Critical Analysis of *WorldKG: A World-scale Geographic Knowledge Graph*

Introduction

OpenStreetMap (OSM) is a public spatial database that stores a large amount of geographic data across the world. One of its most important features is that users are free to make their own contributions to the OSM. As long as official guidelines are followed, anyone can edit and update the content in OSM. It provides the OSM project a huge potential for covering every aspect of geodata as much and detailed as possible, but as a downside, content defined in OSM usually differs in many ways, and that becomes an essential problem for data organization and retrieval. To overcome this challenge, an integration of OSM and knowledge graphs has been created to standardize the way geodata is defined and categorized. This architecture, as called *WorldKG*, takes advantage of knowledge graphs to enrich contextual information of geodata so that the network of geographic entities can be reflected in a more advanced fashion.

Before WorldKG, there were already several other similar works of reorganizing spatial data in a format of knowledge graphs. LinkedGeoData also uses OSM data as the source for constructing a spatial knowledge graph. YAGO2Geo extends the coverage of the original YAGO to include more geographic information. However, the amount of spatial data they covered is still limited. There has been increasing studies on ontology building with spatial data for knowledge graphs as well, and some of them specially focused on OSM data. WorldKG was created given the background of all research and studies mentioned above.

Problem Statement

The problem with the OSM database is that the entities within it are only partially structured with no strict schema. Each entity or “corpus” is defined as a node N and a set of tags T, and each node is defined by an identifier i, location l, and another set of tags T_n which

consists of key-value pairs. The WorldKG ontology ignores the set of tags T making up a corpus and treats each entity as if it were just a node. While the identifier and location for each node is strictly structured, the set of tags for each node is regulated only loosely by a set of guidelines provided by OSM, and there are over 80,000 unique keys which can each have many distinct values associated with them. WorldKG imposes an ontological structure on the OSM data in order to make it more readily usable for applications and automatic data analysis. This is important in the context of our course because we are concerned with finding the best ways to represent geospatial data and to develop data-driven applications which can solve real-world problems with that geospatial data. In addition to the structural creation of OSM nodes, WorldKG also makes a connection to two widely recognized public knowledge graphs, WikiData and DBpedia. Every single geographic entity created in WorldKG has corresponding classes that are strictly aligned with the structure of WikiData and DBpedia.

Contributions

Specifically, the WorldKG paper mentions four key contributions. The first is WorldKG as a graph to better represent the OSM database. Second is the WorldKG ontology which does not necessarily need to be used in WorldKG, but could be used as a baseline to develop future knowledge graphs. Probably the most important contribution of the project is the SPARQL endpoint and interface for WorldKG which allows researchers to download the data in the graph. Finally, all of the source code for the project and pipeline is available on GitHub so that future work can be done with the project by interested researchers. The contributions listed in the paper are not significantly different from each other and could easily be summed up as the open source WorldKG tool. Outside of the listed contributions, the real value of this project is that it provides a queryable interface to the largely unstructured data within OSM so that it is more readily accessible to researchers.

WorldKG Ontology and Creation

In order to construct the WorldKG knowledge graph from the OSM data, we first need to understand what a knowledge graph is and how OSM works. We already explained the node structure of OSM and so will not spend further time on that. Knowledge Graphs are made up of sets of entities, classes (which are a proper subset of the entities), properties, literals, and relations. The data in knowledge graphs is typically represented as subject-predicate-object

tuples where the subject is an identifier for the entity, the predicate represents what information about the subject is stored in this tuple such as “label”, “instanceof”, or “coordinate”, and the object is the plaintext information about the entity or another entity in the relation. For example the tuple Q3375-”parentpeak”-Q15127 represents the relationship between two mountain peak entities where Q3375 is the id of the parent peak and Q15127 is a child peak. Using this structure we could intuitively query a large dataset to find all subjects with a certain label or find all tuples associated with a particular subject to display various parts of the graph.

Given this background information, the process for constructing the ontology and knowledge graph for WorldKG are explained. The first step is defining the classes and subclasses that can be used. OSM provides a map feature list whose items are used as the top-level classes in WorldKG. Any item listed as a value for a key in the map feature list will be considered a subclass for that key. An unsupervised machine learning algorithm was used to identify links between OSM tags and classes in other schemas such as Wikidata and DBpedia.

To create the graph, all the recent OSM data is collected from a data dump. All key-value tags with keys fitting the OSM ontology are kept and the rest are discarded. Finally, RDF triples are created using python code to fill in the rest of the relevant data.

Evaluation and Methodology

After the graph was created, there were 33 top-level classes and 1,143 subclasses for over 100 million entities now in WorldKG. These statistics alone provide some proof of the value of the project in creating a hierarchy to apply to the OSM data. To confirm accuracy, some stratified random sampling with manual verification is done. Five random classes with analogs to Wikidata and DBpedia entries each are chosen, and for those ten total classes 100 random samples from WorldKG are selected to evaluate for a total of 1000 test cases. After manually inspecting these cases the researchers found over 97% accuracy for all classes and attributed the small amount of errors to incorrect tagging in the OSM data. Aside from random sampling for accuracy check, an example scenario of restaurant recommendation is provided to illustrate the practicality of WorldKG. A SPARQL query is written to find the three closest restaurants to Brandenburger Gate in Berlin, Germany. The query successfully returned three restaurants with their distances specified. With the incorporation of a visualized OpenStreetMap map interface, the returned results can be shown on maps.

The sampling process to verify the quality is a solid method in checking the completeness and accuracy of the WorldKG structure. To make it more convincing, authors

could also have conducted hypothesis testing of whether the accuracy of class alignment or type assertion is past a certain percentage, and specified p-score or other indicators and revealed possible results for different confidence intervals. As for the illustration, more and different scenarios can be included to test the robustness of the structure.

Assumptions

The authors use OSM as the source and the startpoint of this entire project. It is well assumed that OSM is trustworthy for all the data it provided for it being a cauldron of user generated content though, and for the continuing development of WorldKG, it has to be assumed that OSM will be running and well supported for a long time. Such assumption can be extended to other frameworks such as RDF for knowledge graph data representations and SPARQL for knowledge graph data query. All of them have to be in good support for WorldKG to run healthily.

Limitations

Although WorldKG has a relatively large coverage on spatial data, in the first place it filtered the input of OSM data and used OSM nodes only, while in OSM there are different data types such as ways and relations. Also, the source data is not strictly reviewed, so errors in the tagging in OSM will naturally propagate into the WorldKG data. The utility works for assisting the operation of WorldKG, which so far include SPARQL endpoints on its official websites, are not well maintained. Last but not least, some sections in the paper are not clearly explained.

Revisions

For ways and relations in OSM, they have more complicated data structures, and it will be more difficult to process them. We hope in future the authors can develop a framework to handle complicated OSM data and include this layer of complexity in WorldKG. For the tagging error caused by the unstrict OSM structure, one measure to mitigate this error is the use of the map feature list published by OSM as the top level classes and the filtering of any entries that do not fit within the accepted feature types.

We also have some suggestions on phrasing and organizing in the paper. In the introduction when the authors mention other geospatial knowledge graphs (LinkedGeoData, YAGO2Geo), it would be helpful to also point out that WorldKG is more valuable than them because it has many millions more entries than any of them. In the part of the ontology creation, the converting process from OSM data to WorldKG entity is not explicitly revealed, but it seems most properties are listed the same as they are in OSM and represented as links to the WorldKG entity. The additional data in RDF triples is not explained in very much detail and to fully understand this process a reader would most likely need to look through the source code. One possible area to improve on in the ontology is in the naming of other properties derived from the OSM data.