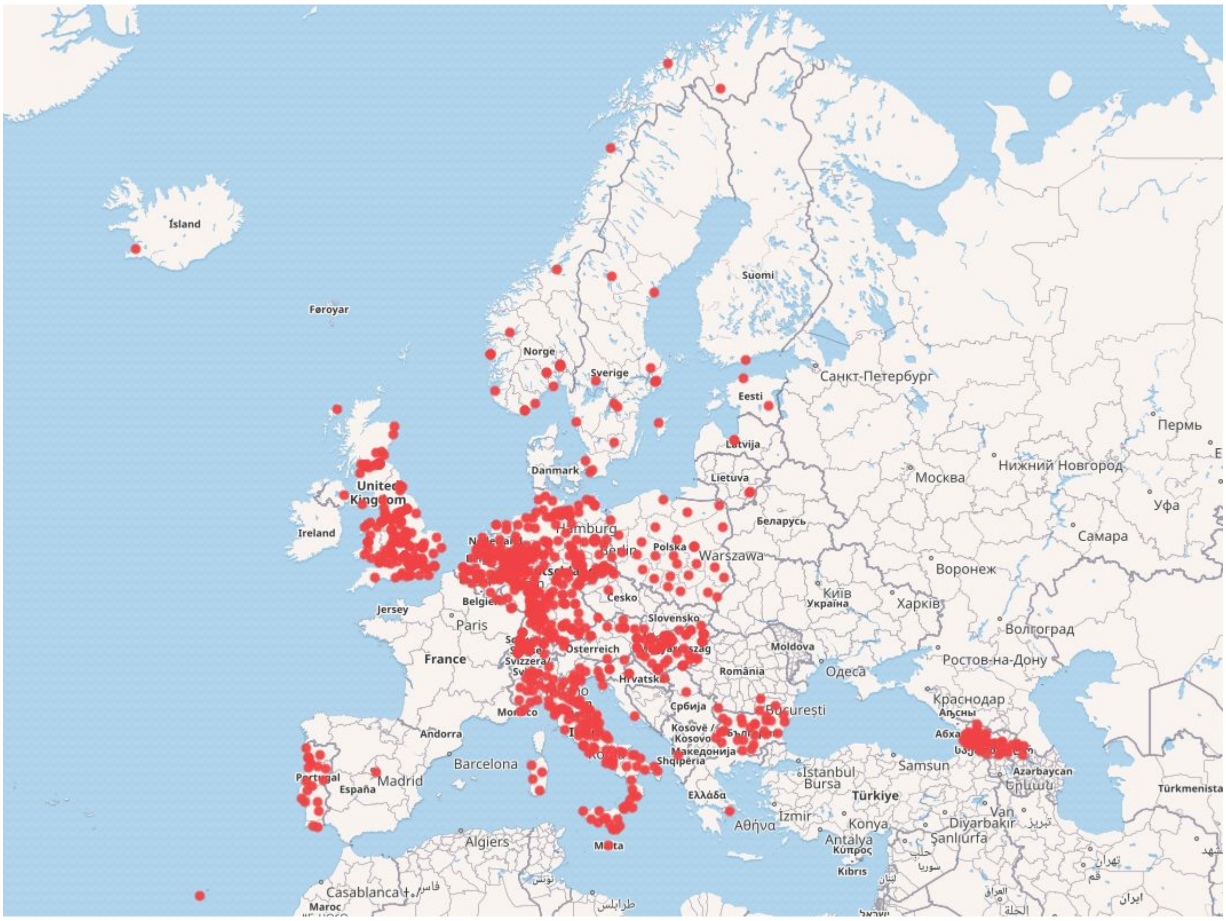
Date
2020-01-28



Archives Portal Europe institutions on Wikidata

project report

Introduction

Wikimedia Sverige

Wikimedia Sverige (WMSE, https://wikimedia.se/) is a Swedish non-profit association and one of the chapters – independent supporting organizations – of the Wikimedia Foundation. WMSE works towards making knowledge freely accessible to all humans, especially by supporting the projects of the Wikimedia Foundation, of which Wikipedia is the most well known. A key element of WMSE's work is collaboration with partners in the culture and education sectors, such as libraries, museums, schools and universities.

Wikidata

Wikidata (https://www.wikidata.org/) is one of the so-called sister projects of Wikipedia. It is a free and open database of structured data. Just like Wikipedia, Wikidata can be edited by anyone, and all the content is available under a free license. The CC0 license used on Wikidata enables anyone to re-use the content without having to attribute the source, a major difference from the Creative Commons Attribution-ShareAlike license used on Wikipedia. The data hosted on Wikidata can be displayed on all language versions of Wikipedia, but it can also easily be accessed, queried and downloaded via an API.

The data model in Wikidata is based on the concept of **items**, each of which is identified with a unique Q ID. For example, the Regional Archives in Göteborg have the Q ID Q10553859. Information about the item is conveyed using property-value pairs. In this example, the item is described with the property *country* with the value *Sweden*.

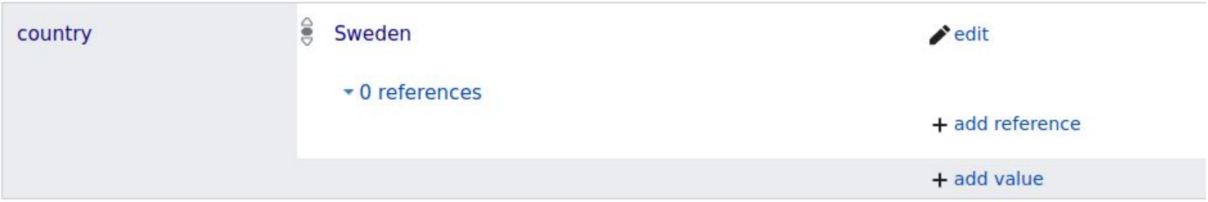


Figure 1. Example of a statement consisting of a property-value pair.

There are over 7,000 Wikidata properties, and new ones are created as needed. An important subset of them are external identifiers, which make it possible to link the items to corresponding entries in other databases, websites and services. For example, the VIAF ID (P214) has been used over 2,000,000 times. Other notable external identifiers include the

Library of Congress Authority ID (P244, 1,100,00 uses) and the International Standard Name Identifier (P213, 1,000,000 uses).

Wikidata can be queried using **SPARQL**, an RDF query language. The Wikidata Query Service (https://query.wikidata.org/) provides an interface for editing and executing SPARQL queries, the results of which can be displayed in several ways (e.g. a map, a picture gallery or a bar chart – depending on the nature of the data) and downloaded in several formats. SPARQL queries can also be executed remotely, making it possible for developers to fetch data from Wikidata using their programming language of choice.

FindingGLAMs

FindingGLAMs (https://meta.wikimedia.org/wiki/FindingGLAMs) is a project led by Wikimedia Sverige in cooperation with UNESCO and the Wikimedia Foundation, financed by a project grant from the Postcode Foundation. The project has a broad aim of improving the coverage of GLAM institutions (Galleries, Libraries, Archives and Museums) and their collections on the Wikimedia projects. A large part of the project is building a global database of cultural heritage institutions on Wikidata.

While partial databases, lists and registers of cultural heritage institutions do exist, there is no centralized, worldwide database of this kind. Most importantly, no such database exists that is available under an open license, free for anyone to access, edit and re-use in their own tools and services. Wikidata, with its flexible, content-agnostic structure and vibrant, multilingual community, has the potential to become such a database.

While volunteer editors have done a tremendous job adding information about cultural heritage institutions, large-scale uploads of existing data are necessary if we want to achieve this goal. This is why WMSE has been working on importing freely licensed data to Wikidata. For example, we have created items for over 21,000 library systems and outlets in the USA sourced from the Public Libraries Survey, which is not copyrighted due to being produced by a US government agency. We have also been reaching out to data owners to encourage them to share their data under a Wikidata-compatible open license.

Our work

In July 2019, we identified the institution directory of Archives Portal Europe (https://www.archivesportaleurope.net/en/directory) as being within the scope of our project, and we reached out to ask about its copyright status and the possibility of downloading the data in a machine-readable format. While the dataset was copyrighted, after sharing some background about our work we learned that releasing it as open data was not out of the realm of possibilities.

In November 2019, we got access to the 6869 data files that had been released under the CC0 license. The majority, 5477, were archives in Italy; other well-represented countries were the United Kingdom (319 institutions), Germany (149 institutions) and the Netherlands (114 institutions). Not all archives included in the directory on APE's website were included in the dataset, as the relicensing did not cover the entirety of the catalog.

We worked with the data in early January 2020. The open source software OpenRefine 3.3 (http://openrefine.org) was used to explore the data, match it against Wikidata items and perform the uploads. Before starting the work, a new Wikidata property was created: the Archives Portal Europe ID (P7764), making it possible to assign the unique APE identifier to any item.

Using the OpenRefine reconciliation function, we established with high probability that 612 of the institutions included in the directory already had Wikidata items. The reconciliation process looked at the labels (the names of the institutions) in connection with *country* statements to identify the items; for example an entry for *Stadtarchiv Esslingen* in Germany would be matched to an item with the same label and a *country* property with the value *Germany*. A small number of items was identified as partial matches, for example if the correct label was present but the item lacked a country statement (it is not uncommon for Wikidata items to lack crucial information); those were few enough to be verified manually. Finally, we added the APE ID to the 612 items.

By implementing this rather strict reconciliation strategy, we decreased the risk of false positives, i.e. adding information to a wrong item. On the other hand, we increased the risk of false negatives, i.e. overlooking existing items and creating new items instead of enriching those. This typically happens when the existing item has a label that is different enough from the label used for searching that the software does not consider them equivalent. This is not uncommon when matching large datasets with Wikidata; indeed, editors normally err on the side of creating duplicate items rather than mis-identifying items. Merging duplicate items is much easier than splitting overloaded items, and can be done by any Wikidata editor.

The remaining 6257 items were created from scratch. The following information was added, where available:

1. Instance of (P31) – either *archives* (Q166118) or a more specific subclass, e.g. *municipal archives* (Q604177).
2. Archives Portal Europe ID (P7764)
3. Country (P17)
4. Coordinate location (P625)
5. Located in the administrative territorial entity (P131)
6. Located at street address (P6375).
7. Official website (P856).
8. E-mail address (P968).
9. Phone number (P1329).

10. Label
11. Description

This information was also added to the existing items, where not already present.

Querying for the APE institutions

Using the APE ID property, it is easy to find the relevant items via the Wikidata Query Service. The following is a simple query that retrieves 1) the Wikidata ID, 2) the label, and 3) the APE ID of all items with an APE ID, sorted by the APE ID.

```
SELECT ?item ?itemLabel ?ape WHERE {
?item wdt:P7764 ?ape.
SERVICE wikibase:label { bd:serviceParam wikibase:language "[AUTO_LANGUAGE],
en". }
} ORDER BY ?ape
```

Results of the query: https://w.wiki/G58.

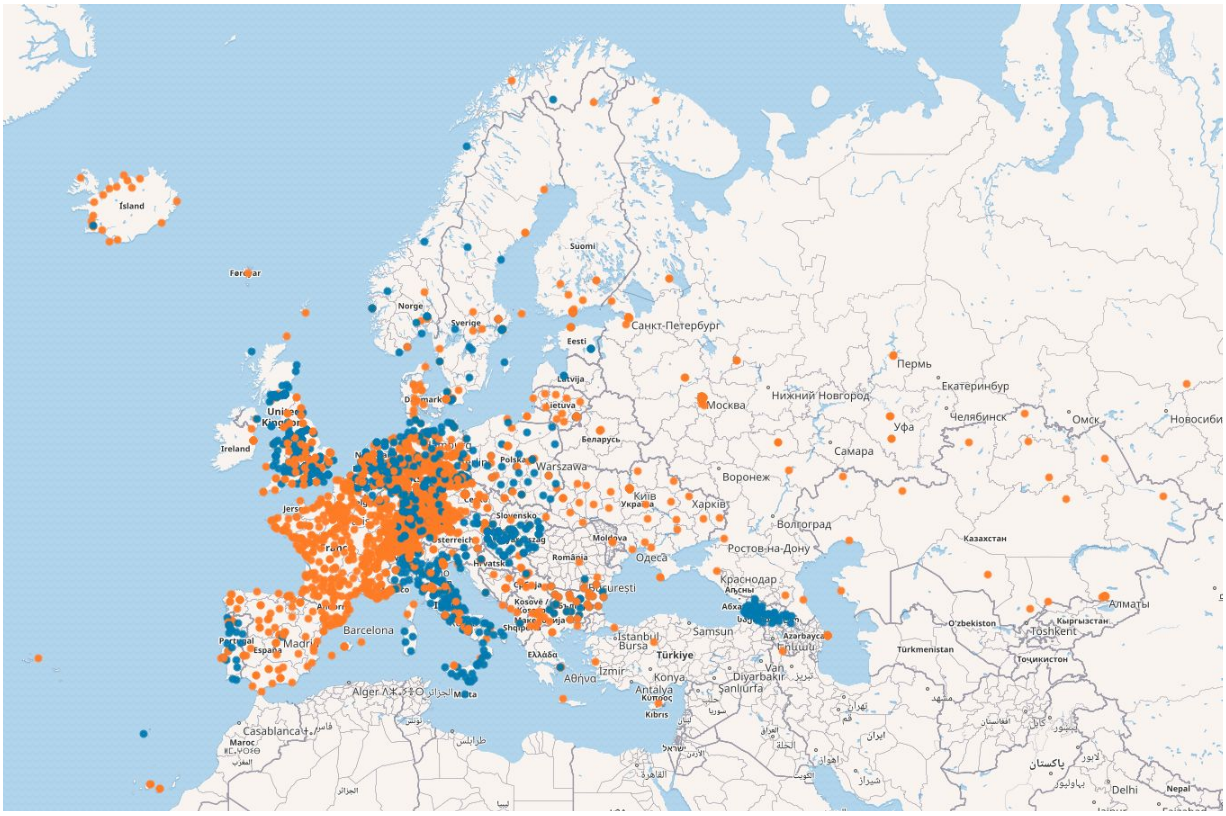


Figure 2. Archives in Europe with coordinates (3581). State on 2020-01-16. Those with an Archives Portal Europe ID are marked in blue; those without – in brown. Generated using the following query: https://w.wiki/Fav.

Suggestions for further work with Wikidata

Integrating Wikidata in the Archives Portal Europe is a possibility. Indeed, cultural heritage institutions around the world have been experimenting with integrating content from Wikidata in their own databases and services. A notable example is the Library of Congress, which in early 2019 added Wikidata identifiers to over one million entries in their authority file, noting that "[w]ith Wikidata the contributing institution is an active open community of editors" (https://blogs.loc.gov/thesignal/2019/05/integrating-wikidata-at-the-library-of-congress/).

Using data from Wikidata in an institutional database makes it possible to enrich the database with additional information, either to fill in known gaps in own data or to provide added value to the users. For example, if an entry in the directory lacks geographical coordinates, but they are present in the Wikidata item, they could be displayed. Examples of data that add value for the visitor but are not currently included in the database are images, links to Wikipedia articles and the institutions' social media handles. By fetching this data directly from Wikidata, the service can become more interesting, without putting extra burden on the developers to maintain it. The information fetched from Wikidata could be visibly marked as such in order to indicate that it was created by volunteers.

Most importantly, by integrating Wikidata into the Archives Portal Europe directory, the institution would be sending a powerful message about the value of the open knowledge movement. This opens the door for further projects and activities, such as encouraging both the data partners and the website's users to edit Wikidata themselves, or highlighting interesting content in the Wikimedia projects, such as Wikipedia articles about the archives and their collections.