Comparative Integration Potential Analyses of OSM and Wikidata – The Case Study of Railway Stations

Alishiba Dsoua[1,*], Moritz Schott[2] and Sven Lautenbach[2,3]

[1] Data Science & Intelligent Systems, University of Bonn, Bonn, Germany; dsouza@cs.uni-bonn.de
[2] Institute of Geography/GIScience, Heidelberg University, Heidelberg, Germany; moritz.schott@uni-heidelberg.de, sven.lautenbach@uni-heidelberg.de
[3] HeiGIT, Heidelberg University, Heidelberg, Germany;

* Author to whom correspondence should be addressed.
This abstract was accepted to the Academic Track of the State of the Map 2022 Conference after peer-review.

OpenStreetMap (OSM) is one of the richest and most diverse sources of geographic information. However, it lacks a fundamental property vital for spatio-semantic analyses: hierarchical structure and semantic linkage. OSM provides links to existing knowledge graphs (structured data that conforms to a specific ontology), e.g., via the *wikidata* key. The usage of these link-tags is currently limited to a small percentage of both OSM and Wikidata objects. Efforts were undertaken to enhance geographic linking (i.e., linking nearby objects of the same type) and semantic linking [1–3]. The WorldKG knowledge graph [4] provides a semantic mapping of a large subset of OSM. While the free and open OSM tagging scheme is an integral part of the OSM project that enabled its success, WorldKG overcomes the inherent lack of structure the tagging scheme represents, paving the way for a knowledge-graph integration of the OSM dataset. Still, open knowledge graphs and OSM are not fully integrated on schema and entity level.

The following analyses provide a series of comparative data insights that help to better understand the potential and implications of integration between knowledge graphs and OSM. This work compares OSM to Wikidata, one of the largest open knowledge graph projects from the Wikimedia Foundation that provides structured storage to other Wikimedia projects such as Wikipedia. Wikidata can, in many aspects, be compared to OSM in terms of its community structure, free and open nature, and simple contribution framework. In this work, the two datasets are first compared in size, data structure, and distribution. Later, we extend our analyses with a community comparison. The analyses also examine how two separate online communities with similar interests have evolved over time.

Grasping the size of the two projects is a straightforward task and visible on their websites: OSM features around 1 billion elements [5], while Wikidata is much smaller with around 97 million objects, of which approximately 9 million have geographic coordinates. Yet, the aforementioned schematic misalignment makes a comparison on dataset level

Dsouza, A., Schott, M., & Lautenbach, S. (2022). Comparative Integration Potential Analyses of OSM and Wikidata – the Case Study of Railway Stations.
In: Minghini, M., Liu, P., Li, H., Grinberger, A.Y., & Juhász, L. (Eds.). Proceedings of the Academic Track at State of the Map 2022, Florence, Italy, 19-21 August 2022. Available at https://zenodo.org/communities/sotm-22
DOI: 10.5281/zenodo.7004483

impractical. OSM tags allow no predefined distinction between the classes and attributes of an object. While lists exist that facilitate the distinction (e.g., Map Features https://wiki.openstreetmap.org/wiki/Map_features), it is not clear on a structural level, e.g., for an object with *highway=elevator* as a tag, the user must know the tag's meaning to extract the correct class. Wikidata classes are defined using an attribute called "instance of" (abbreviation P31). For example, a mosque in Wikidata will have the attribute "instance of religious building". In addition to the incompatible schemata, uneven distribution of classes, community interests, and priorities are some of the biggest challenges for data integration. Wikidata is not limited to geographic entities, wherefore humans (people of interest) represent the largest data type. Therefore, while the analyses at the dataset level as a whole seem possible, they can not be applied to all classes due to varied definitions, incomplete or missing representations, or lack of other comparison parameters such as geographic coordinates.

The topic of railway stations was chosen in the presented study because these objects have a comparable definition and are well represented in both datasets with ca. 130k and 100k elements in OSM and Wikidata, respectively, indicating integration potential. In OSM, railway stations are mapped by the tags *railway=station* or *railway=halt*. In Wikidata, the 'instance of Q55488 (railway station)' value represents Railway Stations.

The presented work (available under the GNU Affero General Public License v3 at https://gitlab.gistools.geog.uni-heidelberg.de/giscience/ideal-vgi/osm-wikidata-comparison) provides a set of generalizable indicators for VGI project description, comparison, and monitoring. Similar approaches have been established for OSM contributors [6], single OSM elements [7], and small geographic regions [8]. For data collection in Wikidata, Wikidata API (https://www.wikidata.org/w/api.php) and Wikidata SPARQL endpoint were used. For Wikidata objects mapped with 'Railway Station', their revision history containing user information, timestamps, and the number of properties was collected. Overall contributions were collected from all users who have contributed to at least one object typed 'Railway Station'. OSM data collection was done using the ohsome framework (https://ohsome.org) to extract all railway stations mapped in OSM, including their history and all edits made by the users who edited these railway stations. In addition to a general comparison between the datasets, we defined five subsets for a more detailed comparison: OSM railway stations with links to Wikidata (59,441 elements), OSM stations without links to Wikidata (74,659), Wikidata stations that have links from OSM (45,050), Wikidata stations without links to OSM but with geocoordinates (54,594) and Wikidata stations without links to OSM and without geocoordinates (6,714).

Our first analysis regarding the growth rate of the two sources showed that OSM is reaching a saturated state regarding the number of railway stations (see Figure 1), where relatively few stations have been added since 2016. Wikidata, on the other hand, still experiences a stable number of new stations that are added to the project. The two datasets depict no clear temporal correlation hinting towards two independent communities, meaning that additions in OSM are not followed by additions in Wikidata and vice versa. This lack of community integration is also true for the subset of linked OSM stations. Yet, this subgroup records a stable growth meaning that more and more OSM stations are explicitly linked to their Wikidata counterpart. Despite the similar size of the two datasets on a global scale, they show significant discrepancies on a country level. For example, in China, Wikidata features only 39% of the stations present in OSM while having more than double the amount

of stations in the United Kingdom. While the lack of stations seems reasonable considering the overall lack of stations in Wikidata, the overabundance of stations in the UK hints toward data issues, such as the misclassification of tram stops, that need more detailed analyses before integration.

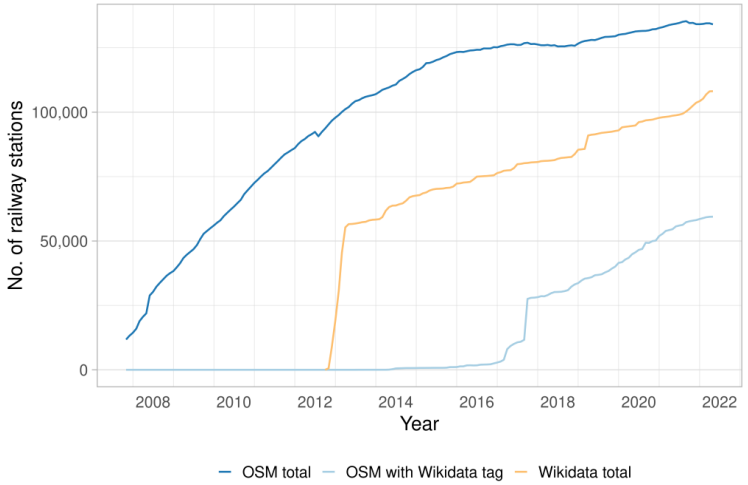

Figure 1. Growth Rate of OSM and Wikidata.

Regarding the properties/tags of each object, we observed that Wikidata has, on average, more properties per object than OSM. One reason could be the project's goal being knowledge collection rather than object location. Since Wikidata is a knowledge graph, it also contains links to other objects that help enrich existing objects increasing this discrepancy even further. OSM objects with links to Wikidatda have almost double the tags compared to those without links. This could be a quality indicator or an indicator that only famous stations, which are very well mapped in OSM, are also linked to Wikidata. Wikidata objects without links from OSM and geocoordinates have the least number of properties, hinting at their lower quality.

Next, we present the community analysis. There were around 1.8 million contributors in OSM in total, and 48k unique users have contributed to either creation, deletion, or updating of the railway station objects. In Wikidata, the number of overall contributors was much smaller, i.e., 24k out of which 14k have contributed to Railway objects. The revisions for Wikidata objects were around 11 times higher than that of OSM revisions. This is evident as Wikidata railway stations have more properties than OSM railway stations. This could also be because OSM contributors have a wide variety of objects to map, ranging from benches to land-use. In addition, adding a new object to the map may take priority over extensive tagging of existing objects. In contrast, Wikidata contributors may focus on details and enrichment of prominent objects of public interest. A similar trend was observed for average stations created by each contributor, wherein Wikidata contributors have created five times and, with median statistics, two times more objects than OSM contributors. This may be due to the higher number of bots and imports in Wikidata. While OSM users generally map a specific area that can only feature a limited number of railway stations, Wikidata users may import railway stations from other sources without limiting themselves to a certain geographical unit. This notion is supported when looking at the specialization of railway station contributors by calculating the share of edits made to railway stations to the

total amount of edits made by a user. OSM users are less specialized than Wikidata users having only 1-2% of their edits in this domain, while Wikidata users had around 5% of their contributions in this domain. OSM users seem to often be more generalists with edits in all domains, in a certain region, while Wikidata users are more topic-driven contributors.

We notice that both communities have great potential to integrate these sources on the topic of railway stations. Although this potential increases daily with other topics reaching a mature data state in Wikidata, it is difficult to generalize our work to the entire datasets as the purpose of the datasets differs widely. OSM is focused entirely on spatial data, whereas Wikidata is a general-purpose knowledge graph. Therefore data content and style will always be tailored towards these goals and make integration a difficult task. To ensure the generalisability of our analyses to other topics, users must ensure that the data is similarly represented in both datasets in terms of the class definition and geometric representation. We also need to acknowledge that there are topics that (currently) have no potential for integration or comparison. E.g., although land-use and land-cover information is a prominent type of data in OSM, this data type is inexistent in Wikidata. The information on "what is a forest" may be present in Wikidata but is incomparable to land use polygons in OSM. To conclude, our observations of the performed analyses show that OSM can benefit from the wide range of semantic information linked to objects, while Wikidata can benefit from the precise geoinformation and completeness of OSM. The analyses can also benefit the semantic web and GIS communities by giving them insights into the datasets which can help integrate datasets. In the future, we would like to expand our work to prominent classes such as places (cities and other named locations) with additional comparison parameters.

Acknowledgements: This work was funded by DFG, German Research Foundation ("WorldKG", 424985896 and "IdealVGI", 424966858).

References

[1] Tempelmeier, N., & Demidova, E. (2021). Linking OpenStreetMap with knowledge graphs—Link discovery for schema-agnostic volunteered geographic information. *Future Generation Computer Systems*, *116*, 349–364.

[2] Dsouza, A., Tempelmeier, N., & Demidova, E. (2021). Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs. In: *Proceedings of the 20th International Semantic Web Conference, ISWC 2021*, 24-28 October 2021, Springer, Cham, 56–73.

[3] Gurtovoy, D., & Gottschalk, S. (2022). Linking Streets in OpenStreetMap to Persons in Wikidata. In: *Proceedings of The Web Conference 2022*, Lyon, France, 25-29 April 2022.

[4] Dsouza, A., Tempelmeier, N., Yu, R., Gottschalk, S., & Demidova, E. (2021). WorldKG: A World-Scale Geographic Knowledge Graph. In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, Queensland, Australia, 1-5 November 2021, 4475–4484.

[5] Schott, M., Herfort, B., Troilo, R., & Raifer, M. (2022). A basic guide to OSM data filtering. *GIScience News Blog*. Retrieved from http://k1z.blog.uni-heidelberg.de/2022/01/20/a-basic-guide-to-osm-data-filtering

[6] Schott, M., Grinberger, A.Y., Lautenbach, S., & Zipf, A. (2021). The Impact of Community Happenings in OpenStreetMap—Establishing a Framework for Online Community Member Activity Analyses. *ISPRS International Journal of Geo-Information*, *10*(3), 164.

[7] Schott, M., Größchen, L., & Lautenbach, S. (2022). OSM Element Vectorisation. Retrieved from https://gitlab.gistools.geog.uni-heidelberg.de/giscience/ideal-vgi/osm-element-vectorisation

[8] GIScience Research Group and HeiGIT (2022). *ohsome quality analyst*. Retrieved from https://github.com/GIScience/ohsome-quality-analyst