Extracting Geo-Semantics About Cities From OpenStreetMap

Mónica Posada-Sánchez
Vienna University of Economics and Business, Vienna, Austria
monica.posada-sanchez@s.wu.ac.at

Stefan Bischof
Siemens AG Österreich, Vienna, Austria
bischof.stefan@siemens.com

Axel Polleres
Vienna University of Economics and Business, Vienna, Austria
axel.polleres@wu.ac.at

ABSTRACT

Access to high quality and updated data is crucial to assess and contextualize city state of affairs. The City Data Pipeline uses diverse Open Data sources to integrate statistical information about cities. The resulting incomplete dataset is not directly usable for data analysis. We exploit data from a geographic information system, namely OpenStreetMap, to obtain new indicators for cities with better coverage. We show that OpenStreetMap is a promising data source for statistical data about cities.

Keywords

Geographic Information System, Linked Data, Data Extraction

1. INTRODUCTION

Access to high quality and updated data is crucial to assess and contextualize city state of affairs. This information is essential for decision makers in cities as well as for the general public. Likewise, infrastructure providers can offer tailored solutions to cities based on such data. The CityDataPipeline (CDP) is a platform to provide integrated access to statistical city data with worldwide coverage with time and provenance context. All data is published in a structured way as Linked Data so that it can be easily accessed through SPARQL queries[1].

So far the CDP uses diverse Open Data sources to integrate statistical information about cities: DBpedia, Eurostat, UN data. The whole integrated dataset is missing many values, especially for smaller cities. Considering the years 2004–2012 the Eurostat dataset alone is missing 71% of the values. The UN dataset is missing 99% of the values considering the same years. The missing rate deteriorates when *combining* the datasets, due to some indicators and cities occurring in only one of the datasets [1]. Values are missing due to several reasons: (i) the indicator was not measured in one year for one city (for example when no air quality sensors are deployed in a city), (ii) computed indicators could not be derived because one of the necessary base indicator values was not available, or (iii) an indicator was only reported by the publisher in a specific time interval.

Bischof et al. [1] use different methods to predict the missing values based on the given ones. However the error of the missing value prediction is often still too high to be used for data analysis. While the *median* prediction error (measured by the normalized root mean square error) is 1.36%, some prediction models return obviously wrong values like negative numbers [1].

One way to improve the prediction error is to perform more sophisticated missing value imputation. The other obvious approach is to collect more data about the cities. The second approach is more promising if we can use a data source with a (nearly) complete coverage of the indicators for the cities in the CDP.

In this work we exploit data from a geographic information system, namely OpenStreetMap (OSM), to obtain new indicators for cities so to improve missing value prediction capability of CDP. OSM is a unique dataset which contains global geospatial information enriched with extensive alphanumeric tagged information. The analytical potential of this dataset is endless although its limitations regarding data quality, lack of standardization and uneven completeness should not be disregarded.

2. METHODOLOGY

To evaluate the contribution of the integration of the OSM dataset to enhance predicting capabilities of the CDP we propose a methodology to systematically extract new KPIs (key performance indicators) or impute missing values with extracted ones.

The process to extract city indicators from open OSM is defined by the following three steps:

1. gather city boundary geometries (see Section 3);
2. perform geospatial queries constrained by the city boundary and compute indicators (see Section 4); and
3. perform data analysis (see Section 5).

Figures 1 and 2 depict two alternative attempted workflows to integrate open geospatial data, specifically OSM, into the CDP. The first one uses the two OSM APIs Nominatim and OverpassQL to gather the city boundaries, while the

[1]http://citydata.wu.ac.at/

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

second one relies on DBpedia and GADM (Global database of administrative areas) for the same task.

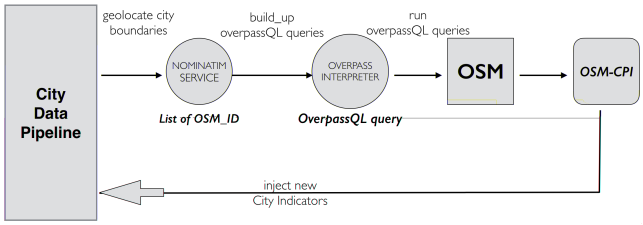

Figure 1: First OSM integration workflow using Nominatim and OverpassQL

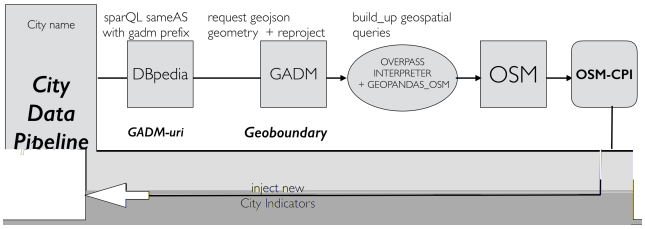

Figure 2: Second OSM integration workflow using DBpedia and GADM

As an alternative using intermediary tools we also evaluated PostGIS. However PostGIS i) did not solve the data quality issues and ii) is more restrictive in terms of data tag and temporal availability.

3. CITY BOUNDARY GEOMETRIES

As a first step to integrate OSM data into the CDP analytical platform, we need to identify the link between the CDP cities data and OSM. To achieve this goal we use the data sources and tools described in Tables 1 and 2.

Table 1: Data sources

	Description	Pros	Cons
CDP	City names and links to DBpedi.	–	–
OSM	Geospatial features enriched with alphanumeric tagged information.	Global dataset, geospatial, alphanumerically rich, open	Uncertainties: lack of formal validation, lack of standardisation, uneven completeness
DBpedia	Structured city information from Wikipedia	Global, open, very rich	Some inaccuracies: (sameAs tags, areas)
GADM	100m resolution Geojson city boundaries file access.	Global, open, geospatial	Only geometrical content

From the CDP we gather the list of city names including the DBpedia links. From this list we followed two different

Table 2: Tools

Tool	Description	Issues
SPARQL	Used to query over DBpedia city content	–
OverpassQL	OSM API used to query OSM data features (nodes, ways and relations)	Tagging is not standard and query definition is tricky, Overpass interpreter prevents query overloading from IP (automatisation issue)
Nominatim geolocator	OSM tool to search for geographic entities. Used to infer OSM_id of city boundaries	Everything named as the city would be returned, filtering of results is not trivial
PostGIS	Used to import city data exported from OSM	Data model differs from the full OSM, only keeps snapshot of OSM

approaches i) obtaining the OSM_id of the *relation* defining the boundary of the municipality by geolocating the city name through the Nominatim service (see Figure 1) and ii) using SPARQL to retrieve the `owl:sameAs` link to GADM geometry (see Figure 2). Once the city geo-boundaries are defined then we construct appropriate OverpassQL queries.

We encountered several hurdles on the city boundary collection process. These issues were mostly due to data quality weaknesses of the datasets and can be summarised as follows:

retrieving OSM_ID The Nominatim geolocation service returns a list of "locations" with those entities named as the city. A preliminary filter of this output list was attempted to isolate those entities corresponding to the boundary of the municipality, still further refinement is needed: OSM lack of standardisation among countries to label different administrative levels together with the uneven completeness intrinsic to the dataset poses a challenge here. In sum, smarter filtering techniques are needed to univocally distinguish between counties, regions, districts and municipalities named alike. In figures, from the geolocation through the Nominatim service of the 4070 cities in CDP, we could only uniquely identify 40% of the cities with our preliminary filter, 35% did not yield a valid OSM relation correspondence and around 2% was not even locatable.

retrieving GADM_ID It was common to find *DBpedia sameAs* links to GADM wrongly assigned to counties or regions named the same as the municipality.

common Due to city name string character encoding errors also prevented to identify the boundary of cities.

We are in the process of using Nominatim together with a simple filtering criterion to select the best city boundary. Furthermore this mapping will allow us to examine the quality of DBpedida-GADM links.

4. KPIS FROM GEOSPATIAL QUERIES

Once city boundaries are identified, we need to assemble relevant queries that will allow us to compute new city data

Table 3: Seven CDP cities OSM indicator results – new indicators

CDP id	#culture	#education	#greenery	#hotels	#libraries	#markets	#police	#schools	#sport_ctr	#theatre	#tourism
3254	10	116	474	33	11	0	8	69	39	5	122850
267	12	121	1093	30	22	0	2	92	51	5	510
3080	6	54	98	54	5	0	7	22	20	3	259869
2363	8	61	778	3	5	0	3	42	19	6	228509
3686	23	179	2660	46	11	0	10	132	58	17	2795557
282	48	284	1159	70	55	0	16	144	52	22	643546
762	32	278	3979	70	48	0	13	195	127	15	940703

Table 4: Seven CDP cities extracted links

CDP id	GADM URI	OSM URI	Wikidata	Wikipedia
3254	gadm_pfx/4/16523	osm_pfx/relation/346810	wdt_pfx/Q240262	wkp_pfx/da:Aarhus Kommune
267	gadm_pfx/1/784	osm_pfx/relation/1784663	–	wkp_pfx/es:La Coruña
3080	N.A.	osm_pfx/way/222220923	–	–
2363	gadm_pfx/1/1328	osm_pfx/relation/1390623	wdt_pfx/Q129610	–
3686	gadm_pfx/2/2682	osm_pfx/relation/897671	–	wkp_pfx/nl:Gent
282	gadm_pfx/2/2515	osm_pfx/relation/109163	–	–
762	gadm_pfx/2/11546	osm_pfx/relation/62518	wdt_pfxQ1040	wkp_pfx/de:Karlsruhe

gadm_pfx	http://gadm.geovocab.org/id
osm_pfx	http://www.openstreetmap.org
wdt_pfx	https://www.wikidata.org/wiki/
wkp_pfx	https://en.wikipedia.org/wiki/

Table 5: Seven CDP cities OSM indicator results

CDP id	name	OSM_id	area_sqkm	population
3254	A Coruña	346810	38.82	244810
267	Aarhus	1784663	477.52	256018
3080	Brasov	222220923	–	253200
2363	Galway	1390623	50.61	50800
3686	Ghent	897671	158.21	237000
282	Graz	109163	133.025	274000
762	Karlsruhe	62518	175.88	283959

indicators. In this preliminary work we focused on area calculations and on the extraction of tag information both for fulfilling already existing CDP indicators such as population and for generating new indicators counting relevant features within the area.

Regarding the area calculation, whenever available, city boundary coordinates are provided in WGC84, and so they need to be transformed to an appropriate projection to allow the calculation of areas in square meters (in this study we used global Mercator).

The proposed new CDP indicators can be classified in the following categories: education, cultural, leisure, tourism, security and green spaces:

education provides the number of entities corresponding to libraries, colleges, schools (children, languages, music, etc.), kindergartens and universities found within the limits of the city;

culture counts theatres, cinemas and art centres;

green spaces accounts for the number of forest, grass field, green fields, meadows, orchards, plant nursery facilities and village green areas;

security provides the number of police stations;

leisure provides the number of sport facilities; and

tourism provides the number of touristic entities which are not accommodation facilities. The number of accommodation facilities is also provided as separate indicator.

Regarding the assembling of OverpassQL queries, several implementation issues were encountered that need further work. Specifically, whenever a city boundary was conformed by a multi-polygon feature the resulting query will halt. Thus it would be necessary to loop over each of the contained polygons and perform the corresponding queries individually. Other issues of OverpassQL query execution relate to the fact that the Overpass interpreter does not allow concurrent requests from the same IP and resulting in aborted queries when launching chunks of consecutive queries.

As a preliminary validation of the proposed methodology we selected seven cities in CDP, extracted the city boundaries geometry (manually ensuring they were correct) and then gathering the target information. This was later injected into CDP workflow for its data analysis. The results of this study case are shown in Tables 3, 4, and 5.

5. DATA ANALYSIS

As a first step of data analysis we integrated the newly acquired geo indicators with the CDP dataset. Integration is implemented as a simple join over the CDP city URIs. This new integrated dataset can then be analysed and used for further processing.

Since we currently have new indicator values for only seven cities, running the missing value imputation or any machine learning algorithm makes no sense. However with a hosted version of the Overpass API we will get indicator values for at least hundreds of cities which makes missing value imputation feasible again.

Instead we performed a preliminary exploratory data analysis. To find out how each of the new indicators relate to the CDP indicators we computed a correlation matrix of all the indicators. As an example Figure 3 shows the correlation distribution of the *#libraries* indicator. Most other new

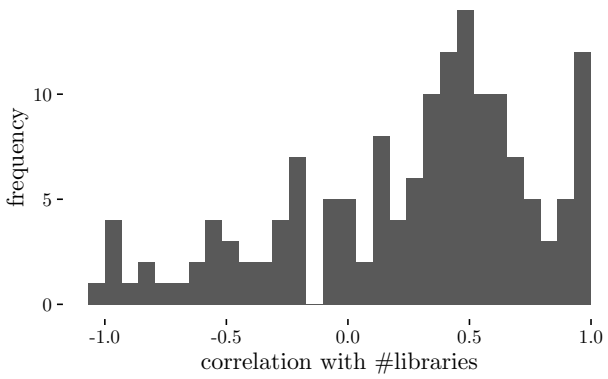

Figure 3: Frequencies of the correlation coefficient of the #libraries indicator with all other CDP indicators showing how the libraries indicator relates to all other indicators

indicators show a comparable correlation distribution. The figure shows a large number of indicators with a moderate correlation with *#libraries*. This could mean that the new indicators relate to some extent to the CDP indicators but still add new information not already present in the dataset.[2]

The results of the first manually matched cities are promising. We expect more reliable and conclusive results once we could generate indicator values for more cities.

6. RELATED WORK

Janowicz et al. [2] give an overview of how to use Semantic Web languages and technologies in the domain of geospatial databases. In this extensive survey paper, which is aimed at readers with a Semantic Web background, they present research questions, contributions and related literature. GeoKnow [3] is an implementation of a process to provide geospatial data as Linked Data. Although GeoKnow could be used to extract geo indicators similary to our approach, we aimed for a more lightweight solution with less overhead compared to the full blown linked data lifecyclye with a plethora of different tools. GeoSPARQL [4] extends SPARQL to a geographic query language for RDF data. This language could also be used to compute indicators similarly the way we do, but gives us no immediate advantage over using the OSM APIs directly. Linked Geo Data [5] also extracts information from OSM and publishes 20 billion triples as linked data. However the indicators are already predefined and thus lacks the flexibility we need for future data analysis. Geonames (http://www.geonames.org/) is a geographic database giving information on millions of spatial entities and might be another interesting data source for the CDP.

7. CONCLUSIONS AND FUTURE WORK

We showed a pipeline extracting structured city indicators from GIS data. The OSM dataset provides worldwide coverage of all sort of geospatial features enriched with alphanumerical and temporal information. These characteristics endow OSM with a great potential for data analysis. In fact, endless new city indicators could be retrieved or analytically inferred from this dataset. As an example, more advanced geospatial analysis of transportation networks were attempted with promising results to further investigate. Another seed for future work is the potential of the OSM data for geospatial and temporal link discovery processes among disparate datasets.

Once we gather more high quality data the results of this project could be used to improve public datasets. For example we could improve the `sameAs` links from DBpedia to GADM or evaluate the quality of numeric indicators such as population and area.

A full evaluation of the extracted indicators with respect to their capabilities of improving the missing value prediction is currently missing. For this evaluation we will compare the error estimates of the current missing value prediction with the error estimates obtained with the extracted indicators. Alternatively comparing the previous and newly predicted values directly for significant changes might give interesting insights as well. Before that we compare the overlapping indicators (those indicators which are already present in the dataset *and* extracted) and thus measure quality for those extracted indicators.

We showed that the method can provide useful data. However, on an operational level, several issues of different nature (data quality, tool usability and reliability, etc.) should be solved before being able to perform any analysis at a global scale in a fully automated manner. About data quality, challenges such as how to overcome i) the lack of standardisation of OSM tagging, ii) lack of data validation and iii) uneven completeness of the dataset remain open for further work. On the tool usability and reliability, issues such as i) how to optimise large geospatial data volume and costly processing, and ii) how to enhance responsiveness and reliability of tools and centralised services so to allow large scale usage, remain still open for further investigation.

[2]A log of the complete data analysis is available at https://github.com/stefanbischof/dds-cdp/blob/master/exploratory_data_analysis/dds.md.

8. REFERENCES

[1] S. Bischof, C. Martin, A. Polleres, and P. Schneider. Collecting, integrating, enriching and republishing open city data as linked data. In *Proceedings of the 14th International Semantic Web Conference (ISWC'15), Part II*, pages 57–75. Springer, 2015.

[2] K. Janowicz, S. Scheider, and B. Adams. A geo-semantics flyby. In *Proceedings of the 9th International Reasoning Web Summer School 2013*, pages 230–250. Springer, 2013.

[3] J. Lehmann, S. Athanasiou, A. Both, A. G. Rojas, G. Giannopoulos, D. Hladky, J. J. Le Grange, A.-C. Ngonga Ngomo, M. A. Sherif, C. Stadler, M. Wauer, P. Westphal, and V. Zaslawski. Managing geospatial linked data in the GeoKnow project. *The Semantic Web in Earth and Space Science. Current Status and Future Directions*, 20:51–78, 2015.

[4] M. Perry and J. Herring. OGC GeoSPARQL – a geographic query language for RDF data. OGC implementation standard, Open Geospatial Consortium, Sept. 2012.

[5] C. Stadler, J. Lehmann, K. Höffner, and S. Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web Journal*, 3(4):333–354, 2012.