another database using this architecture. Of course, large parts of the scripts will be reusable if we ever do another project using materials from the museum’s database. The parts where the image files are converted and collated into DjVu files are also generic.

Pattypan[101] and OpenRefine were both used to upload the public domain articles to Wikimedia Commons. Pattypan is an open-source application for batch file uploading to Wikimedia Commons, designed to be easy to use for GLAM volunteers and staff. Pattypan

[97] https://commons.wikimedia.org/wiki/Commons:Why_we_need_high_resolution_media
[98] http://openrefine.org/
[99] https://github.com/Vesihiisi/NM-harvest
[100] https://kulturnav.org/info/api
[101] https://commons.wikimedia.org/wiki/Commons:Pattypan

uses a spreadsheet to handle the metadata of the uploaded files. That spreadsheet was prepared with OpenRefine, using the output of the Wikidata pre-processing from the previous step as an input and modifying it to fit into the Wikimedia Commons informational template format. It should be noted that OpenRefine is in no way necessary to prepare data for Pattypan ingestion; different users have different preferred tools and workflows, and the spreadsheet format offers great flexibility, down to editing the data by hand in a spreadsheet software of one's choice. In our case, using the two tools in tandem made the most sense, due to having already processed the metadata in OpenRefine in the Wikidata step.

Also used in this project were Mix'n'Match and Quickstatements, the former for the matching of author strings to Wikidata items outlined above and the latter for making batch edits to the Wikidata items of the journal articles, such as adding links to the scanned files on Wikimedia Commons. Both tools are developed by volunteers and popular among Wikidata editors.

Structured Data on Commons (SDC) was not used in this project. At the time of executing this project, the development of SDC was focused on media whose aim is to visualize something, such as artworks, photos and illustrations, unlike scanned text.

Wikisource

Once we had a number of interesting articles uploaded to Wikimedia Commons, we explored how to add them to Wikisource. An early question was whether to do this step automatically as well. It would be technically viable to make a bot that creates the index pages for all the 200 articles. But would it be beneficial to the community?

The Swedish Wikisource community is comparatively small, with a dozen active users. This means the rate of proofreading is slow but steady, and the selection of works that get proofread is determined by that small user group's interests. We decided that dumping 200 new documents on them would not be a nice thing to do (and likely be counter-productive). Instead, we posted an announcement on Wikisource's discussion forum that the documents are available on Wikimedia Commons, and created an index page for one of them.[102] After we initiated the proofreading process, other users came and finished it.

A couple weeks later we found that a dozen more articles from our upload had been proofread,[103] showing that the community was interested in the subject matter.

Outcome

The project resulted in: 204 files on Wikimedia Commons[104], 1,822 Wikidata items[105] and one proofread article on Wikisource[106]. At the time of writing (February 2020), the Wikisource community has proofread an additional 12 articles. We find this impressive since the

---

[102] https://sv.wikisource.org/wiki/Index:Gustaf_Retzius_som_etnograf_%E2%80%93_Fataburen_Kulturhistorisk_tidskrift_(1919).djvu

[103] https://sv.wikisource.org/wiki/Fataburen

[104] https://commons.wikimedia.org/wiki/Category:Files_from_Nordiska_museet:_2019-10

[105] https://w.wiki/HZL

[106] https://sv.wikisource.org/wiki/Fataburen/1919/Gustaf_Retzius_som_etnograf

Swedish Wikisource community is rather small: by comparison in February 2020, there were 18 users who had performed an action within the previous 30 days.[107]

Future

What this project demonstrates is that the way from raw material to Wikimedia uploads can be long and require several tools. This has been especially true since Wikidata became part of the Wikimedia ecosystem and garnered the interest of GLAM institutions. These days, creating Wikidata items for the materials uploaded to Wikimedia Commons and making them as complete as possible can be an important part of a GLAM upload process. The project Sum of All Paintings, which aims to create Wikidata items for every notable painting in the world, demonstrates how important GLAM collections are to the Wikimedia community.[108]

The most obvious advantage of working with Wikidata and Wikimedia Commons simultaneously is that it enables complex queries; like in our example, where we had to identify articles by sufficiently dead authors. On the other hand, with the emergence of Structured Data on Commons, one cannot help but wonder what the division between SDC and Wikidata should be. If we included SDC in our project, we would have had to provide data in three places on the file page on Wikimedia Commons, as SDC statements, and in the corresponding Wikidata item. This might be confusing primarily for new users, but also for experienced Wikimedians who first learned to work with files on Wikimedia Commons and then saw the emergence of Wikidata and SDC. To make things even more confusing, information templates on Wikimedia Commons based on the Artwork module[109] can automatically pull some information (e.g. the accession number) from the linked Wikidata item, if one has been provided. This is not clearly documented and the Artwork module is updated regularly with new functionalities.

This shows that working with structured data about GLAM content is not straightforward. Every uploader can develop their own workflow and strategy, and since both Structured Data on Commons and the way the informational templates pull data from Wikidata are under development, editors who follow the development can experiment and implement new ideas regularly. On the other hand, newcomers or editors who have no interest in technical development and just want to know exactly what is possible to do and how to do it might find this landscape difficult to navigate.

To solve this conundrum, or at least make it more approachable, one imagines documentation might be helpful. There is currently no single user's manual that describes the whole process, focusing specifically on the interplay of Wikidata, (non-structured) Wikimedia Commons and SDC and aimed at newcomers. The primary reason being, one assumes, the fact that the projects have evolved separately from each other and the communities are mostly separate as well. Secondly, technical development on these platforms is rapid and independent from each other. Thirdly, the tools available to editors, like QuickStatements, PattyPan, OpenRefine and the many small gadgets and users scripts

[107] https://sv.wikisource.org/wiki/Special:Statistics
[108] https://www.wikidata.org/wiki/Wikidata:WikiProject_sum_of_all_paintings
[109] https://commons.wikimedia.org/wiki/Module:Artwork

on both Wikidata and WikimediaCommons are developed independently by volunteers, so it's not realistic to expect all of them to engage in a coordinated documentation project.

Wikisource was an interesting part of this project, as it was the only one where we chose to engage manually. That was a conscious choice due to the small size of the Swedish Wikisource community. With such a small group of active users, it is obvious that the content of this platform is very curated and determined by the editors' interests. One might wonder what the reason for the activity being so low is. Is it simply a natural reflection of the low level of interest in crowdsourced transcription and proofreading in the general public, or do technical aspects also come into play?

The latter is definitely a possibility, as we have found. Once one has uploaded or found an interesting literary work on Wikimedia Commons, initializing a proofreading project on Wikisource is not easy – in fact, the documentation classifies it as an "advanced task".[110] In general, Wikisource might be the platform that requires the most technical knowledge to participate in – there are a lot of tags with arbitrary names that have to be memorized. Notably, the tags and formatting customs are determined by each Wikisource community, so the skills gained on Swedish Wikisource do not automatically transfer to English, German etc. Wikisource, which can also lead to editors being less active than they would like to.

We imagine that at least the process of initializing a new proofreading project could be significantly improved by a closer synergy with Wikimedia Commons and by the creation of a user-friendly tool akin to the Upload Wizard. When stumbling upon an interesting literary work on Wikimedia Commons, the user could be offered a direct link to import it into Wikisource. If the file on Wikimedia Commons contains information about the language of the work (in the information template, in SDC or in the linked Wikidata item), the right Wikisource version could be determined automatically. Then, the user would be asked a series of questions, one per screen so as to not risk informational overload, to determine the variables that they currently have to enter manually, like which pages of the work to include. Again, any information already available in structured form, like the name of the author, or the genre of the work, could be imported to Wikisource automatically.

The reason why we included the UNESCO Archives upload in this case study is that it also involves the problem of transcription, albeit approaching it from a different angle: the photo captions were proofread on a different platform, in a drive organized by the GLAM. That made our work considerably easier, as it is unclear how such a project could be undertaken in the Wikimedia environment. Wikisource was developed to proofread multi-paged works like books and articles, and requires, as mentioned previously, a fair amount of internalized technical knowledge. The transcription manual for volunteers published by UNESCO[111] makes clear that the platform they are using for their crowdsourcing project, HeritageHelpers, provides a much simpler environment, making it easier for newcomers to start contributing. Most importantly, the crowdsourcing platform makes it possible for users to enter the data directly into a set of predefined fields, such as *Country*, *Date range*, *Persons*,

[110] https://sv.wikisource.org/wiki/Wikisource:Korrekturl%C3%A4sning
[111] https://heritagehelpers.co.uk/img/uploads/projecten/unesco_tagging_photos/invoerinstructie.pdf

Credits and so on. This means that the output of the process has a structured form, which is much easier to process for Wikimedia Commons or Wikidata than unstructured text.

In this case study, we explored several ways to interact with visual resources with text content in order to maximize their benefits for Wikimedians and other users. A picture says more than a thousand words, but if the picture is a scan of an article or the back of a photo with text on it, the words cannot be neglected. Even though Wikimedia Commons is primarily a visual platform, it can store many different types of media files. If anything, this project demonstrates how many different ways exist to work with them. It makes it clear that there is a need for tools and workflows that do not address only one platform, like Wikimedia Commons. The Wikimedia power users, including those working with GLAM content partnerships, devote a lot of their time to figuring out the right tools for the task at hand, and the best ways to involve the different Wikimedia platforms to truly make the materials useful and used.

A centralized effort to develop tools and documentation for those users could have enormous benefits. It would not only save their time but also and lower the threshold for new editors, including GLAM professionals, who want to work directly with their material on the Wikimedia platforms. A typical GLAM professional who wants to share their expertise and has a basic understanding of the Wikimedia platforms would have to devote a lot of time to researching and learning all the tools and steps we implemented in this case study – especially if they did not know from the beginning what is possible to achieve in the first place. Streamlining the process and enabling seamless connections between the different Wikimedia platforms will give those users more power.

Case Study 5: EMPOWER – Engaging Museums around Problematic data On Wikimedia's Educational Resources

Key facts

Time: 1 December 2018 – 30 September 2019

Organizations involved: Swedish National Heritage Board, Rijksmuseum, The National Library of Sweden, Swedish National Archives, National Museum of Science and Technology, Nobel Prize Museum, National Museums of World Culture

Wikimedia/free knowledge communities involved: Wikimedia Sverige, Wikimedia Finland, Wikimedia Foundation

Keywords: Problematic data, outdated terminology, bias, Wikimedia Commons, Wikidata

Key conclusions

- Handling problematic data is an ongoing process and historic data and information should not be deleted, but instead updated and supplemented.
- In cases where we add new data and information we should use tools that support transparency in order to increase knowledge.
- We should ensure traceability in the process by presenting all the arguments on why new data and information needs to be added.
- There is a shared ownership of open platforms that makes collaborative learning and collaboration possible.
- Open platforms create opportunities for targeted interventions by affected groups, e.g. by adding information in their own languages.

Participants in the problematic data workshop. Axel Pettersson, CC-BY-4.0, via Wikimedia Commons.

Background

Collections at GLAM institutions have been built up over a long period of time and the documentation describing the content is often a reflection of different time periods and their values. The collections' information and data can be regarded as time capsules from the past with descriptions that can nowadays be regarded as unexplained, unclear and problematic. This applies not only to Sweden, but is generally applicable to all countries where collections are being built up over long periods of time.

The discussion of sensitive material in the collections is a recurring topic in the GLAM sector. There is a lot of documentation of issues through previous seminars and work. The focus has however largely been on interpretation and management within each GLAM institution's own collection. Open platforms and free licenses mean new problems to be solved. One advantage of using structured data is that you can make better and deeper searches. Open data can also mean that others can do searches from outside of your web pages if you transfer information from a text into accessible structured data.

The book *Words Matter*[112] by The National Museum of World Cultures in the Netherlands is a research publication on potentially sensitive words in the museum sector. *Words Matter* contains a review of some 50 problematic words and concepts, with suggestions of words to use in their place. These issues had a major impact in the media when the Terminology

[112] The book Words matter

group at the Rijksmuseum[113] worked with a selection of titles on works of art in the collection. [114]

There is an increasing debate within the museum community about the morals and ethics, if not legality, of museums creating and copyrighting media based on unethically acquired objects.[115] On the Wikimedia platforms, material may not carry any restrictions beyond those of CC BY-SA on Wikimedia Commons[116] or CC0 on Wikidata[117].

Problems

Problematic material from the institutions can provide knowledge on open platforms, but the handling requires caution and clarity in the process. What makes the open platforms different from when a GLAM publishes the material on their own website is the often drastic increase in visibility, coupled with the expressed ambition of seeing re-use of the material, expressed by highlighting the free licenses. We have defined problematic data as *"the data and information in the collections that can be perceived as abusive, provocative, racist or outdated"*. We have worked on three major general areas during this project:

Questions before uploading data to open platforms

- What are the benefits to the GLAM institutions of making problematic materials available?
- How can we reduce the concerns that exist when handling problematic material?
- How can we get GLAM institutions to make more problematic material available?

Questions after the upload

- How can we create value for, and give ownership to the group that has been exposed?
- What are the needs of processes, guidelines and tools when publishing on open platforms?
- How can the community of the platform aid in updating and contextualising the problematic data?

Follow-up activities

- How to follow up and maintain published material?
- How can you ensure that the material is reused, linked and disseminated in a way aligned with the intentions for making it available?

---

[113] The Research and Terminology group on Rijksmusuem
[114] Media 2015 Why the Rijksmuseum Is Removing Bigoted Terms from Its Artworks' Titles
[115] See for example Copyright Cortex.
[116] https://commons.wikimedia.org/wiki/Commons:Licensing
[117] https://www.wikidata.org/wiki/Wikidata:Licensing

Implementation

We have used an explorative investigative method and contacted organizations that have worked with sensitive data and information. There were 24 participants[118] in the workshop from institutions representing libraries, archives and museums. The selection was made through invitations to Swedish institutions that have had some previous experience of working on open platforms, or have been in contact with Wikimedia Sweden in other projects, and to a selected few international organizations. The workshop was held at Goto 10 in Stockholm, the coworking space where Wikimedia Sverige's office is located.

Communication

The communication has been based on contacts and inquiries with previous partners before the workshop which was conducted on 30 August 2019. In this way, the research work was combined with a discussion of how methods and tools can be developed on open platforms.

We wrote two blog posts, Workshop on problematic data?[119] and – Börja bara! Börja med det värsta ni har! (*Just start! Start with the worst you have!*)[120], on Wikimedia Sverige's website. The first post was an invitation to the workshop and was shared in social media by both Wikimedia Sverige, our partner the Swedish National Heritage Board, and by Digisam, the national coordinator of digitisation. The blog post about the Problematic Data workshop was also included in the September newsletter from Wikimedia Sverige.

On the 16th August the project was presented at the Wikimania conference in Stockholm in the GLAM space section *Structured Data on Wikimedia Commons for GLAM-Wiki*.[121]

Presentations

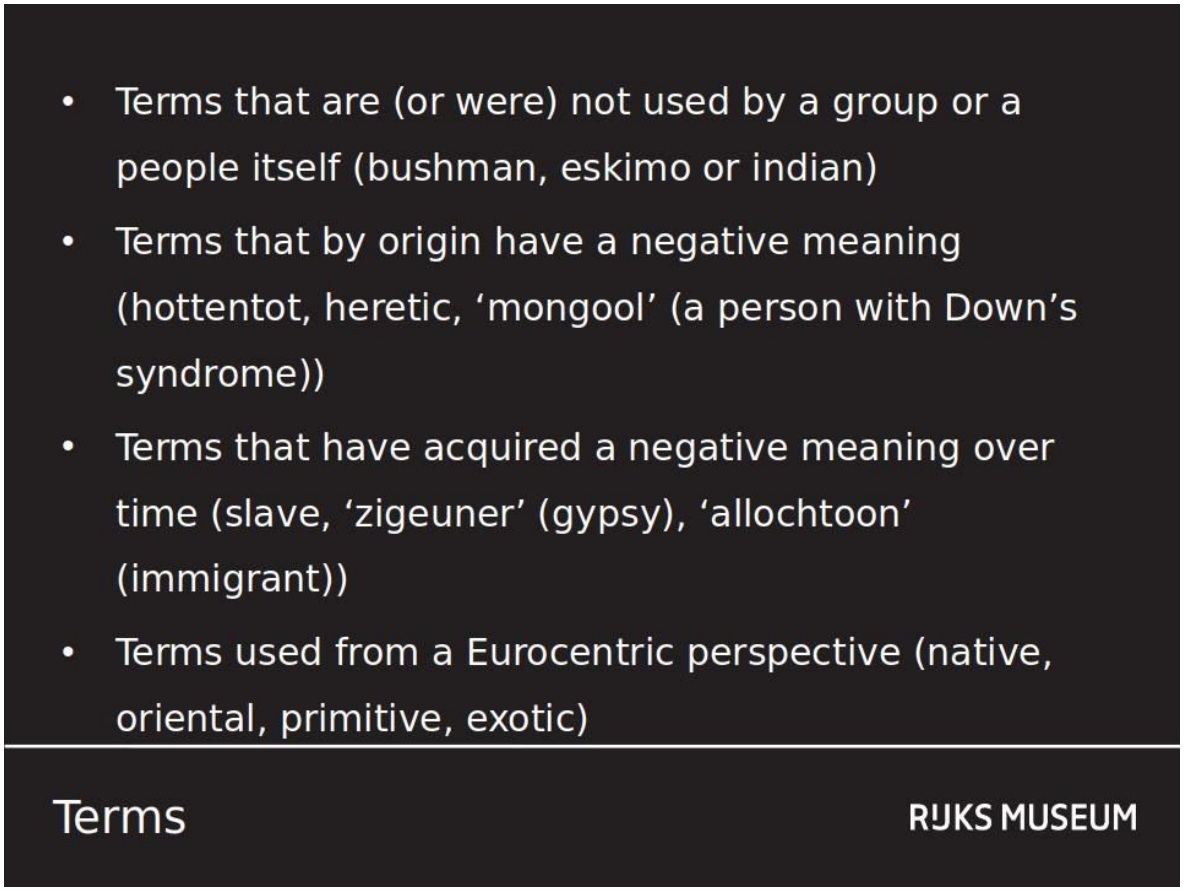
From a larger selection, we invited three lecturers with experience in dealing with problematic data, both in previous work and in current projects. The lecturers oepened the workshop on August 30 with presentations on how they and their organizations work with problematic data. The purpose of the presentations was to give all participants an understanding of how problematic data can look and be handled at different institutions for the workshop part of the day. The recorded presentations and slides from the workshop are available on Wikimedia Commons.[122]

Rijksmuseum's terminology project

Bas Nederveen, an information specialist at the Rijksmuseum, has worked from the start in a special group that has been tasked with critically assessing previously used terminology, and describes this in the lecture "Today's language for today's audience".[123] Problematic terms can be of the following nature:

---

[118] Global metrics Problematisk data, Stockholm
[119] Blog post at Wikimedia Sverige A workshop on Problematic data
[120] Blog post at Wikimedia Sverige (in Swedish) – Börja bara! Börja med det värsta ni har!
[121] Wikimania Stockholm 2019 Structured Data on Wikimedia Commons for GLAM-Wiki
[122] Wikimedia Commons Problematic data workshop 2019
[123] From the lecture "Today's language for today's audience" from Annexet Goto 10 in Stockholm



Terms. An image from Bas Nederveen's lecture on terms and terminology. CC-BY-SA-4.0, via Wikimedia Commons.

The process starts with choosing the term you want to work with. Terms mainly come from our collection database but may come from another work at the museum, such as a work in progress with a book or an exhibition. In working with terms, you research and consult experts in the subject, both internally with the curatorial staff at the museum and externally with the groups. Based on this research you choose an alternative. New titles and descriptions are added to the collection database. As of that moment, these become the preferred titles and descriptions when searching the database and the website. All research is documented in an information sheet with a description of the problem, the solutions and which sources were consulted.

Difficult person museum

"Who has the right to decide what to be remembered or forgotten from the past"

From the presentation by Stefan Bohman. CC-BY-SA-4.0, via Wikimedia Commons.

Stefan Bohman is a former chairman of Swedish ICOM and has written the book *Skelett i garderoben*[124] (Skeletons in the closet) about persons that are popular for some reason, but also have a problematic history and how this is presented in the person's museum. In the lecture at the workshop Stefan describes some conclusions on how difficult person museums[125] work on problematic issues.

There are two main issues to be considered:

1. Who has the right to decide what to be remembered or forgotten from the past?
2. What stories are told and what stories are not told?

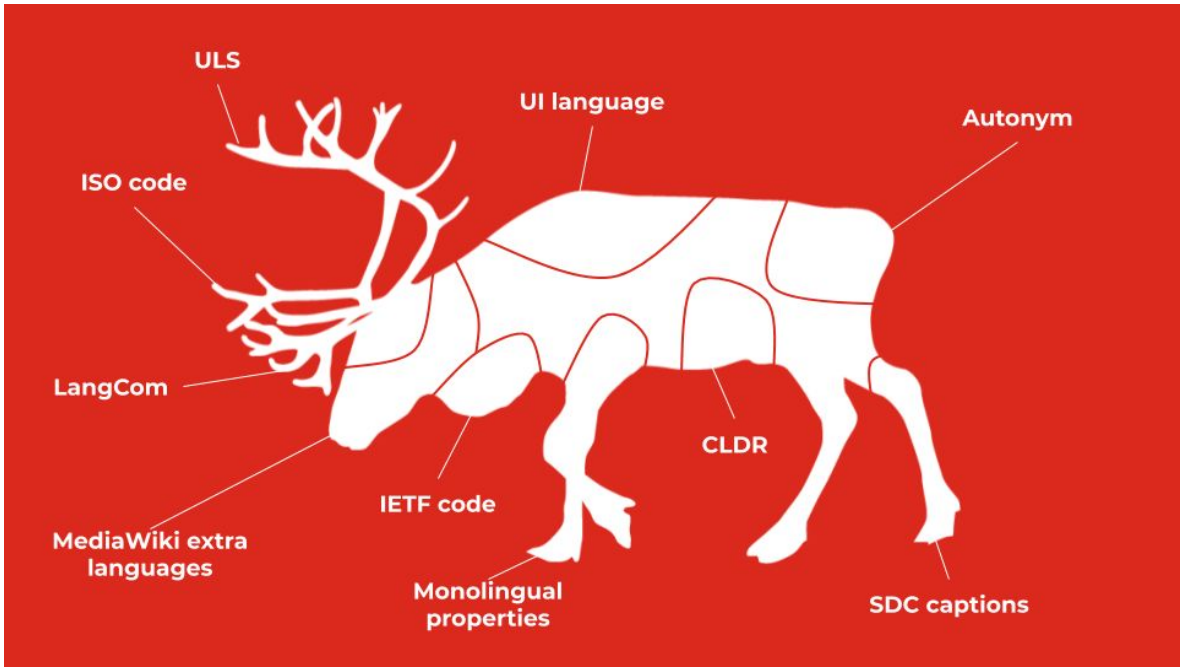
These issues will also become relevant when choosing what to upload to open platforms, and what to withhold. Stefan Bohman presents[126] several strategies for how GLAM institutions deal with difficult questions:

- **Full account:** The museum tells the visitors about the difficult questions in exhibitions and other material.
- **Omitting:** The problematic fact is not included in the museum's exhibitions or in any other museum material.
- **Double bookkeeping:** The museum presents the difficulties in different ways – one for the ordinary public and one for those with a special interest.
- **Minimizing:** The problematic facts are presented, but in a minimized way.
- **Reduction of responsibility:** The museum claims that everyone did the same, the society during the time "was just like that".
- **Comparison:** The person did really do bad things – but in comparison to their contribution to society it's of lesser importance.
- **Change of subject:** The museum concentrates on other subjects than the person's history and work.

WikiProject Saami

Susanna Ånäs is project coordinator at Wikimedia Finland and a project leader for the WikiProject Saami.[127] The project examines and improves knowledge about the Saami representation on Wikimedia platforms.

[124] Skelett i garderoben: svåra museer, http://libris.kb.se/bib/9kktljdz7lrnxz9p
[125] Difficult issues – Difficult Person Museums
[126] The lecture "Difficult person museums" at Annexet Goto 10 in Stockholm
[127] The WikiProject Saami at Meta-Wiki



One solution does not work for every Wikimedia project. An image from Susanna Ånäs' lecture WikiProject Saami. CC-BY-SA-4.0, via Wikimedia Commons.

The vision is to "make the Wikimedia projects more useful to the Saami communities, help the communities control the circulation of the representation of their culture, and to make the Saami communities, languages, and cultures more visible and factual across all Wikimedia projects." Susanna highlights several areas where work can be done to support this vision.[128]

- **Aspects of protection:** Attention to copyright, privacy and personality rights. Culturally sensitive and sacred knowledge. Protection against commercialization, theft and vandalism.
- **Visibility of sensitive data:** Is there a need to exclude information from Wikimedia projects? Ways to describe and filter from display? Remove location data?
- **Ways to correct information:** Remove or tag fake indigenous content. Add and use indigenous names and nationalities. Identify and tag personalities and locations when appropriate. Express consent and restrictions. Add knowledge and data provenance.
- **Decolonising the digital commons and terminology:** Translating and importing concepts into Wikidata enables tagging in minority languages. Initiative to import a multilingual Saami museum thesaurus and Saami place names. Propose the use of Traditional Knowledge labels for Saami communities.
- **Consent requires documentation and infrastructure:** Wikimedia environment has opportunities to store consent. OTRS[129] is used for licensing purposes, and can be repurposed for consent. In events it is possible to ask for consent. Children cannot

[128] From the lecture about WikiProject Saami at Annexet Goto 10 in Stockholm
[129] OTRS – Open-source Ticket Request System

legally consent. Both opt-in and opt-out should be possible, and the right to be forgotten should be respected.[130]

Workshop

In the introduction to the workshop, a brief presentation of Structured data on Commons[131] and a couple of different variations on templates that are already used on Wikimedia Commons was made. Templates can be seen as alerts and extra customized additional information. The purpose was to show the tools that exist and to keep in mind the platforms where development can take place.

The workshop was based on the open space method[132] which takes advantage of the participants' experience and ability to create relevant questions. What was generally desired for future work on problematic data was better discussions, more experiences, good examples, advice and support, more knowledge, how to give back to minorities, strategies and more confidence in what is problematic.

The starting point and the initial question was: *What problematic situations have you encountered or heard about when materials were made available to a larger public?* The participants in the workshop came from several different institutions, for example the Swedish National Heritage Board, Rijksmuseum, The National Library of Sweden, Swedish National Archives, National Museum of Science and Technology, Nobel Prize Museum, National Museums of World Culture.

This question and subsequent questions generated 18 proposals in areas for further discussion divided into two sessions during the afternoon. The person who suggested a problem area was also given the task of documenting the discussion in a report template. Out of the 18 proposals for in-depth discussion, the result was 10 reports that became working material for further analysis. Some examples of focus areas from the group discussions:

- **Trigger warnings:** Problematic expressions and outdated words.
- **Medical images:** Old language and abusive material.
- **Reproduction of old values:** When there are no valid facts.
- **Lack of interest from colleagues:** How we talk about and describe minorities.
- **Relatives:** Relatives or groups that do not like the museum's story.
- **Unconscientious pictures:** Children, bodies, minorities, tortured and dead people.
- **Illegal activities:** Legal considerations.

[130] https://en.wikipedia.org/wiki/Right_to_be_forgotten
[131] The project Structured data on Commons
[132] The method Open Space Technology

Outcome

Thorough preparatory work should be done before GLAM institutions can share problematic data with free licenses on open platforms. The conclusions after the lectures and workshop is that you should not change historical descriptions and, if necessary, expand and add new updated information. In this way, you maintain provenance and traceability over time. The reports were analysed and grouped around five general issues, presented below.

Ethical perspectives

Ethical perspectives seek to resolve questions of human morality by defining concepts. There is data where the initial creator might not have asked for the consent of those depicted, or where they lacked the de-facto ability to refuse consent due to asymmetric power relations. There are photographs of dead children and adults, tortured people, and minorities who do not want to be depicted after death in an online collection. This may not always be a legal issue, but an ethical one. There are already ethical guidelines for many cultural institutions, but those need to be further clarified when it comes to uploading materials to open platforms.

Legal aspects

There may be legal aspects of sharing problematic data and files on open platforms. The laws for abusive and racist material may look different in different countries. When is consent needed and how can it be expressed effectively? These issues are often difficult and have to be solved on a case by case basis. There are risks for institutions to end up in a context that cannot be influenced by themselves. There is also a fear that images can be used and edited by others and where the institution can become indirectly responsible. The intention when making problematic material available is to have transparency and clarity.

Terminology

Terminology serves to facilitate communication between people who are familiar with a subject area. Certain words and phrases used in the past may be perceived as offensive, unintentional or not, to a more contemporary audience and there is no reason to perpetuate racism or sexism. There are several different glossaries but they are not jointly created. Translation is an important issue when the terminology is specialised and concepts do not overlap perfectly in different languages. Working together with terminology on open platforms let institutions mix and influence the interpretation and the shared meaning of words and how they are used.

Labels

Problematic and sensitive material is often about people, specifically from disenfranchised communities. What is the significance of the description of these communities when made by an outside party, and how can this change over time? How do we name groups and what

happens when we use outdated concepts or names? There are problems with power and interpretability that are manifested throughout the communication of society.

Label systems can be seen as a method for tagging different kinds of problematic material. The project Traditional Knowledge Labels[133] is an example of such a system. To add information from a label system can provide guidance in the reuse and access to culturally sensitive content. A conceptual system could be used by indigenous communities to add protocols for access and use of cultural heritage but this must be investigated further.

Labels can be seen as a basis for further discussion. Responsiveness and cooperation are two keywords in the work to agree on commonly agreed descriptions. Tools that support different languages are critical to success here.

Reuse

The point of knowledge on open platforms is reuse and it's part of the concept of *open*[134] that people can edit and improve the content. This can be extra sensitive when it comes to problematic data. High-resolution digital images of materials can be reused in commercial contexts, which can be perceived as tasteless, ignorant, and/or inaccurate. Image agencies can place their own restrictions on free material on open platforms, with uncertainty as a result.

The advantages of working with problematic data on open platforms is that it creates a transparency and visibility that shows that these problems are actively taken seriously. It may be better to be the one who devotes resources to this work at an early stage and shows possible solutions to difficult questions.

It is difficult to generalize problematic data as it deals with different types of problems and each case has to be dealt with separately. But open platforms nevertheless provide opportunities to make materials accessible and to be able to influence how the material is curated. Ownership and curation becomes a joint commitment and responsibility.

Two tools to reduce the negative impact of reuse are; increased public understanding of how freely licensed material can be reused, and that the context in which material is encountered does not necessarily reflect its origins. Increased awareness on the side of content providers about potential reuse, so as to be prepared in the case of unwanted reuse.

Future

The number of uploads will increase as more material is given free licenses. There will continue to be a need for advice and support as the background, material and process will be different in each upload. We see an opportunity to take advantage of all experiences by documenting good examples in the work of developing the management of sensitive material.

[133] Traditional knowledge labels
[134] The open definition

Documentation and processes to develop

Wikimedia has the capability to create a knowledge bank around problematic data. It can support collaboration between institutions working on transparent open methods and assist in disseminating results. We recommend further work, development and projects in the three areas described below. For a good result we suggest working with institutions that have unpublished collections that contain problematic data.

Ethical perspectives

Existing ethical guidelines should be updated to also include open platforms. This can help institutions when preparing a collection with problematic content or when looking at sharing content on open platforms for the first time. We can release a new code of ethics under a free license and focus on the part that affects publishing on open platforms. We can write and organize a code of ethics for open platforms with inspiration from four areas:

- Code of ethics for **galleries**.
- Code of ethics for **libraries**.[135]
- Code of ethics for **archives**.[136]
- Code of ethics for **museums**.[137]

Legal aspects

There are several different laws that are relevant when dealing with problematic data. There is a need for more legal knowledge in this area with an international perspective. It is also an area where changes are constantly taking place so continuous education and platforms where this knowledge is easy to update are desirable. This work should preferably be done by country but can be compared to each other in the form of a table.[138] Working in a project on an open platform (one that allows anyone to freely access, use, modify, and share the content for any purpose[139]) can be used for four areas of legal aspects.

- **Harassment**, **discrimination** and other abusive treatment.
- The law and rules of **copyright**.
- **Integrity** and general data protection regulation.
- Indigenous and minority rights.

Terminology

Defining problematic terminology and preferable alternatives to it is an important step in recognising and addressing problematic data. It is crucial to work on this together on open

[135] Professional Codes of Ethics for Librarians.
[136] The ICA code of ethics for archives.
[137] The ICOM code of ethics for museums.
[138] Example of comparative table Freedom of panorama
[139] https://opendefinition.org/

platforms so that more institutions can influence the interpretation and, by extension, the shared meaning of the terminology. When more actors join the chosen platform becomes an authority. It is important to work directly on open platforms with free licenses so that there are opportunities for several players to contribute. If the terminology is to be used and remain relevant, it must also be maintained and updated. We can work with the terminology directly on three global projects with support for hundreds of languages.

- Add words at **Wiktionary**[140]
- Add information at **Lexicographical data**[141]
- Add knowledge and sources on articles at **Wikipedia**

Wikimedia tools for development

There are specific tools and processes that can be developed to facilitate work in databases and on open platforms. This can be about using the tools in new processes or in new ways that are suitable for uploading and handling problematic data. Our proposal is to continue with processes and development around these three areas, each with an associated set of tools. Preferably with an unpublished collection that contains problematic data where the purpose is to release it on an open platform. The tools can be developed in different ways depending on the system and platform. On the Wikimedia platforms and projects, development takes place continuously with the aim of collaborating and benefiting from each other's contributions and experiences.

Structured data

The technology Structured Data on Commons[142] aims to allow structured and machine-readable metadata to be associated with the free media files on Wikimedia Commons, to make them easier to view, search, edit, organize and re-use. This is one way in which data could be marked as potentially problematic using different statements and properties. The strength of Structured Data on Commons is that it is an open system, meaning anyone can add information. This opens up a discussion about individual interpretations that fit the character of problematic data very well. The result is a combination of domain knowledge of the content and how the technology supports the dissemination of this curated knowledge.

Structured Data on Commons is fairly new and more development is needed to make this mechanism easily usable for the case of problematic data.

Templates

One way to highlight important information can be to work with templates using visual elements, such as warning texts or other labels. This can be done easily in many systems

[140] Wiktionary is a free-content multilingual dictionary.
[141] Lexicographical data – words, phrases and sentences, described in many languages.
[142] Structured Data on Wikimedia Commons is tool for adding structured data on media files

and can be a way to make visible selected problematic parts at an early stage. These templates can be designed in different ways depending on the situation.

Templates[143] on Wikimedia Commons can also be further developed in both form and content. There are opportunities for specialized solutions based on presuppositions. These specialized solutions can be investigated and, above all, tested in real situations. There are ethical guidelines and laws in most areas and one way to add knowledge about problematic data is to actively link to them. Referring to a guideline or law can be a way to initiate a discussion about whether or not certain problematic data and images should be on free platforms. It is also a way to have contact with organizations that are a part of creating the practices in the area. A template can have one or more organizations as senders and refer to one or more relevant authorities in each case. We can also explore more about whether a labeling system as, for example, Traditional Knowledge Labels[144] work on open platforms.

It should be noted that templates, and metadata in general, often get divorced from media representations when these are published or reused on other platforms.

Properties

Properties[145] in Wikidata are a way to describe an item[146]. The use of properties can in a longer perspective become an important aspect in handling problematic data. Wikidata properties that describe terms as problematic might be used in a similar way to what dictionaries do. A description in dictionaries for certain words and concepts can for example be outdated when they want to show that the word is no longer in use. Properties for describing objects can be developed by addressing the need for similar solutions on open platforms. Developing new properties on Wikidata follows a process where several users can participate and in this way the property becomes embedded within the community even before it is used.

Having many properties that describe an object can be a way to increase the granularity of the information. Having the involved actors agree on terms and concepts makes it easier to collaborate. Along with reaching an agreement, you can also link to lexicons, sources, and discussions that describe more about the background of why the property is used for the problematic data.

[143] Templates on Wikimedia Commons
[144] Traditional knowledge labels
[145] Proposal for a Property at Wikidata
[146] See Wikidata glossary

Case Study 6: ORDER – Organized, Reusable Data Enhanced in Relationships

Key facts

Time: January – February 2020

Organizations involved: National Museum of Science and Technology

Wikimedia/free knowledge communities involved: Wikimedia Sverige

Keywords: Structured Data on Commons, metadata

Key conclusions

- Structured Data on Commons is mature enough to be interesting to uploaders of GLAM collections.
- SDC makes editing Wikimedia Commons more accessible to GLAM staff without previous experience.
- SDC cannot be edited directly from the popular upload tools, requiring editors to either write their own scripts or use any of the small volunteer-developed tools with limited scope.
- Being able to work with SDC directly in the metadata processing and file upload stage is crucial if we want to rapidly increase the number of files using SDC.

Hauptstrasse in Ochsenfurt, Germany (1901). Sigurd Curman, Public Domain, via Wikimedia Commons.

Background

The National Museum of Science and Technology (*Tekniska museet*) is a Swedish museum that expressed interest in collaborating with Wikimedia Sverige. They requested help with uploading one of their photo collections, comprising about 500 files by the Swedish architecture historian Sigurd Curman, to Wikimedia Commons. We had had a dialog about the possibility of this project for at least a year, with one museum staff member agreeing to map the photographs' metadata (such as depicted places and architectural objects) to Wikimedia Commons categories. Early in 2020, the selection of the files and the mapping of the metadata was finalized by the museum and the collection was ready to upload.

Structured Data on Commons (SDC) is information about media files on Wikimedia Commons that can be parsed by humans and machines alike. This is what distinguishes it

from the default way of describing media files using wikitext, which has been in use since the advent of Wikimedia Commons. The description is fully free-form which, while offering almost unfettered flexibility to the editor, does not easily lend itself to structure. It is formatted using wikitext, a markup language used across the Wikimedia platforms – yet another technology that newcomers have to learn at least the basics of in order to contribute efficiently.[147] Over the years, editors have developed hundreds of templates of varying complexity, as well as written and unwritten standards for how different sorts of media files should be described. This causes problems for both editors, who can easily and unwittingly deviate from the usual way things are done, and users who want to query Wikimedia Commons using the contents of the wikitext descriptions (which, due to the multitude of existing templates, are not trivial to parse).

From the perspective of years upon years of free-form file descriptions, Structured Data on Commons offers a revolutionary way of describing and querying the millions of resources collected by Wikimedians. Each piece of information, such as the date of creation of a file, or the camera coordinates, belongs in a dedicated field. Furthermore, SDC is tightly coupled to Wikidata; rather than entering an artist's name or describing what objects are depicted in a photo, the user only has to link to their Wikidata items. This in turn facilitates multilinguality, as Wikidata items can have labels in many different languages. Wikimedia Commons users can see them in the language of their choice, benefitting from the work of Wikidata editors. Thus, SDC saves time for both file uploaders and users of Wikimedia Commons.

We had not previously worked with SDC in a systematic way, so this project provided us with a valuable opportunity to explore this area and its suitability for future GLAM uploads.

Problem

The problem this project tried to solve was implementing Structured Data on Commons in a typical GLAM image upload using the available tools and documentation. By doing that, our goal was to get a realistic overview of the existing SDC infrastructure, both in terms of SDC itself (maturity and usefulness of the structured data layer in Wikimedia Commons, which at the time of writing is under development) and the tools available to editors, their strength and weaknesses.

The Curman collection was particularly suitable for such a project due to its consistency. All the photos are authored by the same photographer and have the same copyright status, making the collection a good candidate to test adding structured data in batches. At the same time, its relatively small size – only 500 images – makes it possible to work with structured data using the visual tools available as Wikimedia Commons scripts or gadgets. It also makes it attractive to volunteer editors wishing to make manual SDC edits on an ad-hoc basis. A collection of several thousand items – not unusual in our work with GLAMs – would require more efficient tools and processes.

[147] https://en.wikipedia.org/wiki/Help:Wikitext

Implementation

The material provided by the National Museum of Science and Technology consisted of 558 TIFF files with accompanying metadata supplied in a spreadsheet. We used OpenRefine to process the metadata for both Wikidata and Wikimedia Commons, and Pattypan to upload the files to Wikimedia Commons. The process did not differ from that outlined in Case Study 4 and thus does not need to be elaborated upon in more detail here. Instead, we will focus on the SDC work which was done after uploading the files to Wikimedia Commons.

Neither OpenRefine nor Pattypan currently support Structured Data on Commons. This was an important factor in our choice of workflow and tools. There are, to our best knowledge, no tools or applications for uploading files to Wikimedia Commons that offer SDC integration. We thus faced two alternatives: writing our own tools – either for uploading files together with SDC statements or, more simply, to interact with SDC in existing files via the API – or surveying and using the available tools. The advantage of the latter solution was that it would give us a good overview of what is available to regular users, which we would need anyway before embarking on tool development on our own.

Add to Commons Descriptive Claims (ACDC) is a gadget to add SDC statements to a set of files.[148] It can be used on a manually defined list of files, on the contents of a category or on a PagePile[149] (list of files generated by one of several search tools such as Petscan[150]). The tool was developed and is maintained by a volunteer. It does not have a predefined list of statements; users are free to add any statements currently supported by SDC, including qualifiers. We used this tool to add the following SDC statements:

- creator → Sigurd Curman, with the qualifier object has role → photographer,
- copyright status → public domain, with the qualifier determination method → 50 years after creation of (non-artistic) photographic image
- copyright license → Public Domain Mark

Adding these statements was straightforward. However, the changes were applied at a very slow rate, taking several hours for the whole batch of 500 files being processed.

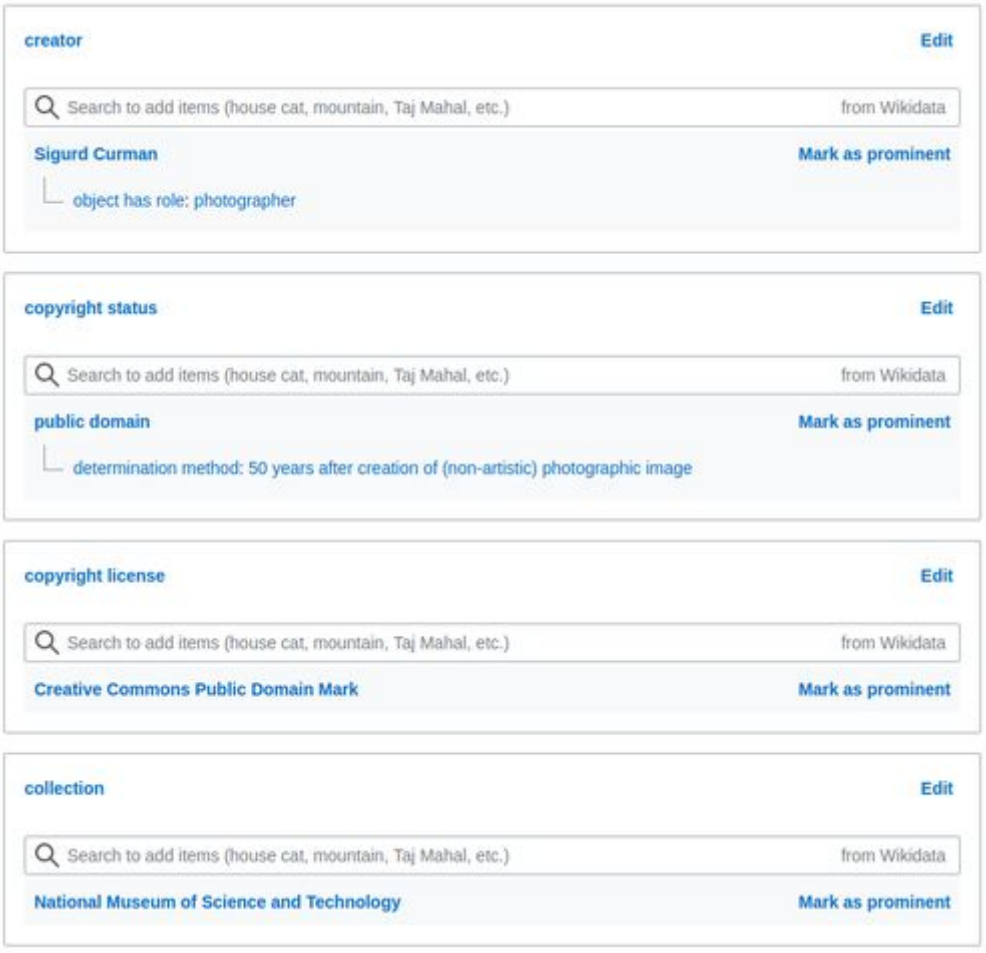
The SDC Tool[151] is a user script to add SDC statements to a set of files. It can be used on categories, galleries and search results. Just like the Descriptive Claims tools, it was developed by a volunteer. The main difference from Descriptive Claims is that it only allows users to add statements from a predefined list, which in February 2020 consisted of: depicts, depicted part, collection, significant event, location of creation, and location of the point of view. We used this tool to add the following SDC statements:

- collection → National Museum of Science and Technology

[148] https://commons.wikimedia.org/wiki/Help:Gadget-ACDC
[149] https://tools.wmflabs.org/pagepile
[150] https://petscan.wmflabs.org/
[151] https://commons.wikimedia.org/wiki/User:Magnus_Manske/sdc_tool.js

We did not find more functionalities in this tool than the Descriptive Claims tool had to offer. While the execution was noticeably faster, the limited number of statements available makes this tool much less flexible. Advanced users can copy the source code to their own user space and modify the tool in accordance with their needs, but this option might not be accessible to everyone.

Outcome



Structured Data in one of the photographs included in the upload. CC-BY-SA-4.0, via Wikimedia Commons.

We uploaded 558 photographs to Wikimedia Commons[152] as well as created the same number of Wikidata items[153]. Each file on Wikimedia Commons is linked to the corresponding Wikidata item, and the other way round. We added 4 structured data statements to each of the files, as detailed above.

[152] https://commons.wikimedia.org/wiki/Category:Media_contributed_by_Tekniska_museet:_2020-02
[153] https://w.wiki/JEf

The selection of statements to add was one of the big considerations of this project.The provided metadata contained information that would have been interesting to add but that we ended up omitting, such as:

- The dates the photos were taken. At the point of working with the Curman collection, adding time and date statements had not been implemented in the UI and was thus not possible for Wikimedia Commons editors.[154] Currently (February 2020) it is possible to add time and date statements directly via the API, even though users are prevented from editing them due to the data type not being supported in the UI.
- File-specific data, i.e. data not common for the whole collection. This includes the aforementioned dates, but also information about where the photo was taken and what it depicts. This information is present in the metadata, and was included in the form of categories, as mapped by the museum staff, so it was not lost. A possible workflow to add this information as SDC based on the categories is as follows:
    - Create a Pagepile in Petscan from the cross-section of the upload category (Media_contributed_by_Tekniska_museet:_2020-02) and the desired content category, e.g. Adlerstraße_(Nuremberg)
    - Using the Descriptive Claims tool, add depicts statements to all the files in the Pagepile.

This solution is quite time-consuming, as it requires semi-manual work to create the Pagepiles, which is why we did not follow this way.

The major outcome of this project was learning about the possibilities and limitations of the available tools from the perspective of a medium-scale GLAM upload. Regrettably, our selection of SDC statements was determined by the technical solutions available. We felt limited by the tools and unable to use the full potential of the data shared by the GLAM. While a lot can be done already, thanks to the contributions of volunteer tool developers, the lack of fully automatic solutions, integrated with the existing upload software, is palpable.

Future

Following the development of Structured Data on Commons has been very interesting from the point of view of GLAM uploads. SDC has an enormous potential to make Wikimedia Commons more understandable and manageable for GLAM staff who might feel overwhelmed by the complex wikitext descriptions with their dozens of templates and the category system that has evolved organically over the years. GLAM staff are usually expert users of databases and catalogs, used to well-thought-out categorizing systems. They expect to be able to query the collections using different criteria, which is not easy today. If all files had structured data statements, making even extremely fine-tuned queries would be trivial (e.g. all paintings by female European artists depicting cats created in the 19th century).

However, in order to actually achieve this dream – providing detailed structured data for the backlog of millions of files that are already up on Wikimedia Commons, as well as making it easy to include in new uploads, especially large-scale ones – the development of efficient

[154] https://phabricator.wikimedia.org/T231979

and robust tools is crucial. This is not something that can be left to volunteer developers. Ideally, SDC support should be implemented in any of the tools that are currently used by Wikimedians, of which OpenRefine and Pattypan are obvious candidates, due to their popularity among Wikimedia Commons uploaders.

Indeed, the developers of OpenRefine are aware of how important implementing SDC support in their application would be for the Wikimedia community.[155] This is a very realistic goal, seeing that OpenRefine already supports Wikidata reconciliation and is thus a popular tool among Wikidata editors. Both Wikidata and SDC use Wikibase as their underlying platform; extending the Wikidata support in OpenRefine to any Wikibase instance[156] would open up possibilities including, but not limited to SDC.

[155] https://github.com/OpenRefine/OpenRefine/issues/2144
[156] https://github.com/OpenRefine/OpenRefine/issues/1640

Case Study 7: AROUND – Advance Return Of User-generated New Data

Key facts

Time of the Case study: November 2018 – June 2019

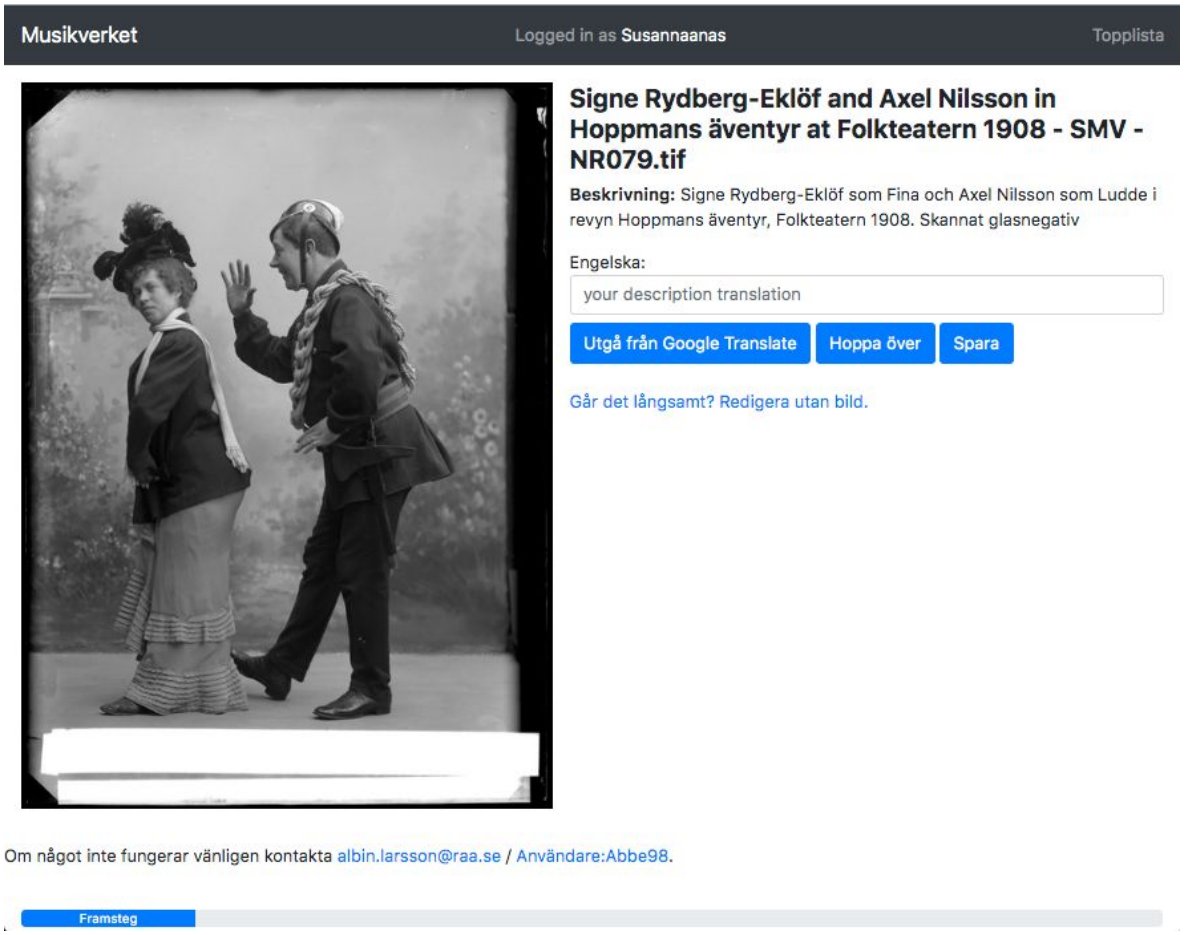
Organizations involved: Swedish National Heritage Board, Swedish Performance Arts Agency, Nationalmuseum, Nordic Museum

Wikimedia/free knowledge communities involved: Wikimedia Sverige

Keywords: GLAM, Wikimedia Commons, metadata, crowdsourcing

Key conclusions

- There is an interest among cultural heritage institutions in extracting enriched metadata from Wikimedia Commons.
- Extracting and ingesting this data is not easy for most organizations.
- Lack of trust and verifiability is a common obstacle to ingesting data from volunteer contributors.
- Insufficient technical resources are an obstacle to ingesting data from other institutions and from authority files.



The Roundtripping translation tool built for the Swedish Performing Arts Agency. Albin Larsson, Public Domain, via Wikimedia Commons.

Background

Wikimedia Commons contains a lot of material shared by cultural heritage institutions. When included in Wikipedia articles in many languages, this material can reach a larger and more diverse audience than on the GLAM website. Additionally, the Wikimedia audience has the ability to edit the material, improving the descriptions and translating them into other languages, categorizing the files, pointing out errors, and so on. Those additions, however, only benefit other Wikimedians, as they are not reflected back at the source – in the GLAM's own database. This is a net loss for the cultural heritage institutions that miss out on the fruits of the work of volunteers around the world. Instead, they could ingest the improved metadata back into their collection management systems, which would not only enrich their collections, but also send a strong signal to Wikimedia editors that their work is noticed and valued. This process is called *data roundtripping*.

A research project undertaken in mid-2019 by Wikimedia Deutschland, focusing on how cultural heritage institutions use Wikidata, also explored the question of ingesting data from Wikidata into the institutions' internal systems.[157] A significant proportion of the participants in

157 https://meta.wikimedia.org/wiki/Research:Wikidata_Use_in_Cultural_Institutions_(2019,_Qualitative_Research)

the study, representing different types of GLAMs mostly in Europe, were positive to the idea, mentioning that the data on Wikidata might be of better quality than the institutions' data. However, despite the broad interest, only a few institutions were found to have implemented data roundtripping solutions of some sort.

This case study is a short summary of a project exploring the interest in, and possible practical implementations of, data roundtripping that was carried out by the Swedish National Heritage Board with partial funding from the European Union within the Europeana Common Culture project. See https://meta.wikimedia.org/wiki/Wikimedia_Commons_Data_Roundtripping for the full documentation.

Problem

The project aimed at researching to which degree cultural heritage institutions can use the Wikimedia platforms to 1) engage contributors to actively interact with their content, and 2) create the data created by Wikimedians to enrich their own data repositories.

Implementation

The project was initialized with a research stage to get an insight into how GLAMs use third-party metadata in their collection management systems, consisting of a survey and a set of qualitative interviews.

Three pilot studies were then designed and carried out in collaboration with three Swedish GLAMs.

The **Swedish Performing Arts Agency** pilot focused on translating photo descriptions from Swedish into English. A dedicated tool was developed, in which users were presented with photographs from the Swedish Performing Arts Agency's collection on Wikimedia Commons, and could translate the descriptions either from scratch or by improving the output from Google Translate. The edits were saved directly in Wikimedia Commons. The results of this project were not ingested into the GLAM's database due to quality issues.

The **Nationalmuseum** pilot focused on retrieving authority ID's from Wikidata into the museum's collection management system. The pilot targeted the items of artists represented in the museum's collections, identified by the Nationalmuseum Sweden artist ID property (P2538), and the ID's to import included the Wikidata ID, KulturNav, VIAF and ULAN.

The **Nordic Museum** pilot made use of Structured Data on Commons and focused on adding *depicts* (P180) statements to fashion paintings from the museum's collection. A dedicated tagging tool was developed, limited to a set of controlled vocabulary, to make this task easy for the users.

Outcome

Ingesting third-party contributions into the institutional data repositories is not easy to implement for GLAMs, even if they are generally interested in it. Issues of trust and verifiability are notable obstacles cited when discussing crowdsourced contributions. On the technical side, there's a great variety of content management systems used by cultural heritage institutions, which affect how easily such ingestion can be done and whether external assistance is necessary.

Future

The Wikimedia platforms are an important part of the global knowledge landscape. Cultural heritage institutions around the world are becoming increasingly more interested in sharing their materials there, which is why further research in the area of third-party data ingestion is needed. Best practices need to be developed on how to manage third party information and communicate its provenance to end users. Efficient validation processes for crowdsourced information are needed in order to ensure the verifiability and quality of information in the institutions' content management systems. On the technical side, content management systems should become more flexible in allowing easy metadata ingestion, and open data exchange standards should be developed and/or implemented more widely.

Overall conclusions from the case studies

Herding cats, the wiki way

Malmö City Library. Johannes Jansson, CC-BY-2.5-DK, via Wikimedia Commons.

What these case studies demonstrate, first and foremost, is that GLAM collections are extremely varied. Even though we worked predominantly with material from Swedish institutions, we had an opportunity to engage with different material types and themes: sound recordings, sheet music, structured lexicographic data, photographs, scanned literature and databases. This is but a selection of the different types of material that a Wikimedia chapter with a strong contact network in its local community might encounter.

What all the projects presented in this paper have in common is that in order to execute each of them, we had to conduct research into the existing tools and similar uploads done by others. That was a necessary step to pick the right tools and plan our work in the quickly changing Wikimedia landscape. In general, we have noticed that the tools available to Wikimedians become more robust, stable and versatile every year. This means that there is less need for individuals to build and maintain one-off tools for processing and uploading specific collections, which makes large-scale contributions to the Wikimedia platforms feasible for people without programming knowledge – such as GLAM professionals.

Another common theme has been that there is no one way of doing things. This stems from the nature of the Wikimedia movement. The GLAM community of practice is large and diverse, uniting people with varied backgrounds, skills and interests. Codifying procedures and standards goes against the wiki way of doing things, where suitable solutions grow organically from discussions and experiences, where volunteer developers create small tools to solve very specific problems and where everyone has the right to make their voice heard and be bold.

But there are also negative consequences of the current situation: new or less technically experienced community members face high learning curves, volunteer developers are put under undue pressure if their tools become essential to other users, and outsiders, such as GLAM partners, require training to understand the landscape of Wikimedia platforms, gadgets, scripts, Toolserver tools and documentation.

Building a hub to strengthen and empower the GLAM communities of practice

The 2030 strategic direction states that Wikimedia will become the essential infrastructure of the ecosystem of free knowledge, and anyone who shares our vision will be able to join us.[158] In order for this to become reality, attracting and involving GLAM institutions is crucial. The world's museums, galleries, archives and libraries are protectors of knowledge, and sharing it is part of their mission. The Wikimedia platforms should be the obvious choice for cultural heritage institutions wishing to share their free resources.

As this paper shows, this is easier said than done. Over the years, Wikimedia Sverige has been serving as a middleman between GLAMs and the Wikimedia platforms. Initially, we would build one-off tools to process and upload specific collections in a process that was opaque to outsiders and required programming knowledge. Today, Wikimedians have flexible tools at their disposal that they can use for large-scale data processing and uploads to Wikidata and Wikimedia Commons. Volunteer developers build scripts and programs that make tasks easier. The GLAM community of practice is becoming stronger and more efficient every year.

We want to take this even further.

Since 2019, Wikimedia Sverige and Wikimedia Foundation have been working towards building a thematic hub for institutional content partnerships – a socio-technical support structure for GLAM communities of practice.[159] The initiative is currently in its early stages, and reflecting upon our experience with GLAM uploads and collaborations is a large part of the process. The GLAM hub will provide an arena for all Wikimedia communities and partners to work together and be part of a support network. And maybe most importantly, it will provide them with user-friendly tools and technical infrastructure.

As mentioned previously, this top-down approach differs from the way Wikimedians have traditionally worked. While the culture of freedom and innovation has resulted in a powerful

[158] https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2018-20
[159] https://meta.wikimedia.org/wiki/WMSE-WMF_joint_initiative_concerning_GLAM_communities_of_practice

and varied landscape of tools and workflows, we believe that focused, centralized work on GLAM tools and services is a worthwhile investment that will benefit Wikimedia contributors, affiliates and GLAM partners alike.

In this paper, we have demonstrated some areas where content uploaders face problems due to lack of technology or documentation. It is important to note that many of those problems are not specific to the particular data sets or collections we worked with. Quite the opposite, they affect the work of anyone working with large-scale uploads. By addressing those problems, we can make uploading material easier, faster and accessible to more people – which in turn will make the Wikimedia platforms more interesting and attractive to content partners.

An advantage of building a GLAM hub to address those issues is that we have a unique position and perspective. First and foremost, by serving as a catalyst for ideas and building links between Wikimedia contributors, affiliates and GLAM partners, we can collect experiences and views from all around the world. We realize that our own experiences are not sufficient to make generalizations about the movement. The GLAMs in our contact network have a lot in common – they all have experience with digitization and online audience engagement. This is not the case everywhere, but we alone cannot reach the GLAMs who need more help in this respect due to language and local knowledge barriers. We want to become the to-go place for everyone interested in large-scale content partnerships to share their ideas, reach out for support and learn, facilitating the flow of knowledge from more experienced institutions and Wikimedians. We believe everyone involved in this line of work has stories such as those outlined in this paper that they can share with others.

Furthermore, both Wikimedia Sverige and Wikimedia Foundation have a strong position in the movement, which is an excellent springboard for this venture. Wikimedia Sverige has a decade of experience of content partnerships, a team with technical expertise and a contact network in the GLAM sector both in Sweden and internationally. Wikimedia Foundation is not only responsible for the MediaWiki software that Wikimedia Commons is built on, but also has a knowledgeable GLAM team with an international contact network and strong capabilities for community engagement. Moreover, WMF is developing the Structured Data on Commons infrastructure, which deserves a separate mention due to its potential to revolutionize working with GLAM collections. Our collaboration will make it easier to continue the development of SDC based on real needs of GLAM institutions.

Archives – rather than "vast cemeteries" – are places where one may find the "experiences, adventures, risks, and dramas" of society, as Jaime Torres Bodet, a Mexican writer and director-general of UNESCO once said.[160] This applies to all the world’s cultural heritage institutions. For years, the Wikimedia community has been working hard to make those resources accessible to everyone, for free. We now have an opportunity to support them even better.

[160] https://unesdoc.unesco.org/archives/focus