Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge

Nikolaos Karalis, Georgios Mandilaras, and Manolis Koubarakis

Dept. of Informatics and Telecommunications, National and Kapodistrian University of Athens, Greece {nkaralis, gmandi, koubarak}@di.uoa.gr

Abstract. We extend YAGO2 with geospatial information represented by geometries (e.g., lines, polygons, multipolygons, etc.) encoded by Open Geospatial Consortium standards. The new geospatial information comes from official sources such as the administrative divisions of countries but also from volunteered open data of OpenStreetMap. The resulting knowledge graph is currently the richest in terms of geospatial information publicly available, open source, knowledge graph.

Keywords: knowledge graphs · YAGO · geospatial knowledge

1 Introduction

Many intelligent applications are driven today by knowledge graphs (KGs) such as the Google KG[1], DBpedia[2] and YAGO [9]. The first version of YAGO was released in 2007 [16,17]. YAGO was created by combining knowledge from WordNet [13] and Wikipedia, and it is one of the first open and free knowledge graphs. The entities of YAGO were created from pages of Wikipedia, whereas WordNet was used to create its classes and their hierarchy. YAGO knowledge is encoded in triples SPO where S is the subject, P is the predicate and O is the object.

YAGO2 [6,7], the second version of YAGO, was released in 2011. YAGO2 introduces geospatial and temporal information to the YAGO knowledge graph by introducing *geoentities*. Geopatial information in YAGO2 comes not only from Wikipedia but also from GeoNames[3]. GeoNames is a gazetteer[4], whose data and accuracy have been studied in [1,2,5].

The geospatial information in YAGO2 is represented with the properties `hasLongitude` and `hasLatitude` which give the longitude and latitude of the center of a geoentity. In YAGO2, the coordinates of Greece are represented with the following triples: `<Greece> <hasLatitude> "39.00"^^<degrees>` and `<Greece> <hasLongitude> "22.00"^^<degrees>`.

[1] https://developers.google.com/knowledge-graph/

[2] https://wiki.dbpedia.org/develop/getting-started

[3] https://www.geonames.org/

[4] A gazetteer is a geographical dictionary that is used, in most cases, together with a map. Given a name (i.e., a city or a river) a gazetteer gives geospatial information about that name.

Temporal information is introduced in YAGO2 to entities of type `people`, `groups`, `artifacts` or `events`. Temporal information is represented using dates. Dates in YAGO2 follow the ISO 8601 format (YYYY-MM-DD) and represent time points. If we want to model intervals e.g., the lifetime of an entity such as a person, we can use pairs of properties e.g., `wasBornOnDate` and `diedOnDate` which connect an entity with a date.

To represent geospatial and temporal knowledge, YAGO2 uses the *SPOTL* data model, which extends the SPO model for knowledge graph triples discussed above: T stands for time, L stands for location, and S, P and O as defined above. The SPOTL model not only allows temporal and geospatial relations between entities, but also temporal and geospatial relations between facts. For example, the fact that Barack Obama was inaugurated as president of the USA can be associated with a place (Washington D.C.) and a date (2009-01-20).

YAGO3 [12], the latest version of YAGO, came out in 2015. YAGO3 is multilingual since it combines information from Wikipedias in multiple languages.

The main technical contributions of this paper are the following.

We develop a new version of YAGO2, called YAGO2geo, with more precise geospatial information. YAGO2geo contains *640 thousand polygons* and *137 thousand lines*. The line and polygon information introduced in YAGO2geo makes, in many cases, more sense than the coordinate pairs that exist in YAGO2. For example, we do not need to model any more the longitude/latitude center of a stream or another geoentity for which it is not clear what the center is. Also, YAGO2geo can be used to answer questions for which precise geospatial information is required. This has not been possible with YAGO2. For example, such questions are "what is the city of Germany where two streams meet at a lake", or "which are the neighboring municipalities of the municipality of Athens?".

The extension, in combination with the 12 million coordinate pairs of YAGO2, creates a geospatial KG much richer, in terms of geospatial knowledge, compared to DBpedia which contains 1 million coordinate pairs and Wikidata which contains almost 2 million coordinate pairs and only 2 thousand shapes. This makes YAGO2geo the richest, in terms of geospatial information, publicly available, open source, knowledge graph.

We draw the new geospatial information from two sources. First, we utilize administrative data taken from official datasets of three countries: the Greek Administrative Geography (GAG) dataset, the administrative divisions dataset for the United Kingdom obtained from Ordnance Survey (OS)[5] and Ordnance Survey Northern Ireland (OSNI)[6], and the administrative division datasets of the Republic of Ireland obtained from Ordnance Survey Ireland (OSI)[7]. To obtain the geometries of administrative divisions of countries of the whole world, we also utilized the latest (2018) version of the Global Administrative Areas dataset (GADM)[8]. We also introduce to YAGO2geo geospatial information from the

[5] https://www.ordnancesurvey.co.uk/
[6] https://www.nidirect.gov.uk/campaigns/ordnance-survey-of-northern-ireland
[7] https://www.osi.ie/
[8] https://gadm.org/

biggest volunteered, crowdsourced and open dataset with geospatial information, OpenStreetMap (OSM)[9].

While introducing more precise geospatial information to YAGO2, we follow the following methodology. If the geoentity we enrich is already in YAGO2, we augment its geospatial information by defining its geometry more precisely (e.g., by a multipolygon for a city which we take from GADM, as opposed to a latitude/longitude pair that exists in YAGO2). We also keep the existing information (e.g., the old coordinate pair that gave the center of the city). Interested practitioners can use our methodology to enrich YAGO2geo with even more geospatial information (e.g., administrative divisions of their own country from official datasets, the European land cover and land use dataset CORINE[10] etc.).

We make a detailed comparison of the geospatial information available from YAGO2 and the geospatial information in OSM and the administrative datasets GAG, OS, OSNI, OSI and GADM.

We make YAGO2geo available publicly at `http://yago2geo.di.uoa.gr`. The free and open dataset there includes the extended KG encoded in RDF. The geospatial information follows the standards of the Open Geospatial Consortium, hence YAGO2geo can be queried using GeoSPARQL.

The rest of this paper is structured as follows. Section 2 discusses related works. Section 3 gives detailed information about the data sources that were used in order to extend YAGO2 with geospatial information. Sections 4 and 5 present the methodology that we followed and demonstrate the knowledge in YAGO2geo with examples. Last, in Section 6 we summarize our contributions, present our conclusions and discuss future work.

2 Related Work

In this section we discuss in some detail which of the existing well-known KGs contain geospatial and temporal knowledge. In GIS terminology which we often follow in this paper, a *geographic feature* (or simply feature) is an abstraction of a real world phenomenon and can have various attributes that describe its *thematic* and *spatial* characteristics. For example, the country Greece is a feature, its name and population are thematic attributes, while its location on Earth, in terms of polar coordinates, is a spatial attribute. Knowledge about the spatial attributes of a feature can be *quantitative* or *qualitative*. Quantitative geographic knowledge is usually represented using *geometries* (e.g., points, lines and polygons on the Cartesian plane) while qualitative geographic knowledge is captured by *qualitative binary relations* between the geometries of features (e.g., Greece is south of Bulgaria).

DBpedia, like YAGO2, contains latitude and longitude pairs for the center of cities, towns etc. extracted from Wikipedia. There are 1 million coordinate pairs available in DBpedia. In addition, DBpedia contains knowledge about some thematic attributes that can be used to infer knowledge about spatial attributes of

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

[10] https://land.copernicus.eu/pan-european/corine-land-cover

features. For example, for each country, the neighboring countries are given, or for each city, the country to which the city belongs is given. In this way, one can infer knowledge about the corresponding geospatial attributes of features e.g., "the geometry of Greece externally connects with the geometry of Bulgaria" using the vocabulary of Region Connection Calculus RCC-8 [14]. Recently, DBpedia has been attempting to add cardinal direction knowledge (e.g., Athens is north of Crete) via properties `dbp:north`, etc.

Grütter et al. in [5] carried out an extensive evaluation of topological relations found in DBpedia and GeoNames about the administrative divisions of Switzerland and Scotland. The authors present two different approaches for the evaluation of the topological relations: the *single dataset approach* and the *interlinked datasets approach*. In the first case, the topological relations of DBpedia are evaluated. In the second case, the topological relations of GeoNames are evaluated, which can be obtained from the `owl:sameAs` links that exist between the entities of DBpedia and GeoNames. The results of their work show that the values of recall and precision are relatively high when DBpedia is queried via GeoNames (i.e., in the second approach) and the links between these two sources are replaced by manually created links, that the authors created based on their expertise on Swiss and Scottish administrative divisions. In the case of Scotland, these values are really low when only the information of DBpedia is used or DBpedia is queried via the original links of GeoNames.

Wikidata [18], is an open and free knowledge graph and the successor of Freebase [3]. It is an activity of the Wikimedia foundation and it is used to serve many other projects of Wikimedia. Wikidata is developed collaboratively by members of its community. The users of Wikidata are able to add new knowledge to the underlying graph but also modify its schema. Wikidata is a multilingual knowledge base, and unlike DBpedia which has different versions for every language, the information of the entities of Wikidata is translated to multiple languages and is part of the same graph. When it comes to quantitative geospatial information, Wikidata provides two data types: `Globe Coordinate` and `Geographic Shape`. The coordinates of an entity can be obtained using the property `coordinate location` for that entity. There are currently over 7 million triples that contain this property (i.e., over 7 million entities for which Wikidata knows their coordinates). The data type `Geographic Shape` has the property `geoshape` which can be used to associate a knowledge graph entity (e.g., the entity for Athens) with a geometry. Geometries in Wikidata are encoded using the GeoJSON format. Currently, Wikidata contains only 2000 geometries which are mostly polygons and multipolygons. Apart from quantitative geospatial information, Wikidata also contains rich topological information, that is represented with various properties, such as `shares border with` and `country`.

Similarly to YAGO2, both DBpedia and Wikidata provide temporal information in the form of dates. One key difference is the fact that YAGO2 has a specific schema for the representation of temporal knowledge, whereas in DBpedia and especially in Wikidata there is a plethora of properties that are used in temporal facts. On the one hand that makes YAGO2 easier to query and to

	DBpedia	Wikidata	YAGO2	YAGO2geo
Coordinates	1M	7.2M	12M	12M
Lines and Polygons (Shapes)	-	2K	-	137K Linestrings and 640K Polygons and Multipolygons
Date of Birth	1.7M	3.5M	1.6M	1.6M
Date of Death	721K	1.7M	797K	797K

Table 1. Geospatial and temporal Information in current knowledge graphs.

comprehend, but on the other hand Wikidata provides larger amount of temporal information. Moreover, the dates in YAGO2 follow a specific pattern, which is not the case in DBpedia and in Wikidata. Last, time intervals in DBpedia and Wikidata can be represented just like in YAGO2.

Table 1 summarizes the geospatial and temporal information that is currently available in YAGO2, DBpedia and Wikidata. In order to compare the quantity of temporal knowledge, we show the number of birth and death facts that appear in each knowledge base, because they are the most common date facts. We can observe that YAGO2 contains the most coordinate pairs, because of the facts that come from GeoNames. Wikidata also contains a significant amount of geographic points and is the only knowledge base that contains geographic shapes. In addition, it provides more temporal information than DBpedia and YAGO2. Table 1 also stresses out the importance of YAGO2geo, since detailed geographic information (i.e., lines and polygons) is currently very limited.

3 Data Sources

YAGO2geo is built from YAGO2 and new geospatial knowledge from multiple sources. First, we use geographical administrative data provided by official sources of Greece, the United Kingdom and the Republic of Ireland. We also extract geospatial information about the administrative units of every country from the GADM dataset as well as for other types of features, such as lakes, from OpenStreetMap. Apart from the geometries, each data source provides additional information (e.g., population for cities) that we include in YAGO2geo.

The geospatial information about the administrative divisions of Greece that we introduce in YAGO2geo comes from official sources of the Kallikratis law which defines the administrative divisions of Greece in 2011. The administrative divisions of Greece, according to Kallikratis, consist of *decentralized administrations*, *regions*, *regional units*, *municipalities*, *municipal units* and *municipal communities*. The Kallikratis administrative divisions have been defined as linked data and called *Greek Administrative Geography* (GAG) by our group in the past and has been publicly available[11].

Ordnance Survey is the national mapping agency of the United Kingdom. It provides data about the countries of England, Scotland and Wales that form

[11] http://linkedopendata.gr/dataset/greek-administrative-geography

Great Britain. For our purposes we used the Boundary-Line dataset[12], which contains the administrative boundaries of Great Britain. More specifically, we used the information about the following administative divisions: *European regions*, *counties*, *districts and metropolitan districts*, *unitary authorities*, *boroughs*, *wards*, *parishes*, and *communities*.

Ordnance Survey Northern Ireland is the official cartographic agency of Northern Ireland. Users are able to obtain its data using the ONSI Open Data portal[13]. In this work we use the datasets *NI Outline*, *Local Government Districts 2012*, *Wards 2012* and *Townlands*.

The Ordnance Survey Ireland is the national mapping agency of the Republic of Ireland and it provides multiple products and datasets. The authors of [4] transformed the geospatial data about the boundaries of the administrative areas of Ireland into RDF. For the extension of the geospatial information of entities that belong to the Republic of Ireland, we consider the datasets (i.e., administrative areas) *city and county council*, *county council*, *city council*, *municipal district*, *barony*, *parish*, *townland* and *rural area*.

GADM provides geographic data about the administrative divisions of every country in the world. Administrative units are divided into six different layers (i.e., administrative levels `level-0` to `level-5`) and there are over 386,000 administrative areas in total. GADM does not only provide the boundaries of every administrative area, but it also provides additional useful information about them (e.g., administrative division and the upper administrative units). Version 3.6 of GADM was released in May 2018 and for our purposes we transformed the provided shapefiles into RDF using our tool GeoTriples [10]. GADM is a very useful dataset but its web site reveals little about it. For example, which group of people have constructed it, where did they find their data for various countries (e.g., Greece), etc. To the best of our knowledge there are also no studies that evaluate the quality of GADM. However, our experience with this and a previous version of the dataset since 2012 tells us that GADM has very good quality geospatial information (see also Section 4.4).

OpenStreetMap is a volunteer project, whose goal is to provide free geographic data and maps to its users. OSM provides geospatial information about multiple features. Such features are natural features (e.g., beaches, lakes, etc.), land use features (e.g., vineyards, etc.), places (e.g., villages, cities, etc.), points of interest, water bodies, waterways and more. We obtained OSM data from Geofabrik[14], which is a company that provides free, regularly-updated extracts of OSM. Geofabrik provides compressed OSM files (osm.pbf) and free shapefiles. After examining both types of files we came to the realization that there are some classes that are not included in the free shapefiles (e.g., airports)[15]. In addition, the OSM files provide every available name of each entity, which is very important in our work. Last, the free shapefiles do not follow the *key-value*

[12] https://www.ordnancesurvey.co.uk/business-and-government/products/boundaryline.html
[13] http://osni-spatial-ni.opendata.arcgis.com/
[14] https://www.geofabrik.de/
[15] http://www.geofabrik.de/data/geofabrik-osm-gis-standard-0.7.pdf (Section 8.3)

schema of OSM. For these reasons, we obtained the necessary information from the compressed OSM files using the tool TripleGeo[16] extended with a plug-in implemented by the second author of this paper.

Note that we have *not* used the OSM data provided by the LinkedGeoData [15] project which was the first attempt to make OSM data available on the web as linked data. The OSM data currently on the LinkedGeoData web site are *not* the most recent and they are not maintained actively, to the best of our knowledge. Thus, Geofabrik was the best portal available to obtain OSM data.

4 The Knowledge Graph YAGO2geo

The main goal of this work is to extend the YAGO2 knowledge graph with detailed geospatial information without duplicating existing knowledge. To ensure that, we try to match geoentities of YAGO2 with entities of the data sources that we have presented in Section 3. For example, the resource `geoentity_Hellenic_Republic_390903` and the entity with identifier `GRC` represent Greece in YAGO2 and GADM respectively. Therefore they should be declared to be identical using an `owl:sameAs` triple. The matching phase for identifying identical entities consists of applying two filters: (i) the label similarity filter and (ii) the geometry distance filter. Our methodology is based on the methodology that was used in YAGO2 when integrating information from GeoNames [7]. A similar approach has been used in LinkedGeoData [15].

The first filter of the matching phase is the *label similarity filter*. It produces matches between the geoentities of YAGO2 and the entities of the specified data source (e.g., GADM) that have similar names. For this purpose we experimented with the Levenshtein distance [11] and also the Jaro-Winkler similarity [8] and found out that, for our task, the latter produces more matches while maintaining high precision. In order for two resources to be matched, the similarity between their labels must be higher than a specific threshold, which we have set at 0.82. We examine every label of each entity, without considering its language tag like in [15]. Here, an entity of YAGO2 can be matched with multiple entities.

After the label similarity fitler is completed, we apply the geometry distance filter. The *geometry distance filter* is applied on the matches that were produced by the first filter and its goal is to eliminate any false matches. Since there are many geographic entities that share the same name (e.g., Athens, Greece and Athens, Alabama), the geometry distance filter is also a disambiguation step. The geometry distance filter checks if the Euclidean distance in the WGS:84 coordinate system[17] between the geometry provided by GADM, OSM, or an official country data source and the point provided by YAGO2 is smaller than a specific threshold, which is set at 0.2 degrees. In case there are multiple entities

[16] https://github.com/SLIPO-EU/TripleGeo

[17] A *coordinate reference system* (CRS) is a coordinate system that is related to an object (e.g., the Earth, a planar projection of the Earth) through a so-called datum which specifies its origin, scale, and orientation. WGS84 is the latest version of the *World Geodetic System* (WGS) and was established in 1984.

GAG	YAGO2	# Matches	Precision
decentralized administration	administrative_division	6/7	1.000
region	first-order	11/13	1.000
regional unit	administrative_division	21/74	1.000
municipality	third-order	325/325	1.000
municipal unit and community	populated_place and locality	530/1037	0.907

Table 2. Greece: results of the matching phase

of YAGO2 that are matched with the same resource, we keep the entity that is closest, in terms of distance, to that resource.

The number of matches produced by the two filters presented above is typically very large and, consequently, it is not possible to manually check if every match is correct. As a solution to this problem, we randomly selected a subset of the matches[18] and manually check if these matches are correct, by checking the label of the matched resources. This methodology has also been used in [7,15].

Let us now apply the matching methodology we just discussed to the problem of matching geoentities from YAGO2 and entities from each data source.

4.1 Greece: GAG dataset

In this section, we try to find matches between the entities of GAG and the geoentities of YAGO2. To achieve this, we carry out the matching phase on pairs of official administrative divisions (Section 3) and classes of YAGO2, as shown in Table 2. More specifically, the decentralized administrations and few regional units are instances of the class `geoclass_administrative_division`. The regions of Greece are matched with the geoentities of YAGO2 that are instances of the class `geoclass_first-order_administrative_division`. The second administrative level of YAGO2 does not appear in the results because it contains *prefectures*. Prefectures are no longer an administrative division in Greece and in the Kallikratis law they are replaced by regional units. Greek municipalities are found in the third administrative level of YAGO2. Last, since municipal units and communities are not found in the administrative levels of YAGO2, we try to match them with `populated places` and `localities`.

Table 2 summarizes our results. The number of administrative units that are found in each division of GAG is shown on the third column. The third column also presents the number of matches we were able to generate, whereas the fourth column shows the quality of the generated matches (i.e., the recall and precision of the matching phase). For the case of Greece we evaluated every match manually. The results show that our methodology was able to match perfectly the majority of the decentralized administrations and regions, that passed the label similarity filter, and all municipalities of GAG. The class `geoclass_administrative_division` of YAGO2 contains 21 Greek regional units and we

[18] For each matching phase we evaluate $max\{300, \#matches * 0.01\}$ matches.

YAGO2, UK (# Entities)		**OS and OSNI**	**# Matches**	**Precision**
first-order	(4)	euro. region	2	1,000
second-order	(185)	counties, unitary auth. metr. districts and boroughs	182	0,953
third-order	(3852)	unitary auth., districts boroughs, wards and parishes	3718	0,933
fourth-order	(7717)	wards, parishes and communities	7642	0,913
populated_place and locality	(15719)	wards, parishes, communties and townlands	1272	0,897

Table 3. United Kingdom: results of the matching phase

matched all of them. Here, we have to mention that we found some regional units in the second level of YAGO2, which however are labeled as prefectures. For this reason, we were not able to match them. Regarding the municipal units and communities, even though the precision is not high, it is very satisfying. As we have explained already, not only we extended matched geoentities of YAGO2, but we also include in YAGO2geo all unmatched entities of GAG.

4.2 United Kingdom: OS and OSNI datasets

The data that is provided by OS and OSNI is used to enrich the geoentities of YAGO2 that belong to the UK with official geospatial information. The countries of the UK are instances of the class `geoclass_first-order_administrative_division` of YAGO2. Counties, metropolitan districts, unitary authorities and the Greater London authority are found in the second administrative level of YAGO2. The third level of YAGO2 has entities that are communities, civil parishes, districts, London boroughs, metropolitan district wards or unitary authority wards. Communities and civil parishes are also found in the fourth administrative level of YAGO2, which also contains district wards.

The results (Table 3) show that we were able to match most of the geoentities of the UK that are found in the administrative levels of YAGO2. In order to match more entities of OS and OSNI, we carried out the matching phase using the class `geoclass_populated_ place` of YAGO2. We can also observe that the quality of the produced matches across all classes of YAGO2 is really high. In the majority of the false matches we have entities of the official datasets that contained words such as *North* and *Lower*, and entities of YAGO2 that did not contain these words in their labels (e.g., Carryduff East of OSNI is matched with Carryduff of YAGO2 and Carryduff West of OSNI remains unmatched). There are many entities of OS and OSNI that are not matched. We extended matched geoentities and introduced unmatched entities to YAGO2.

4.3 Republic of Ireland: OSI dataset

Even though the provinces of Ireland (i.e., Ulster, Connach, Leinster and Munster) are no longer considered as administrative units, they can be found in

YAGO2, Ireland	(# Entities)	**OSI**	**# Matches**	**Precision**
first-order	(4)	-	0	-
second-order	(31)	councils	31	1,000
populated_place and locality	(13175)	baronies, parishes, townlands and rural areas	7193	0,786

Table 4. Republic of Ireland: results of the matching phase

YAGO2	(# Entities)	**GADM**	(#entities)	**# Matches**	**Precision**
countries	(233)	level-0	(256)	221	1.000
first-order	(3958)	level-1	(3610)	3127	0.987
second-order	(44554)	level-2	(45958)	31632	0.974
third-order	(121648)	level-3	(144608)	47579	0.972
fourth-order	(124729)	level-4	(137983)	46511	0.952
fifth-order	(51112)	level-5	(51427)	14	0.571

Table 5. GADM: results of the matching phase

the first administrative level of YAGO2. Irish city and county councils, county councils and city councils are instances of the class `geoclass_second_order_administrative_division` of YAGO2. The remaining administrative levels of YAGO2 do not contain any Irish entities. Like in the case of the UK, we try to match entities of OSI with `populated places` and `localities` of YAGO2.

Since provinces are no longer administrative units, we have zero matches in the first administrative level of YAGO2. In the second level we were able to match all councils. Regarding the remaining administrative divisions of Ireland (Section 3), we were not able to match any municipal districts, but we were able to match almost half of the baronies, parishes and rural areas. There are over 50000 townlands provided by OSI and we matched almost 7000. Table 4 shows the number of geoentities of YAGO2 that were matched.

4.4 GADM

Here we present the results of the matching phase between YAGO2 and GADM. Similalry to the previous cases, we manually align the classes of YAGO2 and the administrative levels of GADM, as shown in Table 5. The classes `geoclass_independent_political_entity`, `geoclass_dependent_political_entity` and `geoclass_semi-independent_political_entity` of YAGO2 contain countries and are combined together in order to be matched with the administrative `level-0` of GADM, that also contains countries.

The results of the matching phase between YAGO2 and GADM (Table 5) show that the precision of the generated matches, in most administrative levels, is really high. We also observe that at higher administrative levels the number of matches is close to the number of geoentities that exist in YAGO2. At lower levels the percentage of the geoentities that were matched drops.

After examining both YAGO2 and GADM and also the results of the matching phase, we see that each data source has its own view of the administrative hierarchies of a country. We also conjecture that these views might not fully reflect the current administrative situation of a country, especially a small one or one where the administrative divisions were reorganized recently. Let us consider the example of Greece with which we are very familiar and for which both properties hold (it is a small country and its latest administrative reorganization was done in 2011). As we have already mentioned in Section 4.1, municipal units and communities are not found in the administrative levels of YAGO2. The second level of YAGO2 contains some outdated information (i.e., prefectures) and some regional units that are not labeled properly. GADM does not provide information about the municipal units and communities, like YAGO2. In addition it does not contain the level of the regional units. Moreover, the Greek entities that are instances of *geoclass_first-order_administrative_division* of YAGO2 (i.e., regions), are found in the second administrative level of GADM. The first level of GADM contains (correctly) decentralized administrations.

We also closely examined the information provided by YAGO2 and GADM about German administrative units. We observed that the units that belong in the third administrative level of YAGO2 are found in the second level of GADM. On the other hand, both sources have almost the same German administrative units in the first and fourth levels and we were able to match almost all of them.

The results of the fifth administrative level are not satisfying, due to the fact that GADM contains only French and Rwandan administrative units in this level. That is not the case with YAGO2, which contains only a few entities that belong to France and Rwanda. Furthermore, French entities that are found in both YAGO2 and GADM (e.g., arrondissements of Paris) are not matched because the provided labels are not similar enough.

In YAGO2geo matched geoentities of YAGO2 are extended with information that is provided by GADM. The information GADM provides for Greece and the UK, even though it is missing administrative units, is of high quality. For that reason, we chose to bring into YAGO2geo unmatched entities of GADM.

4.5 OpenStreetMap

OpenStreetMap has geospatial information for many types of features, such as natural features (e.g, rivers, lakes, etc.) and man-made features (e.g., airports, restaurants, bars, etc.). For YAGO2geo, we focus on features that have a *permanent location*. The majority of these entities are features of nature (e.g, water bodies, waterways, etc.), but we also take into consideration other types as well, such as cities and islands. In Table 6 the groups of features, that are used in order to extend YAGO2, are shown. The group `natural` contains types of water bodies (i.e., lakes, reservoirs and lagoons), as well as beaches and bays. Streams and canals are part of the group `waterways`, whereas `landuse and leisure` consists of forests, parks and nature reserves. Last but not least, `places` contains islands, cities, villages and towns. The types of these groups are manually matched with the classes that are available in YAGO2. For example, forests

OSM Groups	# OSM Entities	# YAGO2 Entities	# Matches	Precision
natural	138640	437209	37447	0.957
waterways	1927776	1132239	137523	0.947
landuse and leisure	527403	131863	37502	0.963
places	189554	4674127	98444	0.952

Table 6. OpenStreetMap: results of the matching phase

of OSM are matched with forests (e.g., `geoclass_forest`) that are found in YAGO2, whereas cities, towns and villages are matched with populated places.

Regarding the results of the matching phase, Table 6 shows that the quantity of matches is relatively low compared to the number of entities provided by both data sources. There are two reasons that led to this situation. Firstly, as we already mentioned in Section 3, OSM is a volunteered, crowdsourced project, which means that it may contain noisy data. Such data ultimately do not contribute to our cause. Secondly, the labels for the majority of the entities of OSM are not available in multiple languages and in most cases they are written in the language of the country that they belong to. This problem affects our results negatively, even though YAGO2 provides the names of many geoentities in multiple languages. We could have produced more matches if we have used looser constraints in our filters but that would have had a negative impact on the quality of our results. Our main goal is to bring information of high quality to the YAGO2 knowledge graph and the results show that, regardless of the groups and the number of produced matches, the quality is always high. Since OSM contains noisy data we chose, unlike the case of GADM, to *only extend matched geoentities* and not bring unmatched entities to the knowledge graph.

4.6 Wikipedia and GeoNames

The geospatial information that already exists in YAGO2, as we have already mentioned, comes from Wikipedia and GeoNames. In this section we present the impact that both sources had during the matching phase. For each individual case, we count the number of matched geoentities of YAGO2 that come from Wikipedia as well as the number of geoentities that come from GeoNames.

The results are shown in Figure 1. In the case of GADM, we see that both Wikipedia and GeoNames have almost equal contribution to the produced matches. In the case of OpenStreetMap, over 90% of the matched geoentities come from GeoNames. Table 6 shows that most matches come from waterways and places. This means that Wikipedia does not contain enough information about these features. Consequently, information about these features found in YAGO2 was extracted from GeoNames. More specifically, the cities, villages and towns of OpenStreetMap are matched with the populated places of YAGO2. This agrees with the findings of [1] that states that most features of GeoNames are populated places and that streams are one of the most common natural features. This also explains the results we have for OSI and OSNI, since the majority of

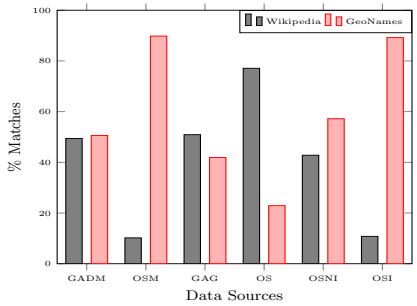

Fig. 1. The comparison between Wikipedia and GeoNames knowledge in YAGO2

their entities are matched with populated places. Last, we see that in the cases of GAG and OS, that most entities come from Wikipedia. It seems that Wikipedia provides rich information about the administrative units that belong to higher administrative levels for both Greece and Great Britain.

5 The Geospatial Knowledge in YAGO2geo

YAGO2geo is publicly available[19] at `http://yago2geo.di.uoa.gr` and its knowledge can be queried using GeoSPARQL[20] and can be visualised using the linked spatiotemporal data visualization tool Sextant[21]. Currently YAGO2geo is curated and used by our research group only but we expect other groups to use it once this paper is published.

YAGO2geo is structured as follows. For each official data source and GADM, we provide a file that contains the matched geoentities of YAGO2 extended with new knowledge and a file that contains the new entities of YAGO2geo. For the official datasets we also provide the topological relations between the administrative units that can be inferred by the geometric knowledge. For OSM, we only provide a file that contains extended geoentities of YAGO2. Last, for each data source we provide the generated matches and the new ontology.

This section discusses how YAGO2geo is enriched with new geospatial knowledge. We present a detailed example for the case of Greece and the GAG dataset. As we already discussed in Section 4, we follow the following uniqueness principle. For every geoentity g of YAGO2, only entities g', which are different than g and are not already in YAGO2 are introduced in YAGO2geo. If a geoentity is already in YAGO2 then its geospatial knowledge is *enriched* in YAGO2geo.

Let us consider the city of Lamia in Central Greece. In YAGO2, we have the following knowledge about Lamia (`<geoentity_Dimos_Lamia_8133738>`[22]):

[19] Published under the license found at https://creativecommons.org/licenses/by/4.0/
[20] http://test.strabon.di.uoa.gr/yago2geo
[21] http://test.strabon.di.uoa.gr/SextantOL3/?mapid=m95dp4hsgkafoe40_
[22] Dimos comes from "Δήμος" which means municipality in Greek

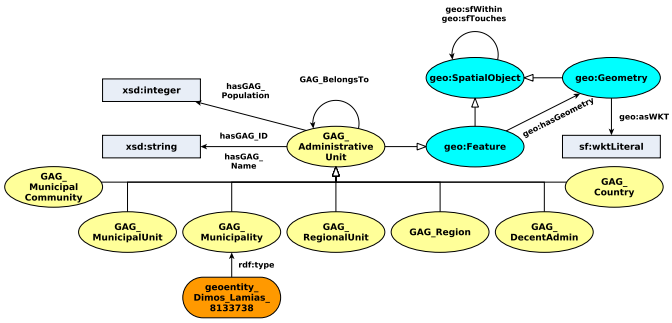

Fig. 2. The ontology of YAGO2geo: GAG part

1. `<geoentity_Dimos_Lamia_8133738> rdfs:label "Dimos Lamia"@eng .`
2. `<geoentity_Dimos_Lamia_8133738> <hasLatitude> "38.86649"^^<degrees>.`
3. `<geoentity_Dimos_Lamia_8133738> <hasLongitude> "22.36735"^^<degrees> .`
4. `<geoentity_Dimos_Lamia_8133738> rdfs:label "Lamia" .`
5. `<geoentity_Dimos_Lamia_8133738> rdfs:label "Lamieon" .`
6. `<geoentity_Dimos_Lamia_8133738> rdfs:label "Λαμιέων" .`
7. `<geoentity_Dimos_Lamia_8133738> <isLocatedIn> <Phthiotis> .`
8. `<geoentity_Dimos_Lamia_8133738> rdf:type <geoclass_third-order_administrative_division> .`

The triples 2 and 3 above give us latitude and longitude of the center of Lamia. Lamia is part of Phthiotis (`<Phthiotis>`), which is a prefecture of the former administrative divisions law Kapodistrias which preceded Kallikratis. This knowledge is encoded by an `isLocatedIn` relation, as shown in triple 7. Lamia includes the municipal units of Lamia, Gorgopotamos, Leianokladi, Pavliani and Ypati. In the Kapodistrias law, these units were previously municipalities themselves, but, since 2011, according to the Kallikratis law, they are no longer municipalities and they all belong to the municipality of Lamia. As it is expected, although YAGO2 contains these toponyms (`<Gorgopotamos>`, `<Leianokladi>`, `<Pavliani>` and `<Ypati>`), we do not have any `isLocatedIn` relations between these four municipal units and Lamia.

Figure 2 shows how the class hierarchy of YAGO2geo is extended with the GeoSPARQL ontology[23] and the ontology of GAG so that the geospatial knowledge extracted from the GAG dataset can be represented. Similar additions to the YAGO2geo class hierarchy have been done for OS, OSNI, OSI, GADM and OSM, but they are not shown here due to space.

As shown in Figure 2, a geoentity, like Lamia, becomes a GeoSPARQL feature (white arrows denote the `rdfs:subclassOf` property) and it is also associated with a geometry (see triples 5 and 6 below). This matched entity is extended with geospatial knowledge and also with additional knowledge provided by the GAG dataset. Here, we enrich the geoentity municipality of Lamia with its population,

[23] We do not show the complete GeoSPARQL ontology due to space

its identifier in the GAG dataset, its administrative division and its official name. This knowledge is encoded in YAGO2geo with the following triples:

1. <geoentity_Dimos_Lamia_8133738> y2geo:hasGAG_Populaton "71693"^^xsd:integer.
2. <geoentity_Dimos_Lamia_8133738> y2geo:hasGAG_Name "ΔΗΜΟΣ ΛΑΜΙΕΩΝ".
3. <geoentity_Dimos_Lamia_8133738> y2geo:hasGAG_ID "9160".
4. <geoentity_Dimos_Lamia_8133738> rdf:type y2geo:GAG_Municipality.
5. <geoentity_Dimos_Lamia_8133738> geo:hasGeometry y2geo:Geometry_GAG_9160.
6. y2geo:Geometry_GAG_9160 geo:asWKT "MULTIPOLYGON(((...)))" .
7. <geoentity_Dimos_Lamia_8133738> geo:sfWithin y2geo:gagentity_804.
8. <geoentity_Dimos_Lamia_8133738> geo:sfTouches <Amfikleia-Elateia>.
9. <geoentity_Dimos_Lamia_8133738> y2geo:GAG_BelongsTo y2geo:gagsentity_804.
10. y2geo:gagentity_916001 y2geo:GAG_BelongsTo <geoentity_Dimos_Lamia_8133738>.
11. <Gorgopotamos> y2geo:GAG_BelongsTo <geoentity_Dimos_Lamia_8133738>.
12. <Leianokladi> y2geo:GAG_BelongsTo <geoentity_Dimos_Lamia_8133738>.
13. <Pavliani> y2geo:GAG_BelongsTo <geoentity_Dimos_Lamia_8133738>.
14. <Ypati> y2geo:GAG_BelongsTo <geoentity_Dimos_Lamia_8133738>.

Triples 1, 2 and 3 contain thematic attributes (i.e., the population, the official name and the identifier) extracted from the GAG dataset. Triple 4 gives us the administrative division to which Lamia belongs. The geospatial knowledge obtained from GAG is encoded by triples 5 and 6. The detailed geometries that we bring into YAGO2geo allow us to make use of the GeoSPARQL topological vocabulary, as shown in the triples 7 and 8 above. The last four triples above model the information, that is missing from YAGO2, about the municipal units that are part of Lamia. The property `GAG_BelongsTo` is crucial, because the geometries of the municipal units and communities are not available in GAG, hence we are not able to generate topological relations that involve municipal units and communities. Last, triples 8 and 14 show entities (the regional unit of Pthiotis and the municipal unit of Lamia respectively) that are not part of YAGO2 and are created from unmatched entities of the GAG dataset. Extended and new entities of YAGO2geo follow the same schema.

6 Summary and Future Work

In this work we presented YAGO2geo, an extension of YAGO2 with precise geospatial knowledge. The new geospatial knowledge comes from official sources (e.g., Ordnance Survey), open source projects (e.g., GADM) and volunteer data sources (e.g., OSM). We expect that other users of YAGO2geo will want to add administrative divisions of their countries or other geospatial data (e.g., Natura 2000 areas, etc.) to the KG. Sections 4 and 5 will serve to guide these users.

In future work we plan to show how to model geospatial data that changes over time in YAGO2geo (e.g., evolution of administrative areas). For this purpose, we will extend the temporal dimension of YAGO2 with official data. Last, we will develop a geospatial question answering system on top of YAGO2geo.

Acknowledgments

We acknowledge the comments of G. Weikum, J. Hoffart and F. Suchanek on YAGO2geo. Part of this work was done while the third author was visiting Max Planck Institute for Informatics, Saarbrücken.

References

1. Acheson, E., Sabbata, S.D., Purves, R.S.: A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. Computers, Environment and Urban Systems **64** (2017)
2. Ahlers, D.: Assessment of the accuracy of GeoNames gazetteer data. In: GIR (2013)
3. Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J.: Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In: SIGMOD (2008)
4. Debruyne, C., Meehan, A., Clinton, É., McNerney, L., Nautiyal, A., Lavin, P., O'Sullivan, D.: Ireland's Authoritative Geospatial Linked Data. In: ISWC (2017)
5. Grütter, R., Purves, R.S., Wotruba, L.: Evaluating Topological Queries in Linked Data Using DBpedia and GeoNames in Switzerland and Scotland. Trans. GIS **21**(1) (2017)
6. Hoffart, J., Suchanek, F.M., Berberich, K., Lewis-Kelham, E., de Melo, G., Weikum, G.: YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In: WWW (2011)
7. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia. Artif. Intell. **194** (2013)
8. Jaro, M.A.: Probabilistic linkage of large public health data files. Statistics in medicine **14**(5-7) (1995)
9. Jia, Z., Abujabal, A., Roy, R.S., Strötgen, J., Weikum, G.: TEQUILA: Temporal Question Answering over Knowledge Bases. In: CIKM (2018)
10. Kyzirakos, K., Savva, D., Vlachopoulos, I., Vasileiou, A., Karalis, N., Koubarakis, M., Manegold, S.: Geotriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings. Journal of Web Semantics **52-53** (2018)
11. Levenshtein, V.I.: Binary codes capable of correcting deletions, insertions, and reversals. In: Soviet physics doklady. vol. 10 (1966)
12. Mahdisoltani, F., Biega, J., Suchanek, F.M.: YAGO3: A knowledge base from Multilingual Wikipedias. In: CIDR (2015)
13. Miller, G.A.: WordNet: A Lexical Database for English. Commun. ACM **38**(11) (1995)
14. Randell, D.A., Cui, Z., Cohn, A.G.: A spatial logic based on regions and connection. In: KR (1992)
15. Stadler, C., Lehmann, J., Höffner, K., Auer, S.: LinkedGeoData: A core for a web of spatial open data. Semantic Web Journal **3**(4) (2012)
16. Suchanek, F.M., Kasneci, G., Weikum, G.: Yago: a core of semantic knowledge. In: WWW (2007)
17. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A large ontology from Wikipedia and WordNet. Journal of Web Semantics **6**(3) (2008)
18. Tanon, T.P., Vrandecic, D., Schaffert, S., Steiner, T., Pintscher, L.: From Freebase to Wikidata: The Great Migration. In: WWW (2016)