
**NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS**

**SCHOOL OF SCIENCES**
**DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS**


**BSc THESIS**

# Extending YAGO2geo with geospatial information from other countries


**Georgios N. Giatrakos**

**Supervisor:** **Manolis Koubarakis,** Professor



**ATHENS**

**MAY 2020**


**ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ**

**ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ**
**ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ**


**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

# Επέκταση του YAGO2geo με γεωχωρικές πληροφορίες απο άλλες χώρες


**Γεώργιος Ν. Γιατράκος**

**Επιβλέπων:** **Μανόλης Κουμπαράκης,** Καθηγητή



**ΑΘΗΝΑ**

**ΜΑΪΟΣ 2020**

**BSc THESIS**

# Extending YAGO2geo with geospatial information from other countries

**Georgios N. Giatrakos**

**S.N.:** 1115201600036

**SUPERVISOR:** **Manolis Koubarakis,** Professor

**ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ**

Επέκταση του YAGO2geo με γεωχωρικές πληροφορίες απο άλλες χώρες

**Γεώργιος Ν. Γιατράκος**
**Α.Μ.:** 1115201600036

**ΕΠΙΒΛΕΠΩΝ:** **Μανόλης Κουμπαράκης,** Καθηγητή

# ABSTRACT

In recent years there has been a lot of progress in the area of knowledge graphs with a lot of well know graphs being developed. Among them DBPedia and YAGO. The YAGO knowledge graph developed by the Max Planck Institute combines data from various sources.

In our university, we have developed the YAGO2geo knowledge base that extends the administrative unit entities of YAGO2 with high accuracy geographical data from various countries. This work was performed by N. Karalis et. al. under the guidance of professor Manolis Koubarakis.

In our work, we further extend YAGO2geo with data from the National Boundary Dataset for the United States of America.

**SUBJECT AREA:** Semantic Web

**KEYWORDS:** YAGO2geo, knowledge graph, geospatial data, semantic web

# ΠΕΡΙΛΗΨΗ

Τα τελευταία χρόνια έχει υπάρξει σημαντική πρόοδος στον τομέα των γράφων γνώσης με πολλούς γράφους να αναπτύσονται.

Στο πανεπιστήμιο μας έχουμε αναπτύξει την βάση πληροφορίας YAGO2geo που επεκτείνει τις οντότητες διοικητικών μονάδων στο YAGO2 με υψηλής ακρίβειας γεωγραφικά δεδομένα απο διάφορες χώρες. Η δουλειά αυτή έγινε απο τους Ν. Κάραλη κ. α. υπο την επίβλεψη του καθηγητή Μανόλη Κουμπαράκη.

Στην εργασία αύτη, επεκτείνουμε περαιτέρω τον YAGO2geo με δεδομένα απο το National Boundary Dataset για τις Ηνωμένες Πολιτείες της Αμερικής.

**ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ:** Σημασιολογικός Ιστός

**ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ:** YAGO2geo, σημασιολογικός ιστός, γεωχωρικά δεδομένα, γράφοι γνώσης

*I dedicate this work to my parents and brother for supporting me*

# ACKNOWLEDGEMENTS

Για τη διεκπεραίωση της παρούσας Πτυχιακής Εργασίας, θα θέλαμε να ευχαριστήσουμε τον επιβλέποντα, καθ. Μανόλη Κουμπαράκη. Επίσης, τους Νικόλαο Κάραλη και Θοδωρή Στέφου, για τη συνεργασία και την πολύτιμη συμβολή τους στην ολοκλήρωση της.

# CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# PREFACE

Η πτυχιακή εργασία υλοποιήθηκε το Εαρινό Εξάμηνο του 2020. Ο επιβλέπωντας καθηγητής ήταν ο Μανώλης Κουμπαράκης του Πανεπιστημίου Αθηνών.

# 1. INTRODUCTION

The Semantic Web is a term first introduced by Tim Berners Lee. The main goal of the Semantic Web is to represent meaning and structured information in a way that is accessible by computers[3]. In this effort a standard was created by the W3C called RDF[1] and it is widely used to facilitate the Semantic Web. In conjuction with RDF, SPARQL[2] is used as a query language for RDF data.

Using RDF institutions have developed knowledge graphs containing vast amounts of information. One of those knowledge graphs is YAGO first released in 2007. [13]. In its first iteration it combined knowledge from Wikipedia and Wordnet. In its second iteration, YAGO2, it added entities from Geonames to enhance its knowledge with spatial and temporal information[7].

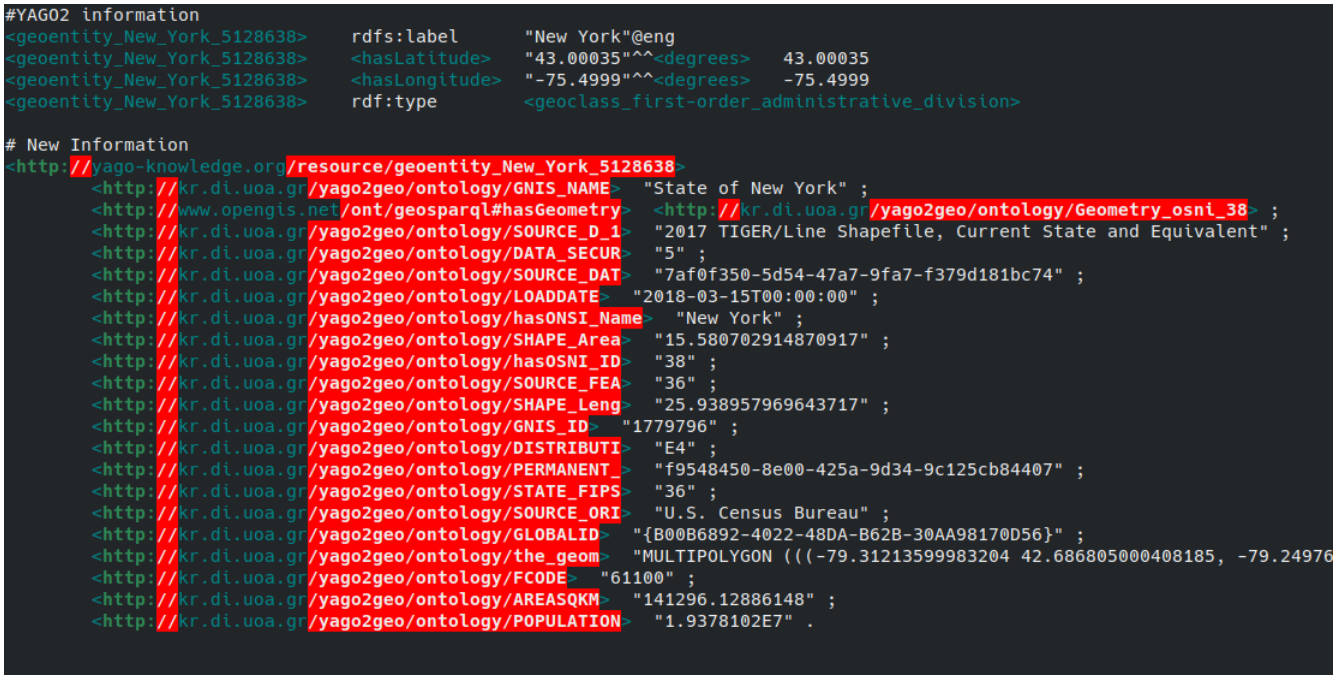
YAGO2geo further extends YAGO2 with qualitative geospatial information[8] about the administrative divisions of various countries. It uses official datasets for the countries of Northern Ireland, United Kingdom and Greece. In addition to thoses datasets it makes use of OpenStreetMap[3] and GADM[4].

In this work we further extend YAGO2geo with qualitative geospatial information using the National Boundary Dataset[5] for the United States of America.

YAGO2 is limited to only point coordinates for its geographical entities. An example of a YAGO2 entity is that of New York State. As we can see in Figure 1.1 YAGO2 only contains the coordinates of the State of New York. In our work we are able to expand the New York State entity with a polygon representing its geometry as well as other useful information present in NBD.

The rest of this thesis is structured as follows. Chapter 2 discusses background and related works. In chapter 3 we describe the data sources used. In chapter 4 we present the methodology for extending YAGO2geo. Finally, in Chapter 5 we summarize our contributions, present our conclusions and discuss future work.

[1] https://www.w3.org/RDF/
[2] https://www.w3.org/2001/sw/wiki/SPARQL
[3] https://www.openstreetmap.org/about
[4] https://gadm.org/
[5] https://www.sciencebase.gov/catalog/item/4f70b219e4b058caae3f8e19



```
#YAGO2 information
<geoentity_New_York_5128638>    rdfs:label      "New York"@eng
<geoentity_New_York_5128638>    <hasLatitude>   "43.00035"^^<degrees>   43.00035
<geoentity_New_York_5128638>    <hasLongitude>  "-75.4999"^^<degrees>   -75.4999
<geoentity_New_York_5128638>    rdf:type        <geoclass_first-order_administrative_division>

# New Information
<http://yago-knowledge.org/resource/geoentity_New_York_5128638>
        <http://kr.di.uoa.gr/yago2geo/ontology/GNIS_NAME>  "State of New York" ;
        <http://www.opengis.net/ont/geosparql#hasGeometry>  <http://kr.di.uoa.gr/yago2geo/ontology/Geometry_osni_38> ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_D_1>  "2017 TIGER/Line Shapefile, Current State and Equivalent" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/DATA_SECUR>  "5" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_DAT>  "7af0f350-5d54-47a7-9fa7-f379d181bc74" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/LOADDATE>  "2018-03-15T00:00:00" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/hasONSI_Name>  "New York" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SHAPE_Area>  "15.580702914870917" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/hasOSNI_ID>  "38" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_FEA>  "36" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SHAPE_Leng>  "25.938957969643717" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/GNIS_ID>  "1779796" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/DISTRIBUTI>  "E4" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/PERMANENT_>  "f9548450-8e00-425a-9d34-9c125cb84407" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/STATE_FIPS>  "36" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_ORI>  "U.S. Census Bureau" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/GLOBALID>  "{B00B6892-4022-48DA-B62B-30AA98170D56}" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/the_geom>  "MULTIPOLYGON (((-79.31213599983204 42.686805000408185, -79.24976
        <http://kr.di.uoa.gr/yago2geo/ontology/FCODE>  "61100" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/AREASQKM>  "141296.12886148" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/POPULATION>  "1.9378102E7" .
```

**Figure 1.1: New York State Entity**

# 2. BACKGROUND AND RELATED WORK

## 2.1 Semantic Web

The main idea behind the Semantic Web has been discussed above. In this section, we present to the reader the basics of RDF.

RDF stands for Resource Description Framework. It is a way to represent data on the internet. The most common way of representing RDF data is in triples. The triple representation is called ”Turtle” which stands for ”Terse RDF”. The three terms are called subject, predicate and object. RDF makes it easy to extend entities with new information and it offers a general way of representing relationships between entities as well as the facts that we may with to represent about an entity.

As an example, lets take the case of the city of Athens. We know that Athens is a city, that it has a population of 664,000 people and that it is the capital of Greece. In the picture bellow we see how this information is denoted in triple format. As we can see we are able

```
<http://domain.com/Athens> rdf:type <http://domain/City>.
<http://domain.com/Athens> <http://domain.com/hasPopulation> 664,000.
<http://domain.com/Greece> <http://domain.com/hasCapital> <http://domain.com/Athens>.
```

**Figure 2.1: Athens entity in RDF format**

to convey both numerical facts about Athens as well as relationships between Athens and other entities. This flexibility makes RDF suitable for expressing general knowledge.

In addition to the triple format we can also represent the triple information as a graph. In this graph each resource is represented by a node. The predicates correspond to edges in the graph that connect the subject and object. Back to our example, the Athens facts are shown as a graph in the picture bellow.

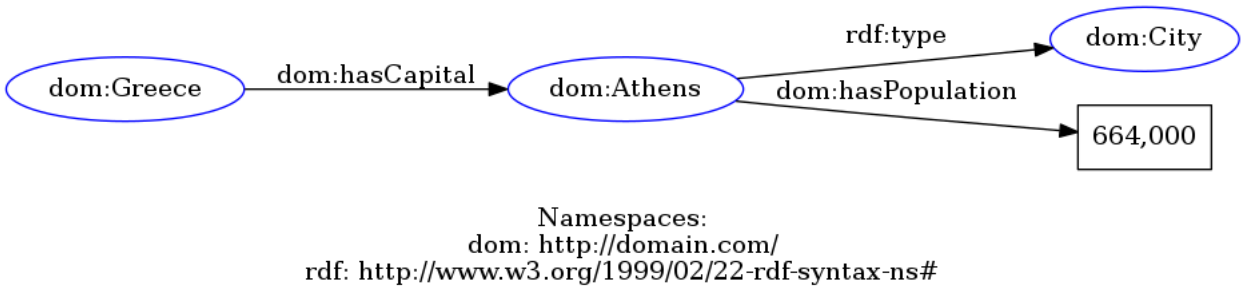



**Figure 2.2: Graph of the Athens entity**

We use the rdf:type predicate to denote that the entity Athens is of type dom:City. The rdf: and dom: are prefixes that point to the domain where we host our ontology. This very basic example of a triples database illustrates the expressive power of RDF. We are able to show information about the Athens entity, and its relationship with other entities.

## 2.2 Knowledge Graphs

The term knowledge graph is used to described a large collection of entities and their relationships organized in a graph[6]. In the previous chapter we described the YAGO2

knowledge graph. In this section we present two popular knowledge graphs, Wikidata and DBPedia.

DBPedia is a knowledge graph deriving its data from Wikipedia[1]. Wikipedia is a community edited encyclopedia. It has high traffic and is very actively maintained by the community. In addition to Wikipedia DBPedia has been linked with other open datasets like YAGO and GeoNames. The 4.5 million entities of DBPedia can be queried via a public SPARQL endpoint.

Wikidata is another popular knowledge graph.[14] Wikidata like Wikipedia is user edited and also actively maintained. The entities as well as the schema can be edited by the users. It boasts 88 million entities as of August 2020. Wikidata can be queried using SPARQL.

Other popular knowledge graphs include: Freebase[4], CYC[10] and many others.

## 2.3 YAGO2geo

The main work of this thesis will focus on extending the YAGO2geo knowledge graph with data from more countries. For that reason it is important to illustrate the structure of YAGO2geo.

The YAGO2geo knowledge graph was developed in our university by N. Karalis et. al., under the guidance of professor Koubarakis. It adresses the lack of qualitative geographical information in YAGO2. As mentioned above YAGO2 contains only coordinates for some of its entities and thus does not fully describe the shape of them. YAGO2geo adds polygons to entities of YAGO2 that more accurately portray the space they occupy. The term polygon refers to a collection of points that describe a two dimensional space. YAGO2geo can be queried using the stSPARQL [9] and GeoSPARQL[12] query languages.

YAGO2geo derives its data from official sources. Furthermore, it currently uses sources for the countries of Greece, United Kingdom and Northern Ireland. In addition, YAGO2geo also links the GADM and OpenSteetMaps datasets to YAGO2.

We will perform a basic query to get the information YAGO2geo has for the municipality of Athens. We will be using the SPARQL endpoint[1]. The query bellow returns all the triples where the ”Dimos Athens” entity appears as a subject.

```
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX yago: <http://yago-knowledge.org/resource/>
PREFIX y2geor: <http://kr.di.uoa.gr/yago2geo/resource/>
PREFIX y2geoo: <http://kr.di.uoa.gr/yago2geo/ontology/>

SELECT ?p ?o
WHERE{
        yago:geoentity_Dimos_Athens_8133876 ?p ?o .
}
```

**Figure 2.3: Municipality of Athens query**

The results contain the labels, population and id of the municipality of Athens as they

[1] http://test.strabon.di.uoa.gr/yago2geo/Query

appear in YAGO2geo. In addition to these information we also get the geometry of the entity. We are also able to display a map (Figure 2.4) of the entity using the Sextant tool . [2]

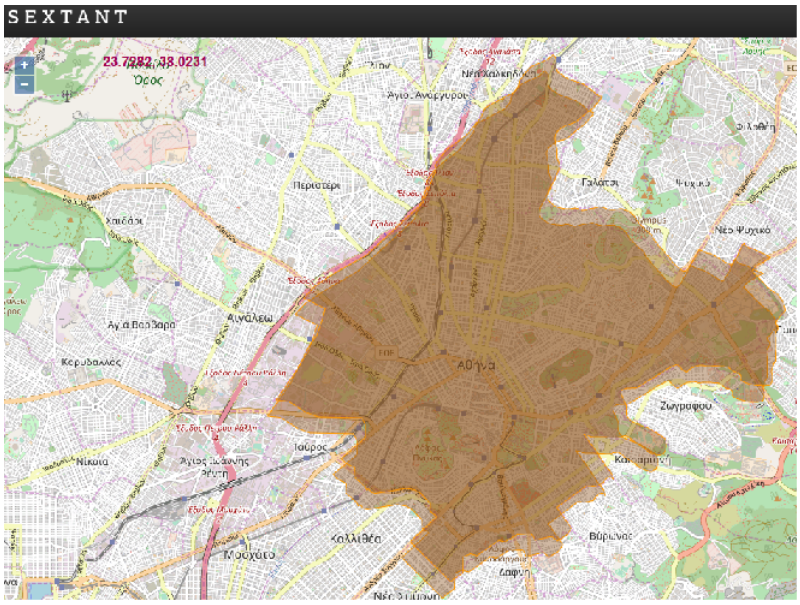



**Figure 2.4: Map of the municipality of Athens**

In addition to displaying the geometry of the Athens municipality we are also able to perform other helpful queries. For example, we can also show on the map all the municipalities that touch the municipality of Athens.

## 2.4 Summary

In this chapter we gave a brief overview of RDF. We also presented two popular knowledge graphs , DBPedia and Wikidata. Finally, we demonstrated the structure and information present in YAGO2geo

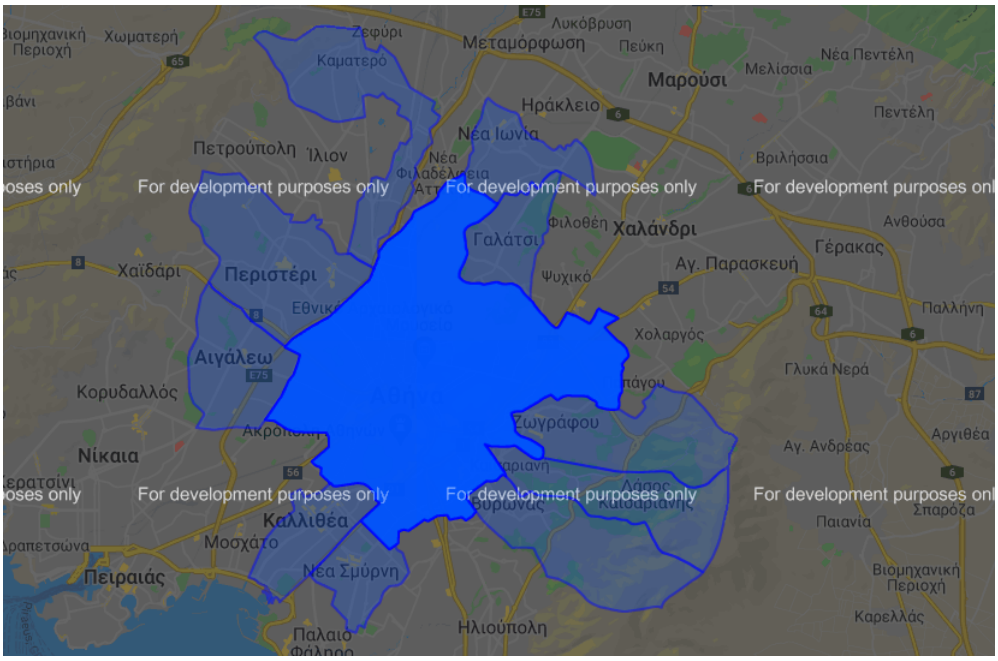



**Figure 2.5: Municipality of Athens and surrounding municipalities**

# 3. PRELIMINARIES

In this chapter, we present the source used to extend YAGO2geo. We also give a brief overview of the structure of YAGO2.

## 3.1 National Boundary Dataset

The National Boundary Dataset (NBD) was published by the United States Geological Service[1] and it contains qualitative information about the administrative divisions of the United Stated of America (USA).

NBD provides information about the following administrative divisions:

- State or Territory (56 entries)
- County (3233 entries)
- Minor Civil Division (36701 entries)
- Incorporated Place (19727 entries)
- Unincorporated Place (10126 entries)
- Reserve (8293 entries)
- Native American Area (857 entries)

In addition to the boundaries of each entity it includes other useful information, like FIPS[2] codes, population and area size.

### 3.1.1 Administrative Organization of United States

An administrative unit in the US can fall into one of three categories: i) federal, ii) state or iii) local. Each one of these levels has its own government unit and every citizen elects their representatives for each one. For example a citizen of New York City, votes for his representative in the Senate (federal level), in the New York State Assembly (state level) and for City council (local level).

There are also places that are of interest but do not have their own governments. These are called unincorporated places and are present in the layer of the same name.

#### Federal

In the federal level the US consists of Native American areas and reserves and the District of Columbia.

[1] https://www.usgs.gov/

[2] https://www.nist.gov/itl/publications-0/federal-information-processing-standards-fips

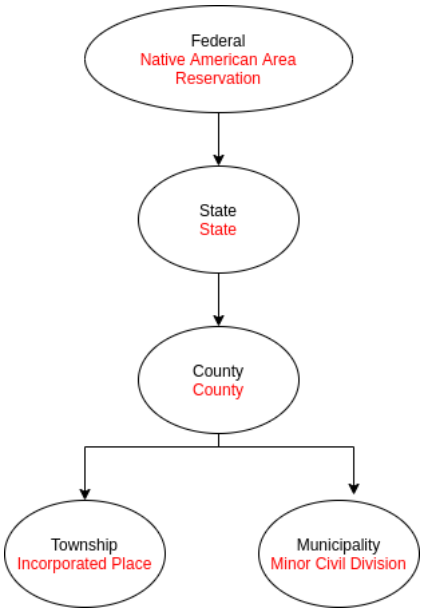



**Figure 3.1: Administrative levels and government units belonging to them**

### State

In the state level the US consists of 50 states and the following 5 incorporated territories:

- American Samoa
- Guam
- Northern Mariana Islands
- Puerto Rico
- US Virgin Islands

### Local

In the local level the US consists of Counties, Minor Civil Divisions, Incorporated and Unincorporated places. These subdivisions aren't always clear with cases where levels might be coextensive with each other.

The county is the first order subdivision of a state. There are two states that don't have counties, Alaska where they are called boroughs and Louisiana where they are called parishes. In the state of Connecticut there is no government in the county level. Counties do not fully cover all states, there are cases of cities that are independent of counties, e.g. Carson City in Nevada that borders 3 counties but is not part of any of them.

There are two main subdivisions of counties municipalities and townships. These are associated with minor civil divisions and incorporated places respectively. In general, incorporated places correspond to a concentration of people whereas minor civil divisions are independent of population.

In the dataset there is also the division of unincorporated place. This includes areas that are not administered by a government, but there is a population. In Figure 3.5 we show maps of the different administrative levels in the state of New York. In this case as we can

see there is overlap between the level of incorporated place (Figure 3.4) and minor civil division (Figure 3.3).

## 3.2 YAGO2geo

We have discussed the information present in YAGO2geo in a previous section. In this section we delve deeper into the structure of the graph.

YAGO2geo matches the data present in the above sources with the entities of YAGO2 using a novel matching algorithm that is based on their labels and coordinates. We will present and apply the algorithm in the following section to the NBD dataset.

Before moving to the matching phase we give a general overview of the classes YAGO2 uses for representing cities and administrative units.

YAGO2 represents hierarchies of administrative divisions with the classes geoclass_*-order_administrative division. In total there are 5 possible levels, but in the US these extend only to the third level.

In addition to the above classes there is also the more general term of populated place and locality. These terms refer to places where people may live and have significantly more entries. A place can be represented more than one time. For example, the state of New York has an entity of type geoclass_first-order_administative_division and one of type populated_place. This fact is important in the matching phase and will be discussed further in the results section.

In Table 3.1 we list the entities with their respective number of entries. The total number of entries is 136,978.

**Table 3.1: Classes of YAGO2 associated with administrative units**

| Class Name | Number of Entries |
|---|---|
| geoclass_populated_place | 166491 |
| geoclass_third-order_administrative_division | 29157 |
| geoclass_second-order_administrative_division | 3131 |
| geoclass_reserve | 1260 |
| geoclass_locality | 103 |
| geoclass_populated_locality | 100 |
| geoclass_first-order_administrative_division | 51 |
| geoclass_reservation | 22 |

## 3.3 Summary

In this chapter , we present the governmental units present in the NBD dataset. We also discuss the classes present in YAGO.

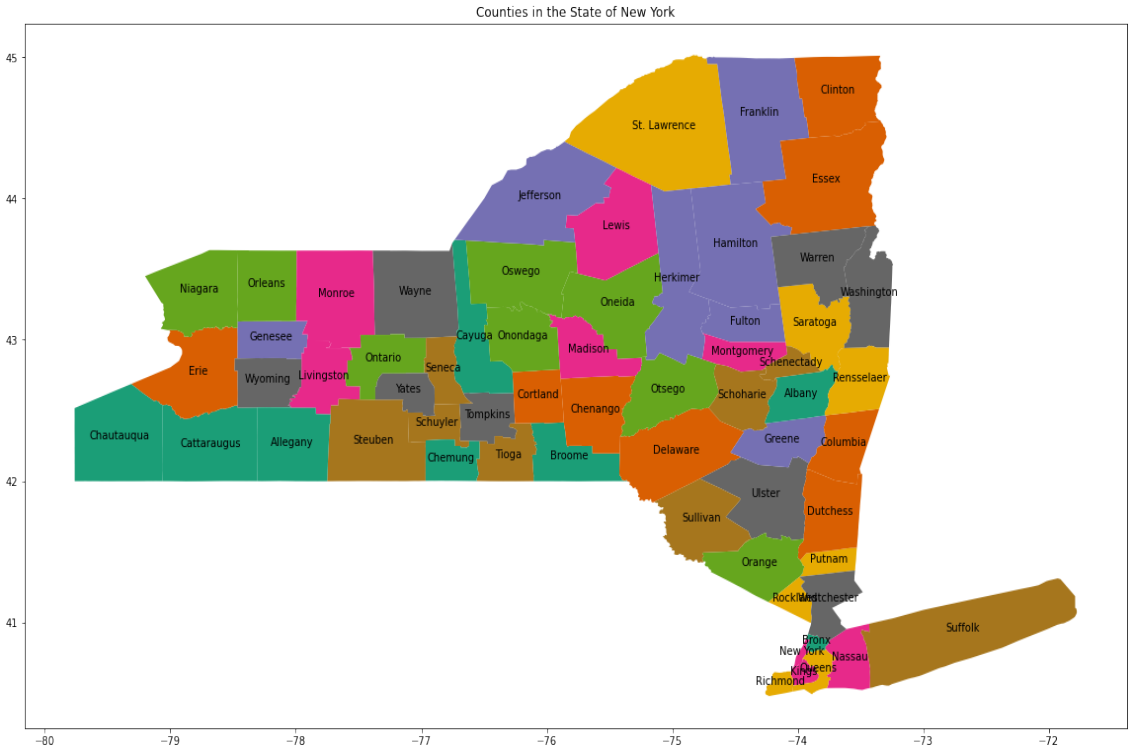
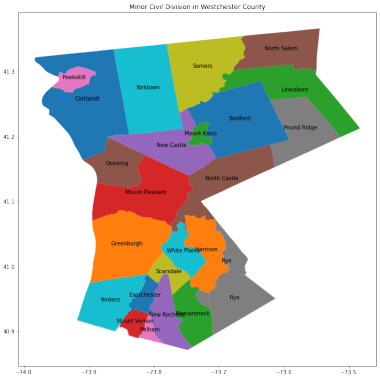
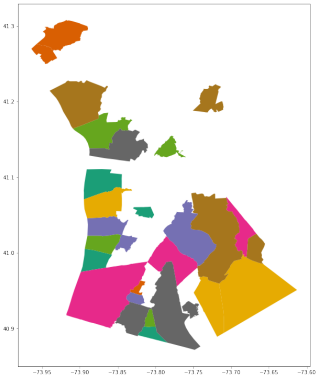



**Figure 3.2: Counties in the state of New York**




[b].4

**Figure 3.3: Minor Civil Divisions in the county of Westchester**



[b].4

**Figure 3.4: Incorporated places in the county of Westchester**

**Figure 3.5: Divisions of New York State**

# 4. EXTENSION OF YAGO2GEO

In this chapter, we describe the methodology we used to extend the YAGO2geo knowledge graph with the entities of NBD. We also discuss the results of our work.

## 4.1 Matching Phase

The first step in extending YAGO2geo is matching the entities of YAGO2 with those present in our dataset. To achieve this task we used the algorithm developed in YAGO2geo. Briefly, we first iterate through the entities of YAGO2 and calculate the label similarity with the NBD entities keeping those that are above a certain threshold. Next we select the NBD entity that is closest geographically to the YAGO2 entity. For the label similarity we tested various metrics such as Levehnstein and Jaro-Winkler distance[11][5]. We used levehnstein for the final results.

We started by filtering the YAGO entities keeping only those with coordinates. We made that choice since the geometry filter wouldn't work without it and the accuracy of our results wouldn't be high. After this initial filter we are left with 136978 entities.

We begin with the YAGO entities concerning administrative divisions. The first order of administrative division is matched with GU_StateOrTerritory. The second order of administrative division is matched with GU_CountyOrEquivalent. Lastly, the third order of administrative division is matched with both GU_IncorporatedPlace and GU_MinorCivilDivision.

The rest of the YAGO entities are matched as follows. The populated place class which has the highest number of entities is matched with all levels of NBD. In the same way populated locality and locality are also matched with all NBD entities. Finally, reserve and reservation are matched with the Reserve layer of NBD.

In order to get better results we apply text processing techniques. First, we remove certain stopwords from the labels to increase our matching rate. For example, the New York State entity in our YAGO dataset has the label ”New York State” but the one present in NBD is called New York. In order to properly match them we remove the stopword ”State”. Similarly, we have chosen other stopwords like City, Town etc to get better results. This technique is quite effective and we are able to match a large percentage of the 3 administrative divisions of YAGO as shown in Table 4.1. We also change the word Saint to St. to match the formatting used in NBD. In Figure 4.2 we see the extended New York entity with the information added from NBD.

## 4.2 Results

In this section we present our results after the matching phase.

As we can see the matching rate is high for the administrative order hierarchy, but lower for the non-hierarchical populated place categories. We are able to match around 98% of the three administrative levels with the corresponding NBD entities. This is due to the high accuracy of the information present in NBD. Another reason is the geometry filter applied after the label similarity features that disambiguates the cases where places have the same label. Also due to the geometry filter picking the closest entity we resolve the cases of adjacent entities with the same name.

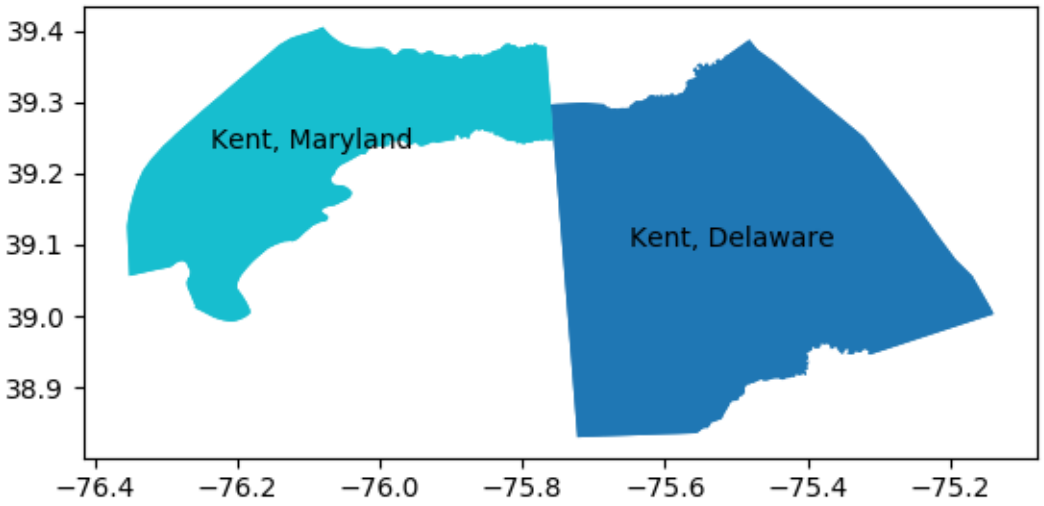



**Figure 4.1: Kent County Maryland adjacent to Kent County Delaware**

**Table 4.1: Matching Phase results**

| YAGO Entity | NBD Entity | Matches | Correct Matches |
|---|---|---|---|
| First order administrative division (51) | State or Territory (56) | 50 | 50/50 |
| Second order administrative division (3131) | County or Equivalent (3233) | 3117 | 3117/3117 |
| Third order administrative division (29157) | Incorporated Place (19728) | 9478 | 3453/3500 |
| Third order administrative division (29157) | Minor Civil Division (36702) | 19082 | 3500/3500 |
| Reserve(1260) | Reserve(3115) | 747 | 746/747 |
| Populated Place(165461) | State or Territory(56) | 26 | 26/26 |
| Populated Place(165461) | County or Equivalent(3233) | 1007 | 1003/1007 |
| Populated Place(165461) | Incorporated Place(19728) | 15654 | 290/300 |
| Populated Place(165461) | Minor Civil Division(36702) | 6857 | 283/350 |
| Populated Place(165461) | Unincorporated Place(10109) | 7563 | 290/300 |

The populated place category is harder to match achieving only 18% matching rate. The populated places category contains a vast number of entities that correspond to any place that people live eg. neighborhoods, beaches, parks etc. This information is not present in NBD and thus not matched.

```
#YAGO2 information
<geoentity_New_York_5128638>    rdfs:label      "New York"@eng
<geoentity_New_York_5128638>    <hasLatitude>   "43.00035"^^<degrees>   43.00035
<geoentity_New_York_5128638>    <hasLongitude>  "-75.4999"^^<degrees>   -75.4999
<geoentity_New_York_5128638>    rdf:type        <geoclass_first-order_administrative_division>

# New Information
<http://yago-knowledge.org/resource/geoentity_New_York_5128638>
        <http://kr.di.uoa.gr/yago2geo/ontology/GNIS_NAME>  "State of New York" ;
        <http://www.opengis.net/ont/geosparql#hasGeometry>  <http://kr.di.uoa.gr/yago2geo/ontology/Geometry_osni_38> ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_D_1>  "2017 TIGER/Line Shapefile, Current State and Equivalent" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/DATA_SECUR>  "5" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_DAT>  "7af0f350-5d54-47a7-9fa7-f379d181bc74" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/LOADDATE>  "2018-03-15T00:00:00" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/hasONSI_Name>  "New York" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SHAPE_Area>  "15.580702914870917" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/hasOSNI_ID>  "38" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_FEA>  "36" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SHAPE_Leng>  "25.938957969643717" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/GNIS_ID>  "1779796" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/DISTRIBUTI>  "E4" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/PERMANENT_>  "f9548450-8e00-425a-9d34-9c125cb84407" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/STATE_FIPS>  "36" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/SOURCE_ORI>  "U.S. Census Bureau" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/GLOBALID>  "{B00B6892-4022-48DA-B62B-30AA98170D56}" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/the_geom>  "MULTIPOLYGON (((-79.31213599983204 42.686805000408185, -79.24976
        <http://kr.di.uoa.gr/yago2geo/ontology/FCODE>  "61100" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/AREASQKM>  "141296.12886148" ;
        <http://kr.di.uoa.gr/yago2geo/ontology/POPULATION>  "1.9378102E7" .
```

**Figure 4.2: Extended New York Entity**

# 5. CONCLUSIONS AND FUTURE WORK

In this thesis we extended the YAGO2geo knowledge graph with entities located in the United States of America. In order to achieve that we used information from an official dataset. This dataset is called National Boundary dataset.

We used the same matching algorithm as in the original YAGO2geo paper to match the entities of the two datasets. Our results are very precise for the hierarchical entities of YAGO2 and we achieve very high matching rates. We were not able to achieve as high matching rates for the un-hierarchical YAGO classes. The knowledge graph is extended with entities following the same Open Geospatial Consortium standard that is compatible with GeoSPARQL and stSPARQL.

For our thesis we used the official US dataset published by the United States Geological Survey. The quality of the information present in it is very high and up to date. The dataset is actively mantained. We are hoping that this work will be useful for further research into linked geo data. It will be available in a public endpoint used to access all the information present in YAGO2geo.

Our future work will focus on adding more countries in YAGO2geo using official data sources.

# ABBREVIATIONS - ACRONYMS

| | |
|---|---|
| RDF | Resource Description Framework |
| SPARQL | SPARQL Protocol and RDF Query Language |
| OWL | Web Ontology Language |
| OGC | Open Geospatial Consortium |
| US | United States of America |
| NBD | National Boundary Dataset |
| YAGO | Yet Another Great Ontology |

# BIBLIOGRAPHY


[1] Sören Auer et al. “Dbpedia: A nucleus for a web of open data”. In: *The semantic web*. Springer, 2007, pp. 722–735.

[2] Konstantina Bereta et al. “SexTant: Visualizing Time-Evolving Linked Geospatial Data.” In: *International Semantic Web Conference (Posters & Demos)*. 2013, pp. 177–180.

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. “The semantic web”. In: *Scientific american* 284.5 (2001), pp. 34–43.

[4] Kurt Bollacker et al. “Freebase: a collaboratively created graph database for structuring human knowledge”. In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 1247–1250.

[5] William W Cohen, Pradeep Ravikumar, Stephen E Fienberg, et al. “A Comparison of String Distance Metrics for Name-Matching Tasks.” In: *IIWeb*. Vol. 2003. 2003, pp. 73–78.

[6] Lisa Ehrlinger and Wolfram Wöß. “Towards a Definition of Knowledge Graphs.” In: *SEMANTiCS (Posters, Demos, SuCCESS)* 48 (2016), pp. 1–4.

[7] Johannes Hoffart et al. “YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia”. In: *Artificial Intelligence* 194 (2013), pp. 28–61.

[8] Nikolaos Karalis, Georgios Mandilaras, and Manolis Koubarakis. “Extending the YAGO2 Knowledge Graph with Precise Geospatial Knowledge”. In: *International Semantic Web Conference*. Springer. 2019, pp. 181–197.

[9] Manolis Koubarakis and Kostis Kyzirakos. “Modeling and querying metadata in the semantic sensor web: The model stRDF and the query language stSPARQL”. In: *Extended Semantic Web Conference*. Springer. 2010, pp. 425–439.

[10] Douglas B Lenat. “CYC: A large-scale investment in knowledge infrastructure”. In: *Communications of the ACM* 38.11 (1995), pp. 33–38.

[11] Vladimir I Levenshtein. “Binary codes capable of correcting deletions, insertions, and reversals”. In: *Soviet physics doklady*. Vol. 10. 8. 1966, pp. 707–710.

[12] Matthew Perry and John Herring. “OGC GeoSPARQL-A geographic query language for RDF data”. In: *OGC implementation standard* 40 (2012).

[13] Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. “Yago: a core of semantic knowledge”. In: *Proceedings of the 16th international conference on World Wide Web*. 2007, pp. 697–706.

[14] Denny Vrandečić and Markus Krötzsch. “Wikidata: a free collaborative knowledgebase”. In: *Communications of the ACM* 57.10 (2014), pp. 78–85.