Automatically Constructing Geospatial Feature Taxonomies from *OpenStreetMap* Data

Basel Shbita
Information Sciences Institute
Department of Computer Science
University of Southern California
Marina del Rey, California
shbita@usc.edu

Craig A. Knoblock
Information Sciences Institute
Department of Computer Science
University of Southern California
Marina del Rey, California
knoblock@isi.edu

***Abstract*—This paper presents a method for constructing a lightweight taxonomy of geospatial features using *OpenStreetMap* (OSM) data. Leveraging the OSM data model, our process mines frequent tags to efficiently produce a structured hierarchy, enriching the semantic representation of geo-features. This data-driven taxonomy supports various geospatial analysis applications. Accompanying the methodology, we release the source code of our tool and demonstrate its practical application with tailored taxonomies for California (US) and Greece, underscoring our approach's adaptability and scalability.**

***Index Terms*—geospatial data, semantic analysis, taxonomy**

In the era of digital mapping and geographic information systems (GIS), the availability of accurate and comprehensive spatial data is crucial for various applications, ranging from urban planning to scientific research [4], [9], [10], [12]. *OpenStreetMap*[1] (OSM) has emerged as a community-driven initiative to provide free and open access to global spatial data, making it the richest publicly available information source on geographic entities worldwide. However, using OSM data in downstream applications is challenging due to the large scale of OSM, the heterogeneity of entity annotations, and the absence of a standardized ontology to describe entity semantics [2]. Our taxonomy supports applications from automated navigation systems, which require precise geographical feature recognition for route optimization, to the classification of remotely sensed data, enhancing both the integration and utility of OSM data in sophisticated GIS applications.

Leveraging the concept of Volunteered Geographic Information (VGI) [3], OSM relies on user contributions to map the geometries and attributes of both natural and urban features. While OSM has proven to be a valuable resource, certain limitations hinder its full potential [7]. The utility of OSM data heavily relies on the consistent tagging of geographical entities by its users, as the platform does not impose restrictions on tag choices. Instead, OSM encourages its contributors to follow a set of best practices for annotation, leading to a highly heterogeneous landscape of tags. The number of tags and the level of detail for individual OSM entities is highly variable. Figure 1 provides an illustration of various building types within a neighborhood, selected from OSM, showcasing the most specific tags associated with them. However, OSM lacks a system that establishes relationships between these tags, hindering the extraction of valuable insights. As a result, the lack of clear semantics not only hinders the interoperability of OSM data with other datasets but also severely limits its usability in various applications. To overcome these limitations, it is crucial to establish a comprehensive taxonomy extractor from this dynamic data. This will enable better integration with other datasets and facilitate the effective utilization of the data for diverse scientific and practical purposes.

We address the limitations above and unlock the full potential of OSM data by proposing an approach to structure a taxonomy of geo-feature types from a given OSM data dump. We demonstrate this approach by creating a comprehensive and well-defined taxonomy of geospatial features derived from OpenStreetMap (OSM) data, utilizing a tool that we have made available as open source[2]. This taxonomy will enable users to understand the connections and categorization of different types of features, facilitating detailed analysis and utilization of the data. This approach enables better integration of OSM data into machine learning models and broadens its application, unlocking new opportunities to harness OSM's rich informational spectrum in diverse domains.

I. Constructing OSM Taxonomies

A. The OpenStreetMap Data Model

To understand our proposed approach, it is crucial to elucidate the structure of the OSM data. Initiated in 2004 as a collaborative project, OSM strives to generate a publicly accessible vector map encompassing the entire world. Remarkably successful, the project has over 10 million registered users as of March 2023. In the OSM data model, each feature is represented as one or more geometries (nodes, ways, and relations) with attached attribute data, which contains meaningful information for the taxonomy construction. Attribute information is stored as tags associated with geographic entities in the form of key-value pairs. As OSM does not prescribe a fixed set of tags, meticulous filtering becomes imperative to include only pertinent information. The comprehensiveness and diversity of features available in OSM can exhibit substantial regional

[1]https://www.openstreetmap.org/

[2]https://github.com/basels/osm-taxonomy

2472-9671/24/$31.00 ©2024 IEEE
DOI 10.1109/ICSC59802.2024.00039

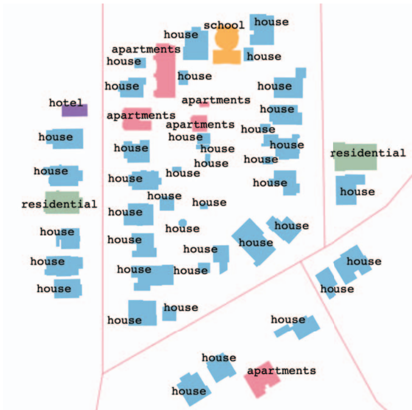


Fig. 1. Simplified illustration of a neighborhood within *OpenStreetMap* with different `building` feature sub-types, depicting instances with type `house` in blue, `apartments` in red, `residential` in green, `school` in orange, and `hotel` in purple.

variations due to the contributions of volunteers. While this data model is adequate for numerous applications, it lacks a meaningful structure between tags or their interrelationships, which constitutes the focal point of this work.

B. Identification of Meaningful Tags

The initial phase of our methodology centers around the preprocessing of OSM and reducing the set of its attribute data into a meaningful one. This dataset consists of a wide range of tags contributed by OSM users, encompassing both suggested and self-defined tags [5]. To illustrate the magnitude of this tag diversity, let us consider a recent snapshot of OSM data for California from March 2023, where we encountered an overwhelming 3,000 unique tags. Given the extensive and heterogeneous nature of these tags, it becomes imperative to establish a concise and representative set of target labels that would serve as the foundation for constructing the taxonomy.

During the identification process, we encounter two distinct challenges. First, we need to address the issue of frequent tags that are non-informative. For instance, the `name` tag, which typically provides the name of a geo-entity (e.g., "The Ritz-Carlton"), or the `maxspeed` tag, commonly associated with road features to indicate the maximum allowable driving speed. To address this issue, we identify the most commonly used tags from a user-centric viewpoint and manually curate a set of "blacklisted" tags. This list is included with the tool and can be easily modified to accommodate different domains or specific user preferences.

Secondly, we confront the challenge of infrequent tags that may possess informative characteristics but are inadequately represented within the dataset. To mitigate this issue, we apply a frequency cutoff threshold, effectively filtering out less common and idiosyncratic tags, and focusing exclusively on the most prevalent and significant ones. For instance, consider the key-value pair `leisure=sauna` which describes a specific subtype of leisure. In the recent California OSM snapshot, this pair appeared fewer than 10 times. Consequently, it was not considered for inclusion in the final taxonomy. Through these meticulous processes, we strike a delicate balance between inclusiveness and practicality, ensuring that the resulting taxonomy faithfully represents the prominent geo-features while avoiding an unwieldy and unmanageable taxonomy structure.

C. Establishment of Hierarchical Relationships

We present our methodology for establishing taxonomic parent-child relationships among various geo-features using the OSM data. The objective is to construct a hierarchical taxonomy of labels based on frequent tag assignments. To accomplish this task, we implemented Algorithm 1, which takes the OSM snapshot data as input and produces a desired taxonomy tree data structure.

The algorithm operates as follows. Following the initial processing and removal of the undesirable tags (as described in Section I-B), we iterate through the dataset, creating a key-value path counter to identify commonly occurring tag assignments (lines 3-8). These paths serve as the foundation for defining parent-child relationships within the tree structure, thereby shaping the taxonomic hierarchy. For instance, consider the set of tags `{highway=service, service=driveway}` which forms the path `highway--service--driveway`. In this case, the unique parent-child paths are `highway--service` and `service--driveway`.

Subsequently, we insert paths into the tree, prioritizing the most frequent ones (line 9). To maintain consistency and address any ambiguities, we handle instances where multiple paths may conflict with the evolving tree structure by favoring the more frequently occurring path and omitting the less common one (line 10). Moreover, when integrating a parent-child path into the tree, if a child tag appears under different parent tags for distinct entities, we replicate it with a unique identifier (line 13). For instance, the tag (key or value) `residential` may pertain to both `highway` and `building`; in such cases, they would be distinctly labeled as `residential_highway` and `residential_building`, respectively.

Algorithm 1: Constructing a lightweight taxonomy.

```
   Data: osmDataset
   Result: taxonomyTree
1  taxonomyTree = create_empty_tree();
2  tagPathsCounter = Counter();
3  for entity in osmDataset do
4      tags = entity.get_tags(); // key-value pairs
5      filteredTags = filter_tags(tags);
6      if filteredTags is not empty then
7          tagPath = create_tag_path(filteredTags);
8          tagPathsCounter[tagPath]++;
9  for (tagPath,count) in tagPathsCounter.sort(order=descending) do
10     if is_path_consistent_with_tree(tagPath, taxonomyTree) then
11         parent, child = extract_parent_and_child(tagPath);
12         if parent is not null and child is not null then
13             insert_parent_child_pair(taxonomyTree, parent, child);
14 return taxonomyTree;
```

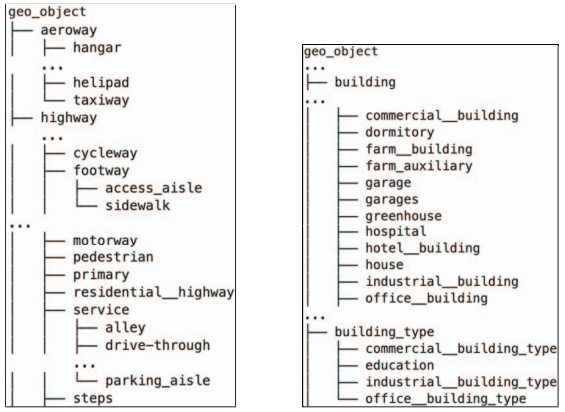
By following this process, we construct a taxonomy tree that encompasses a comprehensive representation of geo-features within the OSM dataset. The resulting taxonomy allows for a more nuanced understanding of their interrelationships.

II. Qualitative Evaluation

To assess the effectiveness of our proposed method for constructing a lightweight taxonomy of geographic features using OSM data, we conducted an experiment utilizing two comprehensive OSM datasets in the form of `.osm` dump (snapshot) files. Our evaluation consists of a qualitative analyses, providing insights into the resulting taxonomies generated from each dataset and comparing them on a surface level. The first dataset we employed comprised the complete California (US) OSM geo-data snapshot from March 2023[3], encompassing approximately 150 million OSM instances. Among these instances, approximately 10 million contained at least one tag, with 1 million being nodes, 9 million being ways, and around 68,000 being relations. The number of tags assigned to each instance varied from 1 to 16, with an average of 2.3 tags per geographic entity. The second dataset consisted of the complete Greece OSM snapshot from March 2023[4], which included approximately 40 million OSM instances. Around 2 million contained at least one tag, with 266,000 being nodes, 1.7 million being ways, and around 18,000 being relations. Similarly, the number of tags assigned to each instance ranged from 1 to 13, with an average of 2.1 tags per geographic entity. Each dataset comprised around 3,000 unique labels. For both datasets, we established a minimum threshold of 500 instances per tag for the purpose of our analysis.

To evaluate the resulting taxonomies, we performed a qualitative analysis, examining them from a user perspective to assess their coherence and utility. The qualitative analysis revealed several positive findings regarding the constructed taxonomies. In both cases, the taxonomies successfully captured the essential characteristics of the geographic features within the OSM datasets, providing a structured and organized representation. A snippet from the resulting taxonomy generated from the California dataset is depicted in Figure 2a, demonstrating the hierarchical relationships between tags that facilitated the classification of diverse types of geo-features, such as `aeroway` and `highway`. This hierarchical structure enhanced the comprehension and interpretation of the roles and functions of these features. Furthermore, the taxonomy facilitated the differentiation of various sub-types within the same feature category, such as `cycleway` and `footway`, as well as different types of `service` ways, including `alley` and `drive-through`, as illustrated in the same figure.

From a human perspective, the resulting taxonomies exhibited both accurate and inaccurate taxonomic relations. It accurately captured hierarchical relationships between categories in certain domains, such as transportation (e.g., `highway - residential_highway`, `aeroway - taxiway`) and amenities (e.g., `amenity - restaurant`, `leisure - park`). The resulting relationships reflected intuitive groupings and aligned with human understanding. However, certain inaccuracies were observed in the taxonomy, likely stemming from its automatic generation. Figure 2b illustrates an example where the taxonomic relation between `building` and `building_type` is redundant and does not provide additional meaningful information.



(a) Snippet from the California taxonomy showing accurate hierarchical relationships

(b) Snippet from the resulting California taxonomy showing redundant taxonomic terms and relations.

Fig. 2. Taxonomy snippets from the resulting California (US) dataset.

Furthermore, we conducted a surface-level comparison between the resulting taxonomies derived from both datasets. This comparison highlighted how different geographical locations, such as countries or states, can yield distinct results. Figure 3 presents a textual comparison between the resulting taxonomies, demonstrating the differences between California and Greece. For example, the `historic` category in Greece encompasses types of features, such as `archaeological_site` and `castle`, which are either uncommon or nonexistent in California. Furthermore, the usage of the `internet_access=wlan` tag by OSM users in Greece was much more prevalent compared to the California dataset. Additionally, the presence of the `kerb` tag category, encompassing different types of the feature (e.g., `flush` and `raised`), was observed in the taxonomy resulting from the California dataset but not in the Greece dataset. These variations in the taxonomies could be attributed to factors such as tagging style, cultural differences, and historical context. The full taxonomy text files generated from both experimental datasets are also available in our repository.[5]

The inaccuracies observed in the taxonomy can be attributed to the limitations of automatic generation. While automated approaches can be efficient, they often lack the contextual understanding and domain knowledge possessed by humans. The absence of human judgment and expert curation during the automatic generation process can result in inconsistencies and illogical relationships within the taxonomy.

[3]https://download.geofabrik.de/north-america/us/california.html

[4]https://download.geofabrik.de/europe/greece.html

[5]https://github.com/basels/osm-taxonomy/tree/main/data

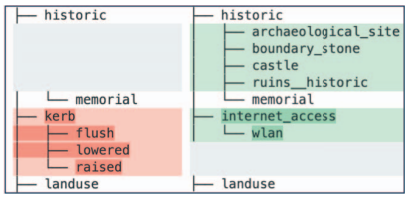


Fig. 3. Textual comparison of snippets from two resulting taxonomies: California (US) on the left and Greece on the right.

To enhance the accuracy of automatically generated taxonomies, while still benefiting from the efficiency of the automated process, it is useful to integrate human oversight and expert input at key stages. Combining automated techniques with strategic human validation and refinement can help identify and rectify inaccuracies without undermining the automation's extensive groundwork. Moreover, incorporating domain-specific knowledge and user feedback can further improve the quality and coherence of the generated taxonomy.

The semantic representation of the taxonomy offers a meaningful utility for OSM, addressing the limitations associated with unstructured tags, noise, inconsistencies, and the requirement of domain knowledge within the OSM suggested schema, which is vast and constantly evolving. Consequently, it provides a comprehensive framework for categorizing different types of features.

III. Related Work

Various approaches have been employed to construct ontologies suitable for geographical data, introducing more structure. Sun et al. [6] have developed a three-level ontology for geospatial data that, although potentially reusable, requires completion and quality assessment through manual work. Similarly, Codescu et al. [1] have created OSMonto, an ontology for *OpenStreetMap* tags that facilitates the exploration of tag hierarchies and relationships with other ontologies, but also requires manual effort. In contrast, our research initially surpasses ontology development by automatically constructing a lightweight taxonomy as a foundational step, which is then refined with minimal human intervention. This balance of our approach being primarily automatic while still benefiting from human expertise not only sets it apart from these works but also leads to a more targeted representation of geospatial features, enhancing the analysis and utilization of OSM data.

In the domain of leveraging OSM data and constructing structured taxonomies for geospatial features, WorldKG [8] is a geographic knowledge graph that provides a comprehensive semantic representation of geographic entities from OSM. Dsouza et al. [11] further leveraged WorldKG to develop a neural architecture that exploits a shared latent space for effective tag-to-class alignment of OSM entities. Building on these pivotal contributions, our methodology enriches this line of research by dynamically constructing taxonomies that can assimilate OSM data from any time frame, ensuring an up-to-date and adaptive representation, thus underlining the enduring significance of data alignment and structure.

IV. Conclusion and Future Work

In this paper, we presented an automatic approach for constructing lightweight taxonomies of geospatial features from *OpenStreetMap* (OSM) data. By leveraging the OSM data model and identifying frequent tags, we established hierarchical relationships to create a structured taxonomy. Our approach addresses the limitations of unstructured tags in OSM and enhances the semantic representation of geo-features. Through qualitative analysis, we demonstrated the effectiveness and utility of the constructed taxonomy in facilitating the classification and interpretation of diverse types of geo-features, such as buildings and highways.

Future work can focus on improving the scalability and efficiency of the taxonomy construction process, incorporating machine learning and natural language processing techniques to handle ambiguous tags, refining the taxonomy with user feedback and domain-specific knowledge, and integrating it into various geospatial analysis applications. These efforts will advance the understanding and utilization of OSM data for a wide range of geospatial applications, paving the way for more efficient and accurate geospatial analysis workflows.

References

[1] Codescu, M., Horsinka, G., Kutz, O., Mossakowski, T. and Rau, R., 2011. Osmonto-an ontology of openstreetmap tags. *State of the map Europe (SOTM-EU)*, 2011, pp.23-24.

[2] Touya, G. and Reimer, A., 2015. Inferring the scale of OpenStreetMap features. *OpenStreetMap in GIScience: Experiences, research, and applications*, pp.81-99.

[3] Kunze, C. and Hecht, R., 2015. Semantic enrichment of building data with volunteered geographic information to improve mappings of dwelling units and population. *Computers, Environment and Urban Systems*, 53, pp.4-18.

[4] Cao, R., Zhu, J., Tu, W., Li, Q., Cao, J., Liu, B., Zhang, Q. and Qiu, G., 2018. Integrating aerial and street view images for urban land use classification. *Remote Sensing*, 10(10), p.1553.

[5] Minghini, M. and Frassinelli, F., 2019. OpenStreetMap history for intrinsic quality assessment: Is OSM up-to-date?. *Open Geospatial Data, Software and Standards*, 4(1), pp.1-17.

[6] Sun, K., Zhu, Y., Pan, P., Hou, Z., Wang, D., Li, W. and Song, J., 2019. Geospatial data ontology: the semantic foundation of geospatial data integration and sharing. *Big Earth Data*, 3(3), pp.269-296.

[7] Vargas-Munoz, J.E., Srivastava, S., Tuia, D. and Falcao, A.X., 2020. OpenStreetMap: Challenges and opportunities in machine learning and remote sensing. *IEEE Geoscience and Remote Sensing Magazine*, 9(1), pp.184-199.

[8] Dsouza, A., Tempelmeier, N., Yu, R., Gottschalk, S. and Demidova, E., 2021, October. Worldkg: A world-scale geographic knowledge graph. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management* (pp. 4475-4484).

[9] Uhl, J.H., Leyk, S., Li, Z., Duan, W., Shbita, B., Chiang, Y.Y. and Knoblock, C.A., 2021. Combining remote-sensing-derived data and historical maps for long-term back-casting of urban extents. *Remote sensing*, 13(18), p.3672.

[10] Li, J., Qin, H., Wang, J. and Li, J., 2021. Openstreetmap-based autonomous navigation for the four wheel-legged robot via 3d-lidar and ccd camera. *IEEE Transactions on Industrial Electronics*, 69(3), pp.2708-2717.

[11] Dsouza, A., Tempelmeier, N. and Demidova, E., 2021, September. Towards Neural Schema Alignment for OpenStreetMap and Knowledge Graphs. In International Semantic Web Conference (pp. 56-73). Cham: Springer International Publishing.

[12] Shbita, B., Knoblock, C.A., Duan, W., Chiang, Y.Y., Uhl, J.H. and Leyk, S., 2023. Building spatio-temporal knowledge graphs from vectorized topographic historical maps. *Semantic Web*, 14(3), pp.527-549.