Towards Open-Source Maps Metadata

Areeg Mostafa*
mosta041@umn.edu
University of Minnesota, USA

Mohamed F. Mokbel*
mokbel@umn.edu
University of Minnesota, USA

ABSTRACT

This paper envisions having an open-source web portal for detailed worldwide road network maps with rich metadata. This would be major advancement from current portals that only have road networks without important metadata, including traffic-related ones. The envisioned portal will not only enable researchers to exploit more practical research, but would also enable practitioners and small/medium enterprises to avoid the high cost of commercial maps. The paper presents eight directions that can be exploited towards realizing the vision and acts as an invitation to the community to exploit these directions.

CCS CONCEPTS

• **Information systems** → Location based services.

KEYWORDS

OpenStreetMap, Metadata, Map, Road Network, Spatial-temporal

ACM Reference Format:
Areeg Mostafa and Mohamed F. Mokbel. 2023. Towards Open-Source Maps Metadata. In *The 31st ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '23), November 13–16, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3589132.3625576

1 CURRENT STATUS

Having access to accurate digital maps have empowered widely used applications, including transportation, road network routing, location-based services, ride-sharing, food delivery, and last-mile delivery. It used to be the case that accurate digital maps are only built and sold by major industry, e.g., NavTeq in USA (later become Nokia, then part of HERE) [11] and TeleAtlas in Europe (now, part of TomTom) [13]. However, the high cost and proprietary nature of commercial maps along with their inherent inaccuracy due to not being able to be frequently updated, made researchers, developers, practitioners, and enterprises turn their attention towards open source maps [12, 21]. A prime example of such maps is OpenStreetMap (OSM) [27], known as the Wikipedia of maps. OSM is a platform for crowdsourcing-based maps that has recently replaced commercial providers in various sectors of academia, government, and industry [22, 24]. For example, Facebook uses OSM as its backbone mapping support [9], Lyft has described OSM as the "Freshest Map for Rideshare" [19], while Tesla uses OSM for its routing [38]. All these companies, and many others including Apple, Amazon, Mapbox, Microsoft, and Uber, are not only using OSM, but are also extensively contributing to it [1, 8].

Even though OSM is deemed accurate and has become the defacto map infrastructure for various governmental and industrial sectors, we acknowledge that an accurate topological map is only necessary but not sufficient for accurate map services. For example, path finding (i.e., routing), as the most commonly used map service, would need to understand map metadata as much as it understands the topological map itself. One particular type of metadata that is immensely needed by all routing algorithms is the edge weights associated with each road segment, which indicates the time needed to travel throughout the road segment. A map service provider that employs the most efficient shortest path algorithms (e.g., [17, 34, 44, 45]), and is equipped with the most accurate topological map, would still provide inaccurate routing, if it has inaccurate edge weights. Other kinds of metadata including turn restrictions, u-turns, directions, and many others would also affect routing, among other map services.

2 THE VISION

We envision having a full open-source map portal, parallel to OSM, but fully enriched with all sorts of metadata. We distinguish between two kinds of metadata, *structural* and *functional*. Structural metadata refers to road segment characteristics, including length, number of lanes, stop signs, driving directions, and turn restrictions. Functional metadata refers to traffic-related data, including average speed, standard deviation, edge weight, and energy consumption. While each *structural* metadata is computed as one value per road segment, each *functional* metadata is computed as one value per time granularity (e.g., hour) per road segment. OSM maintains some structural metadata, yet with very poor coverage in terms of number of segments that have it [16, 23, 33]. Meanwhile, OSM does not maintain any kind of functional metadata.

Should this vision succeed, it would be transformative for several applications and research communities who are extensively dealing with road networks, including, transportation, urban computing, and location-based services. This will be similar to the impact that OSM has made for these communities over the last two decades. Before OSM, obtaining worldwide digital maps was only available to those who can afford buying it or have access to it through their employers. This was a major block hindering the advances in research projects that require digital maps. We believe that our envisioned portal would have similar impact as it would pave the way for researchers, developers, and practitioners worldwide to have

*The work of these authors is partially supported by the National Science Foundation (NSF), USA, under grants IIS-1907855 and IIS-2203553.

SIGSPATIAL '23, November 13–16, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0168-9/23/11…$15.00
https://doi.org/10.1145/3589132.3625576

free access to worldwide maps enriched with accurate structural and functional metadata. This will not only empower academic research, but it will also support small/medium enterprises to make their products affordable by removing the significantly high cost for having access to accurate maps metadata.

For a particular example, consider one of the basic and most common map services, namely, *shortest path*, where the objective is to find the shortest path between source and destination points. With lots of efforts dedicated to this problem [45, 48], only major enterprises (e.g., Apple and Google Maps) are able to provide accurate answers as they own the traffic metadata on top of the topological map. The reason they have such metadata is that they can easily collect it through their own devices. Though there are many attempts to have open-source routing engines (e.g., OSRM [30], GraphHopper [14], among others [25]), they all fall short to commercial ones as they are all based on OpenStreetMap (OSM), which does not maintain any functional metadata. Hence, such routing engines mainly try to guess the weight metadata for each edge, either as the maximum edge speed (if known) or some heuristic value. This a major hindering block for further research in this area. This is why many major industry had to add their traffic metadata layer as their own proprietary layer on top of OSM [10, 18, 39]. Apparently, this is out of reach from academics, practitioners, and small/medium enterprises, whom we target by our vision.

3 THE ROAD TO THE VISION

To realize our vision, a first step would be to build a dedicated web portal, similar to OSM, where users worldwide can collectively enrich the map metadata anywhere in the world. Unlike OSM where only manually curated map updates are allowed [20, 29], our vision is to allow updates coming from data-driven algorithms and machine learning (ML) modules. We believe that the community wisdom will lead to enhanced accuracy over time. In this section, we present nine directions for map metadata inference. The first four directions aim to do so by harvesting publicly available data. The fifth direction calls for incentives for volunteer users. The sixth to eighth directions aim to learn from the metadata inferred from the first five directions to fill in the remaining missing metadata. The last direction aims for quality assessment of the learning approaches. This section acts as an invitation to the community to exploit these directions, and even come up with more directions, all geared towards map metadata inference, as a means of realizing the vision of building an open-source map metadata portal.

3.1 Direction 1: Fine-grained Trajectories

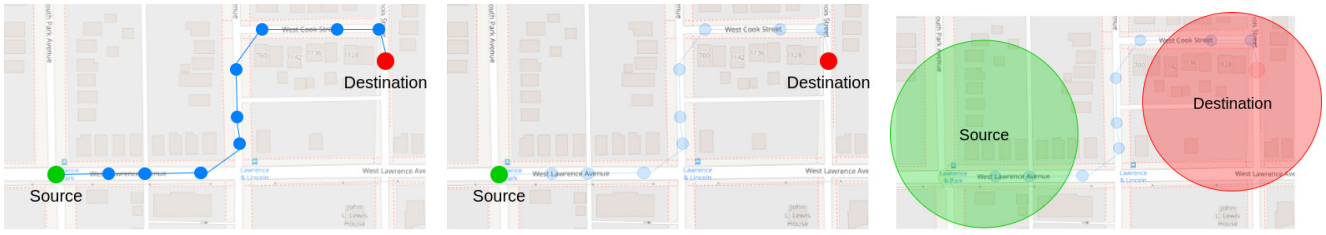
Various groups have recently released detailed trajectory data for few cities, including Athens, Greece [31], Beijing, China [49], San Francisco, USA [36], and Singapore [46]. Figure 1(a) is an example of a detailed trajectory where there is as many GPS points as possible between source and destination points. Inferring some *functional* metadata (e.g., edge weights) from detailed trajectories is pretty straightforward by first doing a trajectory map matching [5, 15] to layout the trajectories on the map, and then use the trajectories to infer traffic-related statistics for the covered road segments. However, to realize our vision, we call on exploiting the trajectory detailed data more for inferring *structural* metadata and *functional* metadata. In particular, analyzing detailed trajectories has the potential to help in inferring directions, turn restrictions, stop signs, and traffic lights. All these can be inferred from the time that vehicles stop at each intersection. Moreover, each road segment can have more functional data, e.g., different weights based on whether a vehicle is going straight, taking a left turn, or a u-turn.

3.2 Direction 2: OD Matrix

Due to the difficulty of collecting and obtaining detailed trajectory data, several companies and authorities start to share an Origin Destination Matrix (OD Matrix) [4], which represents trip information on the form (*origin point*, *destination point*, *starting time*, *duration*). Figure 1(b) gives an example of an entry of such matrix, where only the source and destination points are known. Examples of such public datasets are available for Austin, USA [2], Guangdong, China [47, 50], and Porto, Portugal [32]. OD Matrices have been extensively exploited by the transportation community to understand city traffic. To realize our vision, we would need to exploit OD matrices for map metadata inference. One attempt for doing so mapped each trip entry in the OD matrix to a linear equation as the sum of edges taken for the trip, under the strict assumption that trips use shortest distance path [37]. Hence, the OD matrix has become a set of linear equations with large number of unknown edge weights. With some approximations and assumptions, the equations can be solved to infer edge weights. More techniques are required in this direction. For example, if some edges in between origin and destination already have known weights from Direction 1, that would significantly decrease the number of unknown weights, which would increase the accuracy of solving the equations.

3.3 Direction 3: Coarse-grained OD Matrix

Releasing detailed trajectories and OD matrices have serious privacy concerns, where personal location information and whereabouts can be identified [7]. As a result, a recent trend in releasing datasets focus more on coarse grained OD matrices on the form (*origin zone*, *destination zone*, *time zone*, *statistics*), where origin and destination points are represented as geographical zones rather than exact points, which cloaks the actual trip information as a means of privacy. Figure 1(c) gives an example of this. The time zone could be rush hour, morning, evening, or weekend. Statistics can include mean, max, min, or standard deviation between the origin and destination zones within the specified time zone. Geographical zones can be zip codes, based on traffic, or just by truncating the latitude and longitude coordinates of each GPS point. Examples of such datasets are released for many major US cities, e.g., New York City [26], and Seattle [35]. Uber has also released such data for 51 cities across 6 continents [40]. Due to the coarse granularity of such data, there is almost no work that have exploited it to enrich the map infrastructure. However, we believe that such data is still rich and can benefit our vision, especially if combined with other available data. In particular, though as mentioned in Section 2, open source routing engines like OSRM [30] and GraphHopper [14] lack accurate traffic data, they still have pretty accurate ranking. For example, for any two trips, if OSRM reported their duration as t_1 and t_2, where $t_1 > t_2$, then, even though t_1 and t_2 may not be accurate, it is highly likely that their respective ranking, i.e., t_1 takes



(a) Fine Grained Trajectories (b) OD Matrix (c) Coarse Grained OD Matrix

Figure 1: Trajectory Data

more time than t_2, is pretty accurate. Hence, for a pair of zones in a coarse-grained OD matrix, we can literally compute all possible paths between the two zones via an open-source routing engine to obtain a relative ranking of all paths. The ranking can be then used with the statistics of the OD matrix to fit each path on a curve. This will convert the coarse-grained OD matrix to be just an OD matrix, and then we can use it for metadata inference as in Direction 2.

3.4 Direction 4: Satellite Images

Satellite images, wherever available at any resolution, have been a great source for inferring the underlying road network topology [3, 6]. We would like to go beyond this and exploit satellite images for *structural* metadata. Examples of such metadata would include number of lanes, road width, and road quality. More sophisticated techniques can determine road directions through the color of lines between lanes, the availability of traffic lights or stop signs at each intersection, along with turn restrictions. A main challenge in analyzing satellite images is the cost of obtaining high quality images. Low quality images are more affordable and available, yet, it may lead to less accurate results.

3.5 Direction 5: User Incentives

A major part of OSM success is its wide set of 250+K annual users contributing to it [28]. Users are mostly volunteers who have similar incentives to contributors of Wikipedia and open-source systems. We envision developing a similar set of users to our portal over the years. One way to initially boost the number of contributors is to exploit the gamification concept [41], which is always used as a means for engaging more contributors. Image labeling is one prime successful example of gamification [42]. One attempt to do so is through geospatial data labeling [43], which shows the potential of exploiting gamification for metadata labeling. More techniques need to be exploited, where users can either directly contribute their observed *structural* metadata for road segments, or (parts) of their trajectories that can be used to infer *functional* metadata with any of the first three directions in this section.

3.6 Direction 6: Intra City Learning

All previous directions are geared towards harvesting various forms of available data for metadata inference. This will likely work for all major road segments, e.g., highways and freeways, where there are available data. Unfortunately, this is not the case for less popular road segments, e.g., residential and service roads, where there is no enough coverage in the available data. Hence, this direction aims to complement and take advantage of the inferred metadata through any of the five previous directions to infer the missing metadata for other segments within the same city. One way to do so is to convert each road segment with known metadata to a feature vector. Then, a ML model can be built to get the relation between the feature vector and the known metadata, and then used to infer the unknown metadata for other road segments. A main challenge would be forming the feature vector and considering all the factors that would impact the metadata. Examples of such features would include the length of the road segment, metadata of neighbor roads, distance from major roads, types of buildings on the road sides, and the number of intersections at both ends. Another challenge, which is also applicable to the next two directions, is that state-of-the-art ML models are not spatially-aware. This results in low accuracy when dealing with spatial data, where the spatial neighborhood information is of utmost importance and would have the greatest impact on the result. One way to tackle this is to spatially zone the city in a hierarchical way, where each zone will have its own model, only based on the data within the zone. Then, use such model to infer the metadata for the road segments within the zone.

3.7 Direction 7: Inter City Learning

This directions goes along the same lines of Direction 6, except that we aim to learn map metadata from one city to another rather than from one road segment to another within the same city. The rationale here is that, due to lack of available data, the large majority of cities worldwide have no data whatsoever. The goal of this direction is to exploit the possibility that cities that follow similar road structure would follow similar map metadata. Unlike the case of Direction 6 that targets residential and service roads, this direction actually targets major roads, e.g., highways and freeways, as they are more likely to keep their metadata structure across cities. Similar to Direction 6, a major challenge would be to identify the features that will be fed to a model learning process to get the relations between road structure and its metadata. These features would be different from the ones used in intra city learning, as it needs to focus more on the city structure as a whole rather than road structure. Examples of such features would include the city area size and shape, number of major roads, number of road exits, and downtown area(s). Combined with intra city learning, we envision the possibility of inferring full detailed metadata for a whole city, even if there is no available data for such city.

3.8 Direction 8: Error Learning

All the previous seven directions in this section aim to infer map metadata. Naturally, all of them would have various degrees of

accuracy, and none would be perfect. Hence, in this direction, we are not targeting inferring any new metadata. Instead, we aim to learn the error of the inferred metadata, and then use that error to offset the estimated values for *functional* metadata. This can be used in conjunction with any of the previous directions that estimate *functional* metadata to calibrate their results. For example, assume a road segment with a known weight w that was either entered manually or inferred from Direction 1, hence it has the highest possible accuracy and can be considered as ground truth. Then, assume that applying our efforts using one (or all) of the other six directions have estimated the weight for the same road segment to be w_e. This means that our efforts had an error offset δ=w-w_e, which could be positive or negative value. With this, and along the same lines of Direction 6, we can build a ML model that relates a road segment to its error offset. In particular, we can represent each road segment by a feature vector similar to Direction 6, and add to it the known error offset. We can then use the learned model for those road segments that do not have a ground truth. For any such road segment, whatever weight we will get for it, we add the learned offset δ for a more accurate weight.

3.9 Direction 9: Quality Assessment

Unlike OSM, we allow metadata inferred from ML models alongside the manual addition of metadata. However, this integration raises concerns about quality of the inferred data. We envision using a concept similar to probabilistic knowledge base construction systems, where the confidence probability generated by the ML model of its output is utilized. Metadata with a confidence above a threshold will go through without moderation, while the metadata with less confidence will either go for moderation or flagged as less confident information for the users.

REFERENCES

[1] J. Anderson, D. Sarkar, and L. Palen. Corporate Editors in the Evolving Landscape of OpenStreetMap. *ISPRS Intl. J. of Geo-Information*, 8(5):232, 2019.
[2] Austin Dataset. https://data.world/ride-austin/ride-austin-june-6-april-13.
[3] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. J. DeWitt. RoadTracer: Automatic Extraction of Road Networks From Aerial Images. In *CVPR*, 2018.
[4] E. Cascetta, D. Inaudi, and G. Marquis. Dynamic Estimators of Origin-Destination Matrices Using Traffic Counts. *Trans. Sci.*, 27(4):363–373, 1993.
[5] E. W. Chambers, B. T. Fasy, Y. Wang, and C. Wenk. Map-Matching Using Shortest Paths. *ACM TSAS*, 6(1):1–17, 2020.
[6] G. Cheng, Y. Wang, S. Xu, H. Wang, S. Xiang, and C. Pan. Automatic Road Detection and Centerline Extraction via Cascaded End-to-End Convolutional Neural Network. *IEEE Tran on Geosci. and Rem. Sensing*, 55(6):3322–3337, 2017.
[7] C.-Y. Chow and M. F. Mokbel. Trajectory Privacy in Location-based Services and Data Publication. *ACM SIGKDD Explorations*, 13(1):19–29, July 2011.
[8] C. Dickinson. Inside the Wikipedia of Maps Tensions Grow Over Corporate Influence. Bloomberg. https://www.bloomberg.com/news/articles/2021-02-19/openstreetmap-charts-a-controversial-new-direction.
[9] Facebook AI. Mapping roads through deep learning and weakly supervised training. https://ai.facebook.com/blog/mapping-roads-through-deep-learning-and-weakly-supervised-training/.
[10] Facebook Engineering. MaRS: How Facebook keeps maps current and accurate. https://engineering.fb.com/2019/09/30/ml-applications/mars/.
[11] Geo Awesomeness. What does the acquisition of HERE mean for Nokia, carmakers, TomTom, Google and the industry? https://geoawesomeness.com/what-does-the-acquisition-of-here-mean-for-nokia-carmakers-tomtom-google-and-the-industry/.
[12] GIS Lounge. Businesses Using Open Source GIS. https://www.gislounge.com/businesses-using-open-source-gis/.
[13] GPS World. TomTom-Tele Atlas Merger a Done Deal. https://www.gpsworld.com/consumer-oemnewstomtom-tele-atlas-merger-a-done-deal-2911/.
[14] GraphHopper. https://www.graphhopper.com/.
[15] G. Hu, J. Shao, F. Liu, Y. Wang, and H. T. Shen. IF-Matching: Towards Accurate Map-Matching with Information Fusion. In *ICDE*, 2017.
[16] K. T. Jacobs and S. W. Mitchell. OpenStreetMap Quality Assessment using Unsupervised Machine Learning Methods. *Tran. in GIS*, 24(5), 2020.
[17] L. Li, M. Zhang, W. Hua, and X. Zhou. Fast Query Decomposition for Batch Shortest Path Processing in Road Networks. In *ICDE*, 2020.
[18] Lyft Engineering. How Lyft Creates Hyper-Accurate Maps from Open-Source Maps and Real-Time Data. https://eng.lyft.com/how-lyft-creates-hyper-accurate-maps-from-open-source-maps-and-real-time-data-8dcf9abdd46a.
[19] Lyft Engineering. How Lyft discovered OpenStreetMap is the Freshest Map for Rideshare. https://eng.lyft.com/how-lyft-discovered-openstreetmap-is-the-freshest-map-for-rideshare-a7a41bf92ec.
[20] Machine learning. https://wiki.openstreetmap.org/wiki/Machine_learning.
[21] Money Control News. Uber may shun Google Maps for open source ones: Report. https://www.moneycontrol.com/news/business/uber-may-shun-google-maps-for-open-source-ones-report-2764111.html.
[22] J. Morrison. OpenStreetMap is Having a Moment: The Billion Dollar Dataset Next Door. Medium Artcile. https://joemorrison.medium.com/openstreetmap-is-having-a-moment-dcc7eef1bb01.
[23] M. Musleh and M. F. Mokbel. RASED: A scalable dashboard for monitoring road network updates in OSM. In *MDM*, 2022.
[24] Nextbillion.ai. OpenStreetMap for Businesses: A Primer. White Paper. https://nextbillion.ai/whitepapers/OpenStreetMap-for-Businesses-A-Primer.
[25] N. Nolde. Open Source Routing Engines And Algorithms - An Overview. In GIS Open Source Software. https://gis-ops.com/open-source-routing-engines-and-algorithms-an-overview/.
[26] Nyc taxi zones. https://data.cityofnewyork.us/Transportation/NYC-Taxi-Zones/d3c5-ddgc.
[27] OpenStreetMap. http://www.openstreetmap.org/.
[28] OSM Statistics. https://wiki.openstreetmap.org/wiki/Stats.
[29] Osm automated edits code of conduct. https://wiki.openstreetmap.org/wiki/Automated_Edits_code_of_conduct.
[30] Open Source Routing Machine (OSRM). http://project-osrm.org/.
[31] pNEUMA. https://open-traffic.epfl.ch/.
[32] Taxi Service Trajectory. Prediction Challenge. ECML PKDD 2015. https://www.kaggle.com/c/pkdd-15-predict-taxi-service-trajectory-i.
[33] J. Rapp, F. Barth, and S. Funke. Destination Signs in OpenStreetMap: Quality Assessment and Instrumentation for Routing. In *IWCTS*, 2019.
[34] H. Samet, J. Sankaranarayanan, and H. Alborzi. Scalable network distance browsing in spatial databases. In *SIGMOD*, 2008.
[35] Shared mobility aggregated trips. https://data.seattle.gov/Transportation/Shared-Mobility-Aggregated-Trips/uirh-29ta.
[36] SFMTA - Transit Vehicle Location History. https://data.sfgov.org/Transportation/SFMTA-Transit-Vehicle-Location-History-Current-Yea/x344-v6h6.
[37] R. Stanojevic, S. Abbar, and M. Mokbel. W-edge: Weighing the Edges of the Road Network. In *SIGSPATIAL*, 2018.
[38] Tesmanian. tesla and spacex news. tesla owners improve smart summon routes by updating open street maps. https://www.tesmanian.com/blogs/tesmanian-blog/tesla-owners-smart-summon-routes-open-street-maps-full-self-driving.
[39] Uber engineering. enhancing the quality of uber maps with metrics computation. https://eng.uber.com/maps-metrics-computation/.
[40] Data retrieved from uber movement. https://movement.uber.com/.
[41] L. von Ahn. Games with a Purpose. *IEEE Computer*, 39(6):92–94, 2006.
[42] L. von Ahn and L. Dabbish. Labeling Images with a Computer Game. In *ACM CHI*, 2004.
[43] V. Walter, M. Kölle, and D. Collmar. A Gamification Approach For The Improvement of Paid Crowd-Based Labelling of Geospatial Data. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 4, 2022.
[44] Y. Wang, G. Li, and N. Tang. Querying Shortest Paths on Time Dependent Road Networks. *PVLDB*, 12(11):1249–1261, 2019.
[45] L. Wu, X. Xiao, D. Deng, G. Cong, A. D. Zhu, and S. Zhou. Shortest Path and Distance Queries on Road Networks: An Experimental Evaluation. *PVLDB*, 5(5):406–417, 2012.
[46] Z. Xu, P. Badrinath, X. Huang, and A. Thomas. Grab-Posisi - Southeast Asia's First Comprehensive GPS Trajectory Dataset. https://engineering.grab.com/grab-posisi.
[47] Y. Yang, F. Zhang, and D. Zhang. SharedEdge: GPS-Free Fine-Grained Travel Time Estimation in State-Level Highway Systems. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):48:1–48:26, 2018.
[48] Z. Yu, X. Yu, N. Koudas, Y. Liu, Y. Li, Y. Chen, and D. Yang. Distributed Processing of k Shortest Path Queries over Dynamic Road Networks. In *SIGMOD*, 2020.
[49] J. Yuan, Y. Zheng, C. Zhang, W. Xie, X. Xie, G. Sun, and Y. Huang. T-drive: Driving Directions based on Taxi Trajectories. In *SIGSPATIAL*, 2010.
[50] D. Zhang. Description for ETC Data Release V0. https://www.cs.rutgers.edu/~dz220/Data/ETCData.rar.

Received 07 June 2023; revised 09 August 2023; accepted 08 September 2023