SocietyByte
Wissenschaftsmagazin des BFH-Zentrums Digital Society
https://www.societybyte.swiss

An international Knowledge Base for all Heritage Institutions (Part 2*)

Autoren : Giovanna Fontenelle, Beat Estermann

Datum : 5. Juli 2019

Heritage institutions are places in which works of art, historical records, and other objects of cultural or scientific interest are sheltered and made accessible to the public. The equivalent of that in the digital world, is already taking shape, through digitization and sharing of digital-born or digitized objects on online platforms. In this second part, we describe the different modules of the project in more detail and sketch an avenue for the internationalization of the project. In part 1 of this article, we have described how Wikipedia and related Wikimedia projects play a special role in the emerging data and platform ecosystem, and we have shorty presented the "Sum of All GLAM" project[1], which proposes to improve the coverage of heritage institutions in Wikidata and Wikipedia.

Curation of existing data

Before ingesting new data, it usually makes sense to analyse the existing data on Wikidata and

to correct any instances of bad data modelling. One common problem are Wikidata entries concerning heritage institutions not properly differentiating between “building” and “organization”. Yet to avoid extra work later, it is crucial to make this distinction and correct any other data modelling issues before adding anything to these entries. To coordinate the resolution of data modelling issues, the data cleansing tasks carried out on the Brazilian dataset will be documented and serve as an example to guide similar data cleansing tasks in other countries. The plan is to have these tasks carried out in a coordinated manner by Wikidataists around the world.

In parallel to the cleansing of existing data, some fundamental questions need to be asked about the data:

- To what extent is the data complete? – Is there a Wikidata entry for every existing heritage institution in that country? To what extent is all the information needed for the Wikipedia infoboxes already present in Wikidata?
- How good is the data? – Is the data correct and up-to-date or is it outdated? Is outdated information properly historicized? Are the internal structures of heritage institutions properly represented? Is all the data properly sourced?

After this initial analysis, a strategy for further improvement of the data can be devised on a country-by-country basis. Apart from the manual enhancement of the data by existing members of the Wikidata community, two important avenues need to be pursued to ensure the provision of complete, high-quality data: the integration of existing databases as well as crowdsourcing campaigns targeting both heritage professionals and Wikipedians alike.

Data provision through cooperation with maintainers of GLAM databases

The easiest way to incorporate large quantities of high-quality data into Wikidata and properly reference them to a reliable source is to cooperate with maintainers of official GLAM databases. As the experience in the OpenGLAM Benchmark Surveyhas shown, it is quite easy in some countries to get access to well-curated and complete databases of heritage institutions, while in other countries, such databases are less complete, not that well curated, or may not even exist. In several countries, such as Brazil, Switzerland, or Ukraine, data about all known heritage institutions have already been incorporated. In several other countries, databases are available, but data has not yet been ingested. Itis the project’s goal not only to incorporate data once, but also to establish long-term partnerships with the maintainers of relevant databases to ensure regular updating of the data in Wikidata. At the same time, maintainers of the databases are likely to benefit from many pairs of eyes spotting errors in the data or enhancing existing databases by adding further information.

Data provision and maintenance by means of crowdsourcing campaigns

Where existing databases do not exist, crowdsourcing campaigns are envisaged that will address heritage professionals and Wikipedians alike. For this purpose, data maintenance and improvement tasks need to be documented and broken down into easily understandable, manageable chunks. This documentation will be developed over the coming months in cooperation with test users, and trials will be carried out both in Brazil and Switzerland. Larger campaigns will be scheduled for 2020.

Implementation of Wikidata-powered Infoboxes

To gain more visibility for the ingested data and to close the feedback loop between data provision and data use, Wikidata-powered infoboxes will be rolled-out across Wikipedia. This will require negotiation with various Wikipedia communities, which in the past have adopted differing policies with regard to the use of data from Wikidata inthe article name space. In some Wikipedias, such as the Catalan Wikipedia, Wikidata-powered infoboxes are in widespread use, while other communities, such as the ones on the German or the English Wikipedia, have been more reticent – partly due to quality considerations. Entering a dialogue with the more demanding communities is therefore important to drive efforts to enhance the data quality on Wikidata. While engaging in these dialogues, the project team will document use cases which will provide an empirical basis for the assessment of data completeness and guide further efforts. On the Wikipedia side, transcluding data directly from Wikidata will lead to important benefits, as information that currently must be updated in a myriad of different language versions separately, will be stored in a central place on Wikidata and maintained in a collaborative effort by the various language communities. For smaller communities, this is the only way to cope with an ever-growing amount of structured data in a Wikipedia environment facing a stagnating or shrinking contributor base. And for larger language communities, it is a good way to help provide up-to-date information about their own geographic areas in other languages. To enhance the chances of buy-in from many communities and to facilitate the roll-out of infoboxes across the various language versions of Wikipedia, it is important to make high-quality and properly sourced data available on Wikidata. Furthermore, according to the best practice when creating Wikidata-powered infoboxes, it will always be possible to overwrite information in infoboxes locally by the Wikipedia community if necessary. And last but not least, the roll-out will take place across several language communities in a flexible manner, following the pace of the different communities. Currently, Wikidata-powered infobox templates for museums have already been implemented on the Portuguese(see figure 5) and on the Italian Wikipedias; another one for archives has been prepared in the Portugueseversion. To spread the practice more quickly at an international level, it would be helpful if the templates could be rolled out on English Wikipedia at an early stage of the project.

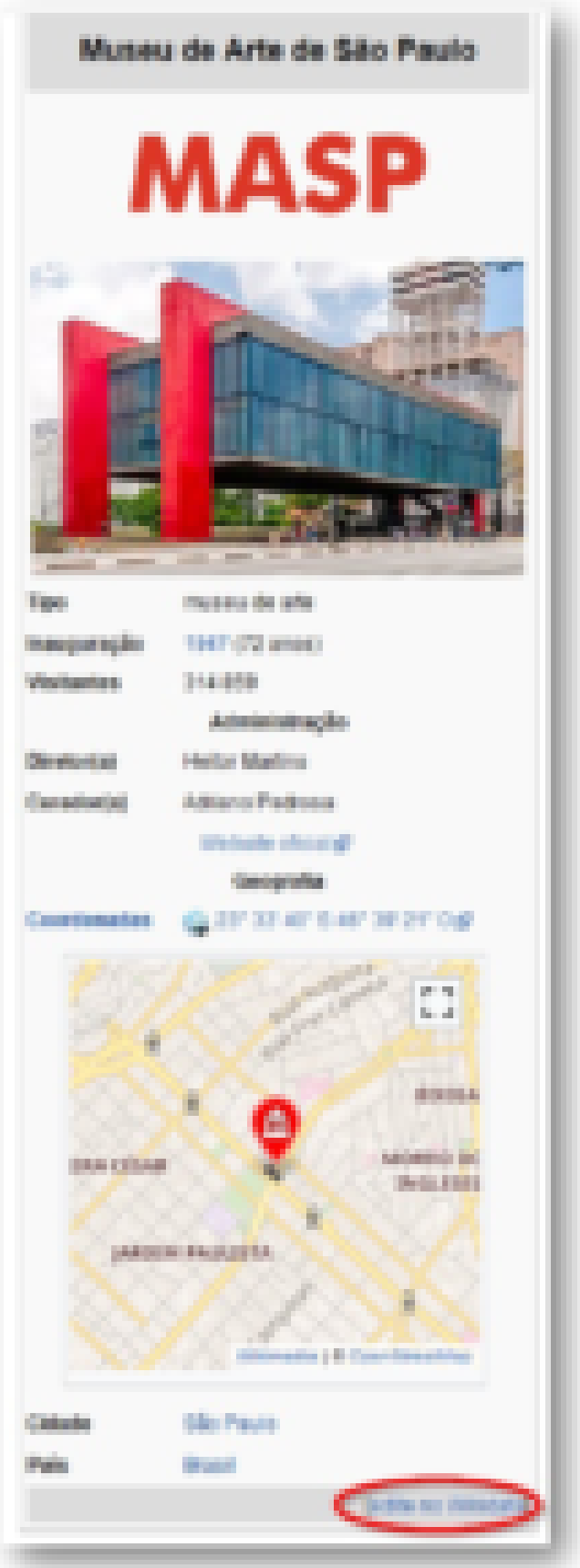

Figure 5: Wikidata-powered Wikipedia infobox for Museums on the Portuguese Wikipedia

Mbabel template to support edit-a-thons or editing campaigns

In addition to providing data for infoboxes, the entries on Wikidata can also be used to create article stubs to aid the creation of new articles about heritage institutions. This is where the **Mbabel tool** comes in; it lets Wikipedia editors automatically create draft articles in their user namespace by providing the structure of an article based on the data contained in Wikidata. This structure includes an introductory sentence and the infobox template prefilled with data from Wikidata. The editors can then complement the draft articles with further information before publishing them in the article namespace. This not only facilitates the work of existing contributors, but also greatly simplifies the job of new editors who participate in edit-a-thons or editing campaigns. By this means, the project team intends to leverage the power of Wikidata to also promote the writing of new Wikipedia articles about heritage institutions that have not yet been covered in a particular language. The tool consists of a template that has so far been implemented on Portuguese Wikipedia for subjects including museums, books, movies, earthquakes, newspapers and the Brazilian elections. In the course of the project, the tool will also be implemented for articles about libraries and archives, before being rolled out in other language versions.

Figure 6: Stub-article automatically created by means of the Mbabel tool

Internationalization of the Project

The internationalization of the approaches described in this article will be facilitated by the model project implemented in Brazil and on Portuguese Wikipedia, which is currently funded by the Geneva-based MY-D Foundationand by a private sponsor. As the current project funding is

limited to the implementation of the Brazilian model project and the provision of documentation, the deployment of the project in other countries and on other language versions of Wikipedia will rely on the involvement of volunteers in various countries as well as local sponsoring and/or funding through Wikimedia Foundation channels, perhaps taking a form similar to the funding of other international outreach campaigns, such as Wiki Loves Monuments.

Outlook

As illustrated in figure 1, the project provides an important cornerstone for any other activity targeting the other layers of information about heritage institutions. Thus, it could serve as a starting point for a more detailed description of archives and collections, and it extends the work that is already been carried out in other GLAM-Wiki initiatives dedicated to the description of specific heritage objects, such as the **Sum of all Paintings Project**, which repertorizes and systematically gathers information about all paintings held by heritage institutions. Another logical extension of the project lies in the development of further cooperation with individual heritage institutions to improve the coverage of their collection on Wikipedia. And, last but not least, the project may be expanded to cover other entities, such as performing arts organizations, historical monuments or cultural venues.

---

*This is Part 2 of this article. Part 1 was published here.

Reference

[1]The working title, GLAM stands for “Galleries, Libraries, Archives, Museums”; the acronym is commonly used to refer to heritage institutions.