SONET: A Semantic Ontological Network Graph for Managing Points of Interest Data Hetereogeneity

Rachel Palumbo
palumborl@ornl.gov
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Laura Thompson
thompsonlk@ornl.gov
Oak Ridge National Laboratory
Oak Ridge, Tennessee

Gautam Thakur
thakurgt@ornl.gov
Oak Ridge National Laboratory
Oak Ridge, Tennessee

ABSTRACT

Scalability, standardization, and management are important issues when working with very large Volunteered Geographic Information (VGI). VGI is a rich and valuable source of Points of Interest (POI) information, but its inherent heterogeneity in content, structure, and scale across sources present major challenges for interlinking data sources for analysis. To be useful at scale, this information needs to be wrangled into a standardized schema. In this work, we tackle the problem of unifying POI categories (e.g. restaurants, temple, and hotel) across multiple data sources to aid in improving land use maps and population distribution estimation as well as to serve as a resource for data analysts wishing to fuse multiple data sources with the OpenStreetMap (OSM) mapping platform. Graph theory and its implementation through the SONET graph database, provides a programmatic way to organize, store, and retrieve standardized POI categories at multiple levels of abstraction. Additionally, it addresses category heterogeneity across data sources by standardizing and managing categories in a way that makes cross-domain analysis possible.

CCS CONCEPTS

• **Computer systems organization → Embedded systems**.

KEYWORDS

big data, openstreetmap, ontology, points of interest, graph database

ACM Reference Format:
Rachel Palumbo, Laura Thompson, and Gautam Thakur. 2019. SONET: A Semantic Ontological Network Graph for Managing Points of Interest Data Hetereogeneity. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Volunteered Geographic Information (VGI) offers researchers a wealth of accessible, geographic data; with that, it is especially valuable in regions of the world traditionally considered data poor. VGI is dominated by user-generated information such as Points of Interest (POIs), that provide insight into the type of facility or land use present and its social function at a specific location. Increasingly, in this era of widely available open source data, the linchpin of analysis is often a disconnection between the schema of disparate data sources. Although the conflation of POIs from multiple platforms would dramatically improve both spatial and categorical coverage, source heterogeneity makes it difficult to examine POI data from more than one or two platforms simultaneously. Maximizing the utility and efficiency of using this data requires addressing the data heterogeneity at the category level by curating, wrangling, and deploying this data in a standardized schema. Each VGI platform classifies POIs into a number of categories such as Zoo, Water Park, Restaurant, or Temple, using a specific, semantic schema, which informs the type or use of the facility or area. With this facility and area level information, land use of cities, regions, and countries can be estimated for areas of the world that traditionally lack land use maps. Because different platforms approach category generation from the top-down (i.e., Facebook and Google) or the bottom-up (i.e., OpenStreetMap (OSM) and Wikimapia), platforms contain between 21 to over 10,000 categories at varying levels of abstraction and consistency. Currently, there exists surprisingly little research exploring the conflation of POIs across different open-source mapping platforms [5]. Our work aims to fill this gap by addressing category heterogeneity across VGI sources. We achieve this through the development of a category-level ontological network that uses a graph database of nested categories to assign POIs from different sources, a category based on a consistent schema. This network manages categories in a way that makes cross-platform analysis possible and supports land use mapping and population modeling. Using the long-standing and popular OSM tag structure, each set of source categories are matched to one or more corresponding OSM category tags and stored in a scalable, extensible graph database. In the database, these category tags are clustered by land use and facility type, creating a hierarchy of land use types. We also connect source categories of original data sources and their OSM-matched tags with each other for querying purposes. Thus, land use types are mapped at the individual scale or at a larger, more generalized scales (e.g., Residential or Non-Residential). The resulting ontological network advances VGI-based research in two ways. First, it enables cross-platform analysis of POI data, thus maximizing both spatial, categorical, and social coverage. Second, the hierarchical network structure allows for the application of POI data in land use and population dynamics research.

2 RELATED WORK

The past several years has been host to a rise in the use of VGI in a variety of research applications, including work evaluating data quality on a variety of VGI platforms. To date, much research exploring POIs has been limited to using data from one or two

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Table 1: Total number of POIs and unique categories for each social media platform maintained in PlanetSense.

Platform	Total POIs	Unique Categories
Google	7,244,756	124
Facebook	41,478,490	1,634
OSM	5,607,516	520
Wikimapia	9,092,770	7,204
Vkontakte	2,007,348	69
Here	3,688,983	458
TomTom	5,251,749	382
Foursquare	9,447,304	938

datasets, or has focused on collecting data for a small area (i.e., a case study) [1–3]. Because different VGI platforms have been used more in some locations than others or to document certain types of information, there may be applications for which one dataset may perform better than another [3]. Though logistically and computationally rigorous, the ability to conflate multiple POI datasets can undoubtedly offer a more comprehensive understanding of geographic context than any source can offer alone. Some attempts to conflate POIs have been made recently, including a few approaches for matching individual POIs from two different platforms [4, 5]. A few tools exist to create a semantic network of tags in OSM [1, 3]. These tools identify relationships and provide structure to key:value pairs in OSM, contributing to our understanding of the semantic web and VGI. Because their intended purpose and scope is different, these tools are limited in their ability to inform land use and population dynamics research; networks do not assign key:value tags to a broader network of facility and land use categories used in these applications, which is a fundamental goal of our research.

3 DATA

We collected public POI data from 8 social media platforms for this work: Google, Facebook, OSM, Wikimapia, Vkontakte (VK), Here, TomTom, and Foursquare. All information was obtained through APIs and stored in the PlanetSense service [10]. Before the POIs can be processed for category matching and analysis, several issues in data variability must be addressed. Missing information such as the POI name or a category, poses an additional challenge to using the data. POIs with no category are collected and given a category that aligns with the schema from its data source of origin, through a machine learning model using Naïve Bayes and Support Vector Machine (SVM) models. This process is beyond the scope of this paper and only marginally applies to the overall semantic network. The data is also enriched with several additional attributes: translated name; reverse geocoded address, and country of origin.

4 POI CATEGORY MATCHING

Once the data has been collected from each platform and ingested into PlanetSense, unique categories from each data source are collected and matched to a set of appropriate OSM tags. OSM is one of the most widely-used and information rich sources of VGI [8]. Each POI within OSM has a set of tags made up of key-value pairs that provides categorical information about the land use or facility present. A maintained wiki provides recommendations for best-practice tagging and a list of commonly used tags, but users may assign and create tags freely. As a result of this open structure, OSM currently has 71,016 keys and 97,420,453 unique tags (key-value pairs) according to taginfo, a database of OSM tags maintained by the OpenStreetMap Foundation. Each POI in OSM is assigned multiple tags that describe various types of information about the POI, such as the building type and what services or activities are expected to occur there. We choose to map our source categories from Google, Facebook, Here etc. to this OSM format because of its broad use and rich source of category types. In keeping with the open tagging format of OSM (more than one tag is associated with each POI) POI matching may be a one-to-many match (**Hospital** matched to *amenity=hospital* and *building=hospital*), or a one-to-one (**City** is matched to only *place=city*). We chose to maintain this standard of consistency to maximize integration of the work with platforms that utilize and render in the OSM schema. There are three desired overall outcomes from this matching:

(1) Standardize category syntax across multiple data sources
(2) Consolidate similar categories within data sources
(3) Match similar categories across data sources

The process of matching source categories to OSM tags is achieved through human encoding. There are three primary criteria for assigning OSM tags to the source category.

(1) Assign at least one tag from the Map Features Wiki
(2) Assign one more specific tag from either Map Features of Tag Info Wiki
(3) Assign one *building = ** tag where appropriate

In some cases, there may not be tags of all three types that are appropriate for a POI, so the priority is to select at least one that is presented on the Map Features Wiki. Given the vast number of tags and the potential to assign multiple different tags to the same POI, we ask at least two questions when choosing an appropriate tag: 1) How is the space used; 2) When is the space used. The temporal aspect of this work is inspired by research indicating temporal differences in open/close times for POIs of certain categories [6] These two factors help classify functionally similar POIs into the same group.

5 MATERIALS AND METHODS

5.1 Building a Graph Database

To address the major challenge of organizing and accessing this large amount of data in a scalable way, several approaches were considered. First, we considered a dictionary, accessible through python or R programming languages. The size and complexity of the data makes this a cumbersome option. A second approach considered was a hierarchical tree data structure. The major drawback for this approach is structure limitations in order to maintain a balanced tree. Due to the nature of this data, there is not always data available to maintain a balanced tree. An crucial aspect of this data is the relationships between the data points; categories that are semantically similar will be more closely aligned than others but also have relationships to other, less similar categories at various levels of abstraction. Because of this, a graph database was chosen as the database structure to organize the POI categories and their matched

Table 2: Non-Residential Super-Categories and Sub-Categories

Retail and Service Public Services	Commercial	Institutions and
Food	Office Building	Religious
Store	Manufacturing	Education
Kiosk	Power Plant	Healthcare
Market	Chemical Refining	Public Service
Hospitality and Tourism	Learning Facilities	
Filling Stations		

OSM tags. Three Level0 categories, 13 Level1, and 44 Level2, and currently 11,329 source categories matched to 2,545 OSM Tags optimized for use in land use mapping and population modeling. As new data is collected, processed, and added to the database, the number of source categories and OSM tags may increase, but it is not expected to be substantial. In creating an ontology of POI categories for land use mapping and population modeling, land use and facility types were adapted from the LandScan project [2] to create the Level0, Level1, and Level2 nodes. At the top of the database structure (see Figure 1), three Level0 nodes, Residential, Non-Residential, and Administrative facility types provide an important base for land use mapping and population modeling. Under Residential, 3 Level2 categories, Single Family, Multi-Family, and Refugee/IDP Settlements, create an additional level of abstraction. The Administrative level contains 2 Level2 categories, National and Sub-National. This accommodates source categories that do not fit precisely into Residential or Non-Residential types, such as City, Country, and Country. Non-Residential takes the bulk of the data, being divided into 13 Level1 categories and 34 Level2 categories. Table 2 illustrates a few examples of the types of Level1 and Level2 categories in this division. Traversing the graph down from the sub-category level, is the Source Category level. This includes source categories from each of the VGI data sources we collect POIs from. Connected to each of these source categories is a set of OSM Tags that describe the type of facility or place it is. There may be some OSM Tags, such as *building=retail*, that are associated with a very large number of source categories; in this case, there are 2,916 source categories associated with that tag. In other cases, only a few source categories may be associated with any given OSM tag. The graph may be traversed to reveal clusters of similar source categories. With this structure, it is possible to easily generate data for source categories that belong to a certain Level0, Level1 or Level2 categories.

6 DISCUSSION

The primary motivation for this research is the construction of a graph database to organize, store, and retrieve standardized POI categories in support of land use and population dynamics research at multiple scales. Researchers have several options for accessing the data including a browser interface or connection through Python, R, or Java drivers. The database can be queried to produce both

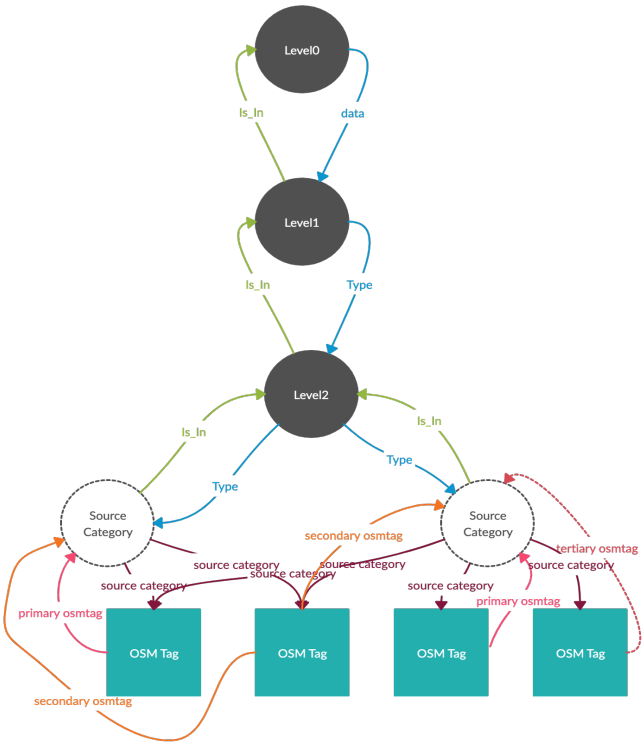

Figure 1: SONET database schema.

a visualization and a table of the desired output. Tables may be exported as csvs for use in other programs. Connections between a source category of interest such as Soccer Stadium, OSM tags that it has been matched to, and other source categories such as Football Stadium that share an OSM tag (i.e., *building=stadium*, can be collected for analysis across multiple data sources. It also has the potential to be used as an aid for adding appropriate tags to new POIs in OSM or OSM dependent platforms.

6.1 Land Use Mapping

Depending on project needs, land use mapping may be done at various scales, from general use classifications such as Residential and Non-Residential, to more specific land use types such as Retail, Commercial, and Military. The ability to approach land use at various scales using VGI data can be optimized with the category standardization, organization, and retrieval abilities afforded by this graph database. In many cases, there are multiple categories within one data source that have a sufficient level of similarity to unite under the same category. Facebook for example, has 96 categories indicating a restaurant with a specific cuisine such as Italian Restaurant, Pakistani Restaurant, Soul Food Restaurant, and Belgian Restaurant. In the absence of a graph database, creating a land use layer identifying types of retail stores would require the collection of all Restaurant categories. Searching for POIs tagged with many different categories is time-consuming and prone to errors of omission. However, the graph database stores all restaurant types, including other eating establishments such as Cafes or Food

Courts, as members of the *Food* Level2 category, which is connected to the larger *Retail/Service* Level1 category. This makes it easy to query the database for all categories (from one or all data sources) in *Food*, retrieve the source categories and their matched OSM tags, and extract all POIs with those source categories or OSM tags from a data storage service such as PlanetSense

Table 3: Sample output for a query extracting POI source, source category, and matched OSM Tags from the Food Sub-Category.

Root	Non-Residential	
Super-Category	Retail/Service	
Sub-Category	Food	
Source	**Source Category**	**OSM Tags**
OSM	bar	amenity=bar; building=retail
Facebook	Whisky Bar	amenity=bar; building=retail
Facebook	Bar	amenity=bar; building=retail
Facebook	Sports Bar	amenity=bar; building=retail
Google	bar	amenity=bar; building=retail
Wikimapia	bar	amenity=bar; building=retail
Facebook	Restaurant	amenity=restaurant; building=retail
Google	food	amenity=restaurant; building=retail
Google	restaurant	amenity=restaurant; building=retail
Wikimapia	eatery	amenity=restaurant; building=retail
Facebook	African Restaurant	amenity=restaurant; cuisine=african; building=retail

6.2 Population Modeling

Population models require very large amounts of spatiotemporal data. VGI sources are valuable sources of such information, especially in areas that are lacking in authoritative data. However, category differences across multiple sources increases the challenges for efficiently using this information to create high resolution models at, for example, the building level. Achieving an appropriate level of precision using VGI generated POIs requires a rectification of the model schema to data source schemas and data source schemas to each other. In most cases, manual matching of categories is required, significantly slowing down project efficiency. In this situation, the graph database can be used to assign standardized categories (i.e., OSM tags) to source categories from diverse sources. Querying for a specific category may be done using POI name and source category. Additionally, taking advantage of the semantic organization of the database, new categories can be quickly matched to existing, similar categories, providing a potential set of OSM tags in the proper format and adding them to the database. This will significantly increase the efficiency and consistency of preparing data for such models.

7 CONCLUSION

To the authorsâĂŹ knowledge this research represents one of the first attempts to create an ontological network to match categories across multiple heterogeneous sources of POI VGI data. To date, research utilizing POI data has been limited to using one or two sources of POI data, often at small spatial scales. Utilizing graph theoretical approaches will allow for the translation of POI categories from multiple sources into the popular and long-standing OSM tag format. Furthermore, the ontological network is constructed with a hierarchical framework, allowing for the retrieval of categories at varying levels of specificity. The resulting ontological network will advance the study of VGI data in two ways: first, by enabling cross-platform analysis of POI data, spatial and categorical coverage globally will improve. Cross-platform coverage will improve geographic understanding in regions of the world that are traditionally considered data poor. Second, this work supports the use of POI data in land use mapping and population modeling applications. In the future, we would like to expand and enrich the ontology graph to include more attributes and make it available to the open source community.

ACKNOWLEDGEMENT

This manuscript has been co-authored by UT-Battelle, LLC, under contract DE-AC05-00OR22725 with the US Department of Energy (DOE). The US government retains and the publisher, by accepting the article for publication, acknowledges that the US government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for US government purposes. DOE will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

REFERENCES

[1] Helen Dorn, Tobias Törnros, Alexander Zipf, Helen Dorn, Tobias Törnros, and Alexander Zipf. 2015. Quality Evaluation of VGI Using Authoritative DataâĂŤA Comparison with Land Use Data in Southern Germany. *ISPRS International Journal of Geo-Information* 4, 3 (sep 2015), 1657–1671. https://doi.org/10.3390/ijgi4031657

[2] Hartwig H Hochmair, Levente Juhász, and Sreten Cvetojevic. 2018. Data Quality of Points of Interest in Selected Mapping and Social Media Platforms. (2018). https://doi.org/10.1007/978-3-319-71470-7_15

[3] David Jonietz, Alexander Zipf, David Jonietz, and Alexander Zipf. 2016. Defining Fitness-for-Use for Crowdsourced Points of Interest (POI). *ISPRS International Journal of Geo-Information* 5, 9 (aug 2016), 149. https://doi.org/10.3390/ijgi5090149

[4] Grant McKenzie, Krzysztof Janowicz, and Benjamin Adams. 2013. Weighted Multi-attribute Matching of User-generated Points of Interest. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL'13)*. ACM, New York, NY, USA, 440–443. https://doi.org/10.1145/2525314.2525455

[5] Tessio Novack. 2018. Graph-Based Matching of Points-of-Interest from Collaborative Geo-Datasets. *ISPRS International Journal of Geo-Information* 7, 3 (2018), 117. https://doi.org/10.3390/ijgi7030117

[6] Kevin Sparks, Gautam Thakur, Amol Pasarkar, and Marie Urban. 2019. A global analysis of cities'geosocial temporal signatures for points of interest hours of operation. *International Journal of Geographical Information Science* 0, 0 (2019), 1–18. https://doi.org/10.1080/13658816.2019.1615069 arXiv:https://doi.org/10.1080/13658816.2019.1615069