NATIONAL AND KAPODISTRIAN UNIVERSITY OF ATHENS

SCHOOL OF SCIENCE
DEPARTMENT OF INFORMATICS AND TELECOMMUNICATIONS

POSTGRADUATE STUDIES
"INFORMATION AND DATA MANAGEMENT"

MASTER THESIS

Extending the YAGO Knowledge Graph with Geospatial Knowledge

Nikolaos S. Karalis

Supervisor: **Manolis Koubarakis,** Professor

ATHENS

October 2018

ΕΘΝΙΚΟ ΚΑΙ ΚΑΠΟΔΙΣΤΡΙΑΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΣΧΟΛΗ ΘΕΤΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
“ΔΙΑΧΕΙΡΙΣΗ ΠΛΗΡΟΦΟΡΙΑΣ ΚΑΙ ΔΕΔΟΜΕΝΩΝ”

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επέκταση του γράφου γνώσης YAGO με γεωχωρική γνώση

Νικόλαος Σ. Καράλης

Επιβλέπων: **Μανόλης Κουμπαράκης,** Καθηγητής

ΑΘΗΝΑ

Οκτώβριος 2018

MASTER THESIS

Extending the YAGO Knowledge Graph with Geospatial Knowledge

Nikolaos S. Karalis
R.N.: M1518

SUPERVISOR: **Manolis Koubarakis,** Professor

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Επέκταση του γράφου γνώσης YAGO με γεωχωρική γνώση

Νικόλαος Σ. Καράλης
Α.Μ.: M1518

ΕΠΙΒΛΕΠΩΝ: Μανόλης Κουμπαράκης, Καθηγητής

ABSTRACT

YAGO is one of the largest knowledge bases that provide their data as Linked Open Data. Spatial information, in the form of points, was introduced in YAGO2, the second version of YAGO. In this work we present an extension of YAGO with qualitative geospatial information (i.e., polygons and lines), which was extracted from multiple sources. We studied datasets that are provided from crowdsourced projects as well as from official sources of several countries. It is important to point out that we do not introduce duplicate information in the knowledge graph of YAGO, by creating entities that already exist. Hence, at first, we try to match entities of YAGO with the entities of the data sources that we used. Our results show that our methodology produced matches with very high precision. This work is concluded with a demonstration of the extended knowledge graph.

SUBJECT AREA: Semantic Web

KEYWORDS: linked open data, semantic web, knowledge graphs, knowledge bases

ΠΕΡΙΛΗΨΗ

Η βάση γνώσης YAGO είναι μία από τις μεγαλύτερες βάσεις γνώσεις, που διαθέτουν τα δεδομένα τους ως ανοιχτά διασυνδεδεμένα δεδομένα. Χωρική πληροφορία, δηλαδή η αναπαράσταση της τοποθεσίας οντοτήτων με ένα σημείο, προστέθηκε στη δεύτερη έκδοση του YAGO. Σε αυτή τη δουλειά έχουμε ως σκοπό να επεκτείνουμε το γράφο γνώσης του YAGO με ποιοτική γεωχωρική πληροφορία (πολύγωνα και ευθείες), η οποία προέρχεται από πολλαπλές πηγές. Μελετήσαμε δεδομένα τα οποία διανέμονται όχι μόνο από έργα που βασίζονται στον πληθοπορισμό αλλά και από επίσημες πηγές διαφόρων κρατών. Είναι σημαντικό να μην προσθέσουμε στο γράφο γνώσης πληροφορία που ήδη υπάρχει σε αυτόν και γι' αυτό το λόγο ψάχνουμε συσχετίσεις μεταξύ των οντοτήτων του YAGO και εκείνων που ανήκουν στα σύνολα δεδομένων που εξετάσαμε. Τα αποτελέσματα δείχνουν πως η μεθοδολογία μας παρήγαγε συσχετίσεις με πολύ μεγάλη ακρίβεια. Στο τέλος της εργασίας αυτής παρουσιάζουμε τον επεκταμένο γράφο γνώσης.

ΘΕΜΑΤΙΚΗ ΠΕΡΙΟΧΗ: Σημασιολογικός Ιστός

ΛΕΞΕΙΣ ΚΛΕΙΔΙΑ: ανοικτά διασυνδεδεμένα δεδομένα, σημασιολογικός ιστός, γράφοι γνώσης, βάσεις γνώσης

I dedicate this work to my parents and brother

ACKNOWLEDGEMENTS

I would like to thank Prof. Manolis Koubarakis for making me a part of his research group and giving me the opportunity to work on this subject. Furthermore, I would like to thank the members of Knowledge Representation, Reasoning and Analytics research group for their assistance, suggestions and feedback. Finally, I would like to thank Prof. Gerhard Weikum and Johannes Hoffart for their valuable input on this work during my visit at Max Planck Institute for Informatics, Saarland.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ALGORITHMS

LISTINGS

1. INTRODUCTION

There are continuous efforts in order to make data publicly available. With the attention that the Semantic Web [4] has received over the last years, large knowledge bases have been created that provide their data as linked open data. DBpedia [3] is one of the largest knowledge bases and its data comes from the structured content of Wikipedia[1]. Wikidata [32] is a collaboratively edited knowledge base and is the successor of Freebase [5]. In this work we focus on the YAGO knowledge base.

The first version of YAGO was released in 2007 [29, 30]. YAGO was created by combining knowledge from two different sources, WordNet [19] and Wikipedia, and it is one of the first knowledge bases, that was created from multiples sources. The entities of YAGO were created from articles of Wikipedia, whereas WordNet was used to create its classes and their hierarchy. Along with the YAGO knowledge base, the YAGO model was introduced, which extends RDFS[2] in order to support relations between facts and relations (i.e., relations between entities).

YAGO2 [11, 12], the second version of YAGO, was released in 2011. YAGO2 introduces spatial and temporal information to the YAGO knowledge graph. Wikipedia is not the only source from which YAGO2 extracts spatial information. The YAGO knowledge base is extended with information from a new source, GeoNames [33]. GeoNames is a gezetteer, whose data and accuracy have been studied extensively [1, 2, 10]. Temporal information was added mainly to entities that represent *people*, *groups*, *artifacts* or *events*. YAGO3 [18], the latest version of YAGO, came out in 2015. YAGO3 combines information from Wikipedias in multiple languages.

The goal of this work is to extend the knowledge graph of YAGO with *more geospatial information*. The spatial information in YAGO is represented with the properties *hasLongitute* and *hasLatitude*. An example of a geo-entity is shown in Figure 1. Our aim is to extend YAGO with geometries from multiple sources. We consider data sources that provide data about the boundaries of administrative units, such as GADM[3] and Ordnance Survey[4], which is the mapping agency of Great Britain. We also use data provided by OpenStreetMap[5] to extend the geospatial information of addtional entities that are not necessarily administrative units (e.g., forests, lakes, etc.)

```
@base <http://yago-knowledge.org/resource/> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
#current geospatial information
<geoentity_Dimos_Athens_8133876> rdfs:label "Dimos Athens"@eng .
<geoentity_Dimos_Athens_8133876> <isLocatedIn> <geoentity_Nomarchía_Athínas_445408> .
<geoentity_Dimos_Athens_8133876> <hasLatitude> "37.98888"^^<degrees> .
<geoentity_Dimos_Athens_8133876> <hasLongitude> "23.73604"^^<degrees> .
#new geospatial information, example
<geoentity_Dimos_Athens_8133876> geo:hasGeometry <Geometry_Athens> .
<Geometry_Athens> geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 1: Geospatial information in YAGO.

[1] https://www.wikipedia.org/
[2] https://www.w3.org/TR/rdf-schema/
[3] https://gadm.org/
[4] https://www.ordnancesurvey.co.uk/
[5] https://www.openstreetmap.org/

The rest of this thesis is structured as follows. Chapter 2 discusses related works. Chapter 3 gives detailed information about the data sources that were used in order to extend YAGO with geospatial information. The methodology that we followed to extend YAGO as well as the results of this work are shown in Chapter 4. In Chapter 5 we demonstrate the extended knowledge graph of YAGO. Last, in Chapter 6 we summarize our contributions, present our conclusions and discuss future work.

2. RELATED WORK

2.1 Knowledge Graphs

Detailed information about the knowledge graph of YAGO was given in the previous chapter. In this section we give detailed information about popular knowledge graphs and knowledge bases.

DBpedia [3] is probably the most popular knowledge graph in the area of the Semantic Web and a very important piece of the Linked Data Cloud. Its data comes from the structured information of Wikipedia (i.e., infoboxes, meta-data of arcticles, etc.). DBpedia is also iterlinked with various open knowledge graphs and datasets, like YAGO and GeoNames. DBpedia can be quired via its online SPARQL endpoints.

Freebase [5] was a collaborative database system that focused on human knowledge. Every real world object in Freebase was represented by a single identifier. For editing purposes it provided a web based appication programming interface. Freebase was shut down in 2016.

Wikidata [32], which is the successor of Freebase, is a free and collaborative knowledge graph, which can be edited by any user. The users of Wikidata are able to control its data as well as its schema. It is a multilingual knowledge base, and unlike Wikipedia which has different versions for every language, the information of the entities of Wikidata is translated to multiple languages. Wikidata is a part of the Linked Data Cloud, since its data is available in RDF.

CYC [16], KnowItAll [9], NELL [6], BabelNet [20], KnowledgeVault [8] and DeepDive [35] are also well-known knowledge bases.

2.2 Geospatial Entity Resolution

Entity resolution is the task of recognizing and linking different representations of a real-world objects that are found in multiple data sources. JedAI [22] is tool that provides an open source library that implements several state-of-the art tasks of entity resolution, which can also be compared with each other. In addition, a graphical user interface is provided, that can be also used by inexperienced users. JedAI can be applied both on structured (e.g., relational databases) and semi-structured (e.g., SPARQL endpoints) data. The second version of JedAI, JedAI 2.0 [23], has improved time efficiency, effectiveness and usability. The authors of [31] created Silk which is a tool that produces links between different data sources that are stored in SPARQL endpoints. SILK is extended in order to support GeoSPARQL [24] functions, in order to create links between geo-entities. [27]. GeoDDupe is a system for geospatial entity resolution that implements data mining algorithms and provides an interactive graphical user interface.

The work of Kamalloo and Rafiei [13] proposes three unsupervised methods for the problem of toponym resolution over documents. Toponyms are names for geographic entities. Toponym resolution is the task of assigning coordinates to toponyms. The first method uses the context of the documents to geotag the toponyms, whereas the second method takes focuses on the spatal hierarchy of the toponyms. The third method is a combination of the previous methods. The results of this work show that the third method outperfoms, in terms of precision, the state-of-the-art unsupervised method, but not the supervised

methods, which on the other hand rely on training data.

NewsStand [26] is a web application that visualizes events on a map. Its data comes from news articles that are retrieved from really simple syndication (RSS) feeds, which are also grouped based on their topic. In order to visualize the proper location of the events, NewsStand geotags the geographic location that are found in the news articles. The task of geotagging consists of two phases, the toponym recognition phase and the toponym resolution phase.

Our methodology is based on the methodology that was carried out in YAGO2 [12]. The core of YAGO is Wikipedia, from which its first batch of spatial data came from. Afterwards, Hoffart et al. [12] matched entities of Wikipedia with entities of GeoNames using their labels and their coordinates. Finally, all unmatched entities of GeoNames were added to YAGO. Stadler et al. in [28] created LinkedGeoData, which provides the data of OpenStreetMap in RDF. This data is interlinked with DBpedia, GeoNames and Geopolitical Data of the Food and Agriculture Organization of the United Nations. LinkedGeoData uses a function, that takes into consideration the labels of the entities and their distance, in order to determine if two entities should be interlinked. Unlike YAGO2, the distance threshold is not constant and it is accossiated to the classes to which each entities belongs.

2.3 Geospatial Information in current Knowledge Graphs

We have already mentioned that the geospatial information that is currently present in YAGO is limited to points, which are represented with *hasLongitude* and *hasLatitude* properties.

Punjani et al. in [25], for the purposes of the evaluation of their geospatial question answering system over linked data, created a golden standard geospatial dataset by interlinking DBpedia, OpenStreetMap and GADM. From each data source they used information about the United Kingdom and the Republic of Ireland. The interlinking process that Punjani et al. followed is based on the iterlinking process of LinkedGeoData. This knowledge graph exploits the qualitative geospatial information that is available in DBpedia and also the quantitative geospatial information of GADM and OSM. Younis et al. [34] created a system that takes as input structured geospatial queries. Their system queries DBpedia which is interlinked with Ordnance Survey.

Grütter et al. in [10] carried out an extensive evaluation of topological relations found in DBpedia and GeoNames about the administrative divisions of Switzerland and Scotland. They also examined whether different versions of DBpedia, namely the English and German versions, provide the same topological relations for Switzerland. The results of their work show that the values of recall and precision are ralatively high when DBpedia is queried via GeoNames 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. The English version of DBpedia provides complete and of high quality information about the topological relations between the administrative units of Switzerland, whereas the original links between DBpedia and GeoNames cover less than 20% of the administrative units. Last, they also found out the the German version of DBpedia does not provide any topological relations.

2.4 Summary

In this chapter we present well-known knowledge graphs and knowledge bases, like DBpedia and Wikidata. Additionaly, we discuss the problem of geospatial entity resolution and refer to works that try to solve this problem. Finally, we present works that use and evaluate knowledge graphs with geospatial information.

3. PRELIMINARIES

In this chapter, we present the various data sources that were used in order to extend the YAGO knowledge graph. The geospatial information that was used to extend YAGO comes from well known projects as well as from official sources.

3.1 GADM

GADM provides spatial data about the administrative divisions of every country. The information is divided into six different layers (i.e., administrative levels) and there are over 386,000 administrative areas. GADM does not only provide the boundaries of every administrative area, but it also provides additional useful information about them. For each entity, along with its identifier and geometry, we also extract the national level and its upper level administrative unit, as well as its name and its administrative division in English and its native language.

Any interested parties are able to download spatial data for the entire world or for specific countries. The data is available in several formats, such as *geopackages*, *shapefiles*, *R files* and *KMZ files*. The geometries are represented in the WGS84 coordinate system. Version 3.6 of GADM was released in May 2018.

3.2 OpenStreetMap

OpenStreetMap (OSM) is a crowdsourced project, whose goal is to provide free geographic data and maps to its users. OSM provides geospatial information about multiple feautures[6]. Such features are natural features (e.g., beaches, springs, etc.), land use features (e.g., forests, etc.), places (e.g., localities, cities, etc.), points of interest, water bodies, waterways and more. Every entity of OpenStreetMap is associated with an identifier, a feature class, a label and a geometry, which are used to extend YAGO.

There are multiple ways to access and export the data of OSM. The Overpass API allows users to extract data about features that are within a specified bounding box. From Planet OSM[7] users are able to get a complete copy of all OSM data, which is weekly updated. In this work, we obtain the data from Geofabrik[8], which is a company that provides free, regularly-updated extracts of OSM in various file formats. The data that we used were released in September 2018.

3.3 Official Data Sources

In the sections that follow we present the official data sources that were used to extend YAGO. For the administrative units of Greece we use the information that is provided by the Kallikratis dataset. We use the data of Ordnance Survey and Ordnance Survey Northern Ireland for the administrative divisions of the United Kingdom and the data of Ordnance

[6] http://sites.pyravlos.di.uoa.gr/dragonOSM.svg
[7] https://planet.openstreetmap.org/
[8] https://www.geofabrik.de/

Survey Ireland for the administative units of the Republic of Ireland. We used these data source because we are familiar with the Greek administrative divisions and because the data of the UK and Ireland are in English. In order to understand the administrative divisions of a country it is important to be familiar at least with its language.

3.3.1 Kallikratis Dataset

Geospatial information about the administrative divisions of Greece is available in the the Kallikratis dataset[9], which has been created from official sources. The Kallikratis law defines the administrative divisions of Greece and is valid since 2011. The administrative divisions of Greece are the following:

- Country
- Decentralized Administrations
- Regions
- Regional Units
- Municipalities
- Municipal Units
- Municipal Communities

We extend YAGO with the identifier, the Greek name of each entity. Most entities also have a population in the Kallikratis dataset. This is information is also added to YAGO.

3.3.2 Ordnance Survey

Ordnance Survey (OS) is a national mapping agency in the United Kingdom. It provides data about the countries of England, Scotland and Wales (i.e., Great Britain). For our purposes we used the Boundary-Line dataset[10], 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

[9] http://linkedopendata.gr/dataset/greek-administrative-geography
[10] https://www.ordnancesurvey.co.uk/business-and-government/products/boundaryline.html

- Communities

Apart from the boundaries, we also extract the names, the description (i.e. administrative division) and the area code of every unit.

3.3.3 Ordnance Survey Northern Ireland

Ordnance Survey Northern Ireland (OSNI)[11] is the official cartographic agency of Northern Ireland. Users are able to obtain its data using the ONSI Open Data portal [12]. We utilize the following datasets:

- NI Outline
- Local Goverment Districts 2012
- Wards 2012
- Townlands

From every entity we obtain its administrative division, its name, its identifier and its geometry. The datasets also provide the area of Northern Ireland, local goverment districts and townlands. The perimeter of every townland is also provided. This information contributes to the extension of the knowledge graph of YAGO.

3.3.4 Ordnance Survey Ireland

The Ordnance Survey Ireland (OSI)[13] is the national mapping agency of the Republic of Ireland and it provides multiple products and datasets. The authors of [7] transformed the geospatial data about the boundaries of the administrative areas of Ireland into RDF[14]. For the extension of the geospatial information of entities that belong in the Republic of Ireland, we consider the following datasets (i.e., administrative areas):

- City and County Council
- County Council
- City Council
- Municipal District
- Barony
- Parish
- Townland
- Rural Area

The datasets also contain the English and Irish name and the type (i.e., administrative division) of every unit.

[11] https://www.nidirect.gov.uk/campaigns/ordnance-survey-of-northern-ireland
[12] http://osni-spatial-ni.opendata.arcgis.com/
[13] https://www.osi.ie/
[14] http://data.geohive.ie/downloadAndQuery.html

3.4 Summary

Here, we give detailed information about the data sources that were used for the extension of YAGO. We used information from well-known projects (i.e., GADM and OSM) as well as data about administrative units from official sources, like Kallikratis and Ordnance Survey. In the next chapter we show how these data sources are used in order to extend YAGO.

4. EXTENSION OF YAGO

In this chapter, we present the methodology that was followed in order to extend the YAGO knowledge graph with geospatial information and also the results of this work. We are focusing on entities of YAGO that have a spatial dimension (i.e., they have a longitude and a latitude).

4.1 Matching Phase

The main goal of this work is to extend the YAGO knowledge graph with qualitative geo-spatial information without duplicating existing knowledge. To ensure that, we try to match entities of YAGO with entities of the data sources that were mentioned in the previous chapter. For example, the resource *geoentity_Hellenic_Republic_390903* and the entity with identifier *GRC* represent Greece in YAGO and GADM respectively and therefore should be matched. The matching phase consists of two filters: *(i) the label similarity filter* and *(ii) the geometry distance filter*.

The first filter of the matching phase is the *label similarity filter*. It produces matches between the geo-entities of YAGO and the entities of the specified data source (e.g., GADM) that have similar names. For this purpose we use the Levenshtein distance [17]. In order for two resources to be matched, the similarity between their labels must be higher than a specific threshold, which is set at 0.8. We examine every label of each entity, without considering its language tag [28]. In this stage of the matching phase, an entity of YAGO 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 false matches of the latter. Since there are many entities that share the same name (e.g., Athens, Greece and *Athens, Alabama*), the *geometry distance filter* is also a disambiguation step. The second filter checks if the Euclidean distance in the WGS:84 coordinate system between the geometry provided by GADM, OSM, or an official data source and the point provided by YAGO is smaller than a specific threshold, which is set at 0.2 degrees. In case there are multiple entities of YAGO that are matched with the same resource, we keep the entity that is closest, in terms of distance, to that resource. Our methodology is shown in Algorithm 1.

The number of the produced matches is very large and consequently it is not possible to manually check if every match is correct. As a solution to this problem, we randomly select a subset of the matches and manually check if these matches are correct, by checking the label of the matched resources. This methodology has been used in [12, 28].

4.2 Results

In the following sections we will present the results of the matching phase between YAGO and the data sources that we presented in Chapter 3.

Algorithm 1 Matching Phase

```
Input: yago, dataSource
Output: geometryDistanceMatches
labelSimilarityMatches ← ∅
geomDistanceMatches ← ∅
/* Label Similarity Filter */
for each yagoEntity in yago do
    for each dsEntity in dataSource do
        /* LabelSimilarity checks every label of both entities */
        /* LabelSimilarity uses Levenshtein Distance */
        if LabelSimilarity(yagoEntity, dsEntity) ≥ labelSimilarityThreshold then
            if yagoEntity not in labelSimilarityMatches then
                labelSimilarityMatches(yagoEntity) ← [dsEntity]
            else
                labelSimilarityMatches(yagoEntity).add(dsEntity)
            end if
        end if
    end for
end for
/* Geometry Distance Filter */
for each yagoEntity in labelSimilarityMatches do
    bestDistance ← ∞
    bestEntity ← ∅
    for each dsEntity in labelSimilarityMatches(yagoEntity) do
        currDistance ← GeomDistance(yagoEntity, dsEntity)
        if currDistance < bestDistance then
            bestDistance ← currDistance
            bestEntity ← dsEntity
        end if
    end for
    if bestDistance ≤ geomDistanceThreshold then
        geomDistanceMatches(yagoEntity) ← bestEntity
    end if
end for
```

4.2.1 YAGO and GADM

In this section we present the results of the matching phase between YAGO and GADM. In order to extend the YAGO knowledge graph with the information provided by GADM, we consider the following classes of YAGO:

- geoclass_independent_political_entity
- geoclass_semi-independent_political_entity
- geoclass_dependent_political_entity
- geoclass_first-order_administrative_division
- geoclass_second-order_administrative_division
- geoclass_third-order_administrative_division
- geoclass_fourth-order_administrative_division
- geoclass_fifth-order_administrative_division

The first three classes represent countries in YAGO. We split the information of YAGO and GADM into levels and we apply the matching phase on the following (YAGO, GADM) pairs:

1. (geoclass_independent_political_entity & semi-independent_political_entity & dependent_political_entity, 0-level)
2. (geoclass_first-order_administrative_division, 1-level)
3. (geoclass_second-order_administrative_division, 2-level)
4. (geoclass_third-order_administrative_division, 3-level)
5. (geoclass_fourth-order_administrative_division, 4-level)
6. (geoclass_fifth-order_administrative_division, 5-level)

The results of the matching phase between YAGO and GADM (Table 1) show that the precision of the generated matches, at all administrative levels, is really high. When it comes to the quantity of the matches, we observe that at higher administrative levels the number of matches is close to the number of entities that exist in YAGO. At lower levels the percentage of the entities that were matched drops. Figure 2 shows the extension of a matched entity, whereas Figure 3 shows a new entity that is created from an unmatched resource of GADM.

After examining both YAGO 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. Let us consider the example of Greece with witch we are very familiar. Neither YAGO nor GADM have any information in their administative levels about *municipal units* and *municipal communities* of Greece. Regarding the *regional units* of Greece, GADM does not provide any information about them, whereas YAGO contains some of

```
@base <http://yago-knowledge.org/resource/> .
@prefix extr: <http://kr.di.uoa.gr/yago-extension/resource/> .
@prefix exto: <http://kr.di.uoa.gr/yago-extension/ontology/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
<geoentity_Dimos_Athens_8133876> geo:hasGeometry extr:Geometry_GRC.3.1.12_1 .
<geoentity_Dimos_Athens_8133876> exto:hasGADM_UpperLevelUnit "GRC.3.1_1" .
<geoentity_Dimos_Athens_8133876> exto:hasGADM_NationalLevel "4thOrder" .
<geoentity_Dimos_Athens_8133876> exto:hasGADM_Name "Αθηναίων" .
<geoentity_Dimos_Athens_8133876> exto:hasGADM_Name "Athens"@en .
<geoentity_Dimos_Athens_8133876> rdf:type exto:Municipality .
<geoentity_Dimos_Athens_8133876> rdf:type exto:Dimos .
<geoentity_Dimos_Athens_8133876> exto:hasGADM_ID "GRC.3.1.12_1" .
extr:Geometry_GRC.3.1.12_1 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 2: YAGO and GADM, extension of a matched entity.

```
extr:gadmentity_Kyritz_DEU.4.12.3_1 geo:hasGeometry extr:Geometry_DEU.4.12.3_1 .
extr:gadmentity_Kyritz_DEU.4.12.3_1 exto:hasGADM_UpperLevelUnit "DEU.4.12_1" .
extr:gadmentity_Kyritz_DEU.4.12.3_1 exto:hasGADM_NationalLevel "4thOrder" .
extr:gadmentity_Kyritz_DEU.4.12.3_1 exto:hasGADM_Name "Kyritz"@en .
extr:gadmentity_Kyritz_DEU.4.12.3_1 exto:hasGADM_Type "Municipality" .
extr:gadmentity_Kyritz_DEU.4.12.3_1 exto:hasGADM_Type "AmtsfreieGemeinde" .
extr:gadmentity_Kyritz_DEU.4.12.3_1 exto:hasGADM_ID "DEU.4.12.3_1" .
extr:Geometry_DEU.4.12.3_1 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 3: YAGO and GADM, a new entity.

them in its second level. Moreover, the Greek entities that are instances of geoclass_first-order_administrative_division of YAGO, are found in the second administrative level of GADM. Although Greek perfactures are no longer considered administrative divisions in Kallikratis, they can be found in the second level of YAGO, which means that the knowledge graph of YAGO contains outdated information. We also closely examined the information provided by YAGO and GADM about German administrative units. We observed that the units, that belong in the third administrative level of YAGO, 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.

4.2.2 YAGO and OpenStreetMap

OpenStreetMap has geospatial information for many types of features, such as cities, lakes, bars, restaurants, etc. For our purposes we focus on entities that have a permanent location. The majority of these entities are features of nature (e.g, water bodies, forests, etc.), but we also take into consideration other classes as well, such as cities and archaeological sites. The feature classes, that we used in order to extend YAGO, are shown in Table 2. Like the case of GADM, we want to apply the matching phase on pairs of features classes of OSM and classes of YAGO. For that reason, we had to find the classes of YAGO that correspond to the features classes of OSM that we are interested in. These pairs are also shown in Table 2, along with the results of the matching phase. In the same table we can observe that in some cases YAGO has more instances than the datasets of Geofabrik (e.g., islands, localities, etc.), but there are also cases where Geofabrik provides

Table 1: The results of the matching phase between YAGO and GADM.

YAGO (#entities)	GADM (#entities)	Matches	Correct Matches
countries (233)	0-level (256)	221	50/50
first-order_administrative_division (3958)	1-level (3610)	3086	200/200
second-order_administrative_division (44554)	2-level (45958)	28162	200/200
third-order_administrative_division (121648)	3-level (144608)	42556	199/200
fourth-order_administrative_division (124729)	4-level (137983)	45693	199/200
fifth-order_administrative_division (51112)	5-level (51427)	9	6/9

The number of entities in each administrative level of YAGO and GADM are shown on the first and second columns respectively. The number of total matches and correct matches for each pair are shown on columns three and four.

more information than YAGO (e.g., nature reserves, forests, etc.).

Regarding the results of the matching phase, Table 2 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 issue. Firstly, as we already mentioned in Section 3.2, OpenStreetMap is a crowdsourced project, which means that it contains noisy data. Such data ultimately do not contribute to our cause. Secondly, the labels 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 YAGO provides the names of many entities 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 YAGO knowledge graph and the results show that, regardless of the class and the number of produced matches, the quality is always high.

Since OpenStreetMap contains noisy data we chose to only extend matched entities and not bring unmatched entities to the knowledge graph. An example is shown in Figure 4.

<geoentity_Sánta_Margaríta_11552842> geo:hasGeometry extr:Geometry_265943516 .
<geoentity_Sánta_Margaríta_11552842> exto:hasOSM_Name "Σάντα Μαργαρίτα" .
<geoentity_Sánta_Margaríta_11552842> exto:hasOSM_FClass "beach" .
<geoentity_Sánta_Margaríta_11552842> exto:hasOSM_ID "265943516"^^xsd:integer .
extr:Geometry_265943516 geo:asWKT "POLYGON((...))" .

Figure 4: YAGO and OpenStreetMap, extension of a matched entity.

Table 2: The results of the matching phase between YAGO and OSM.

OSM Feature Class	YAGO Class	Matches	Correct Matches
archaeological (13333)	archaeological/prehistoric_site (4303)	287	100/100
park (339529)	park (87196)	31286	100/100
beach (24055)	beach & beaches (13157)	2634	99/100
canal (82341)	canal & canalized_stream & canal_tunnel & section_of_canal & navigation_canal (53729)	2787	97/100
stream (1768807)	stream & intermittent_stream & streams & section_of_stream (1078510)	127885	99/100
forest (105809)	forest & forest_reserve (42936)	2847	98/100
island (36474)	island & islands & section_of_island & land_tied_island (161448)	21255	99/100
spring (628)	spring (78354)	23	23/23
nature_reserve (64425)	nature_reserve (1731)	184	96/100
meadow (29364)	meadow (521)	3	3/3
water & reservoir (456600)	lake & crater_lake & section_of_lake & lakes & lagoon & intermittent_lake & oxbow_lake & reservoir (373438)	150016	98/100
wetland (21307)	wetland & intermittent_wetland (3887)	20	20/20
locality (71320)	locality & populated_locality (316234)	3755	95/100
city & town & village (93163)	populated_place & section_of_populated_place (4354916)	67389	98/100
region (445)	region & economic_region & historical_region & lake_region (2192)	23	23/23

The first column contains the feature classes of OSM and the second column contains the corresponding classes of YAGO. The results of the matching phase are shown in the third and fourth columns. The number of entities of each class is shown in the parenthesis.

4.2.3 YAGO and Kallikratis

Similarly to the case of GADM (Section 4.2.1), in this section, we try to find matches between the official administrative divisions of Greece provided by the Kallikratis dataset and classes of YAGO. The administrative levels of Greece are already mentioned in Section 3.3.1. During this process, we found out that the decentralized administrations and few regional units are instances of the class *geoclass_administrative_division*. Some regional units are also found in the second administrative level of YAGO. The Kallikratis dataset does not provide information about the boundaries of municipal units and communities, most of which are categorized as *populated places* and *localities* in YAGO. In order to have correct matches we use the geometry of the municipality, to which each munipality unit belongs. The results of the matching phase between YAGO and Kallikratis are shown in Table 3. This time we present the number of official administrative units that are matched.

The results show that our methodology was able to match the majority of the entities that exist in the higher administrative levels of the Kallikratis dataset. The class *geoclass_administrative_division* of YAGO contains 21 Greek regional units and we matched all of them. On the other hand the regional units that are found in the second class of YAGO did not pass the label similarity filter. Regarding the quality of the results, we evaluated every match and found out that all of them are correct. An extended entity is shown in Figure 5. An example of an unmatched entity of Kallikratis is shown in Figure 6.

Table 3: The results of the matching phase between YAGO and Kallikratis.

Kallikratis	YAGO	Matches
Decentralized Administrations	administrative_division	6/7
Regions	first-order_administrative_division	11/13
Regional Units	administrative_division & second-order	21/74
Municipalities	third-order_administrative_division	324/325
Municipal Units & Municipal Communties	populated_place & locality	377/1037

The first column contains the Greek administrative divisions and the second column contains the corepsonding YAGO classes. On the last column the number of administrative units, that are matched in each level, is shown.

```
<geoentity_Dimos_Athens_8133876> geo:hasGeometry extr:Geometry_9186 .
<geoentity_Dimos_Athens_8133876> exto:hasKallikratis_Name "ΔΗΜΟΣ ΑΘΗΝΑΙΩΝ" .
<geoentity_Dimos_Athens_8133876> exto:hasKallikratis_Population "657701"^^xsd:integer.
<geoentity_Dimos_Athens_8133876> exto:hasKallikratis_ID "9186"^^xsd:integer .
extr:Geometry_9186 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 5: YAGO and Kallikratis, extension of a matched entity.

```
extr:kallikratisentity_1 geo:hasGeometry extr:Geometry_1 .
extr:kallikratisentity_1 exto:hasKallikratis_Name "ΠΕΡΙΦΕΡΕΙΑ ΑΝ. ΜΑΚΕΔΟΝΙΑΣ ΘΡΑΚΗΣ" .
extr:kallikratisentity_1 exto:hasKallikratis_Population "599259"^^xsd:integer .
extr:kallikratisentity_1 exto:hasKallikratis_ID "1"^^xsd:integer .
extr:Geometry_1 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 6: YAGO and Kallikratis, a new entity.

4.2.4 YAGO, Ordnance Survey and Ordnance Survey Northern Ireland

In order to extend the entities of YAGO that belong to the United Kingdom with official geospatial information, we use data that is provided by Ordnance Survey and Ordnance Survey Northern Ireland. The countries of the UK are instances of the class *geoclass_first-order_administrative_division*. Counties, Metropolitan Districts, Unitary Authorities and the Greater London Authority are found in the second adiministrative level of YAGO. The third level of YAGO 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 YAGO, which also contains District Wards.

Table 4: The results of the matching phase between YAGO and the official datasets for the United Kingdom

Class	Number of Entities	Matches	Correct Matches
first-order	4	2	2/2
second-order	185	182	100/100
third-order	3852	3676	100/100
fourth-order	7717	7624	100/100
populated_place & locality & section_of_populated_place & populated_locality	17231	1799	94/100

The first column contains the classes of YAGO and the second column the number of entities that we were able to find in each class. The number of total and correct matches are shown in the last two columns.

The results (Table 4) show that we were able to match most of the entities of the United Kingdom that are found in the administrative levels of YAGO. We also included the classes *geoclass_populated_place*, *geoclass_locality*, *geoclass_section_of_populated_place* and *populated_locality* into the matching phase in order to match more entities of Ordnance Survey and Ordnance Survey Northern Ireland. We can also observe, that the quality of the produced matches across all classes of YAGO is really high. There are many entities of OS and OSNI that are not matched. Similarly to the cases of GADM and Kallikratis, we extended matched entities and introduced unmatched of OS and OSNI to YAGO. Entities that are part of the new knowledge graph are shown in Figures 7, 8.

```
<Oxfordshire> geo:hasGeometry extr:Geometry_8328 .
<Oxfordshire> exto:hasOS_Name "Oxfordshire County" .
<Oxfordshire> exto:hasOS_AreaCode "CTY" .
<Oxfordshire> exto:hasOS_Description "County" .
<Oxfordshire> exto:hasOS_ID "8328"^^xsd:integer .
extr:Geometry_8328 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 7: YAGO and OS, extension of a matched entity.

```
extr:osnientity_WOODSTOCK_N08000359 geo:hasGeometry extr:Geometry_N08000359 .
extr:osnientity_WOODSTOCK_N08000359 exto:hasOSNI_ID "N08000359" .
extr:osnientity_WOODSTOCK_N08000359 exto:hasOSNI_Name "WOODSTOCK" .
extr:osnientity_WOODSTOCK_N08000359 exto:hasOSNI_Type "Ward" .
extr:Geometry_N08000359 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 8: YAGO and OSNI, a new entity.

4.2.5 YAGO and Ordnance Survey Ireland

Even though the provinces of Ireland (i.e., Ulster, Connach, Leinster and Munster) are no longer considereded as administrative units, they can be found in the first administrative level of the knowledge graph of YAGO. Irish city and county councils, county councils and city councils are instances of the class *geoclass_second_order_administrative_division*. The rest administrative levels of YAGO do not contain any Irish entities. For this reason, like the case of the United Kingdom (Section 4.2.4), we try to match entities of OSI with *populated places* and *localities* of YAGO.

Table 5: The results of the matching phase between YAGO and Ordnance Survey Ireland

Classes	Number of Entities	Number of Matches	Correct Matches
first-order	4	0	-
second-order	31	31	31/31
populated_place & locality & section_of_populated_place & populated_locality	13175	7648	99/100

Since provinces are no longer administrative units, we have zero matches in the first administartive level of YAGO, but on the other hand we were able to match all councils. Regarding the rest administrative divisions of Ireland (Section 3.3.4), 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 7500. Table 5 shows the number of entities of YAGO that were matched.

```
<County_Roscommon> geo:hasGeometry extr:Geometry_AE19629149713A3E055000000000001 .
<County_Roscommon> exto:hasOSI_Name "ROSCOMMON COUNTY COUNCIL" .
<County_Roscommon> exto:hasOSI_Name "ROSCOMMON COUNTY COUNCIL"@en .
<County_Roscommon> exto:hasOSI_ID "AE19629149713A3E055000000000001" .
<County_Roscommon> exto:hasOSI_Type "County_Council" .
extr:Geometry_AE19629149713A3E055000000000001 geo:asWKT "MULTIPOLYGON(((...)))" .
```

Figure 9: YAGO and OSI, extension of a matched entity.

```
extr:osientity_AE19629ACEE13A3E055000000000001 geo:hasGeometry
   extr:Geometry_AE19629ACEE13A3E055000000000001 .
extr:osientity_AE19629ACEE13A3E055000000000001 exto:hasOSI_ID
   "AE19629ACEE13A3E055000000000001" .
extr:osientity_AE19629ACEE13A3E055000000000001 exto:hasOSI_Name "CORNAROYA" .
extr:osientity_AE19629ACEE13A3E055000000000001 exto:hasOSI_Name "CORNAROYA"@en .
extr:osientity_AE19629ACEE13A3E055000000000001 exto:hasOSI_Type "Townland" .
extr:Geometry_AE19629ACEE13A3E055000000000001 geo:asWKT "POLYGON((...))" .
```

Figure 10: YAGO and OSI, a new entity.

4.2.6 Wikipedia and GeoNames

The spatial information that already existis in YAGO, 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 entities of Yago that come from Wikipedia as well as the number of entities that come from GeoNames. The results are shown in the following histogram (Figure 11).

In the case of GADM we see that both Wikipedia and GeoNames have equal contribution to the produced matches. In the case of OpenStreetMap, over 90% of the matched entities come from GeoNames. Table 2 shows that the most matches come from water bodies, waterways, and different types of places. This means that Wikipedia does not contain enough information about these features. Consequently, information about these features was extracted from GeoNames with YAGO2. More specifically, the cities, villages and towns of OpenStreetMap are matched with the populated places of YAGO. Acheson et al. state in [1] that most features of GeoNames are populated places and that streams are one of the most common natural features. This also explains the results about OSI and OSNI, since the majority of their entities are matched with populated places. Last, we see that in the cases of Kallikratis and Ordnance Survey, that most extended entities come from Wikipedia. It seems that Wikipedia provides precise information about the administrative units that belong in higher administrative levels for both Greece and Great Britain.

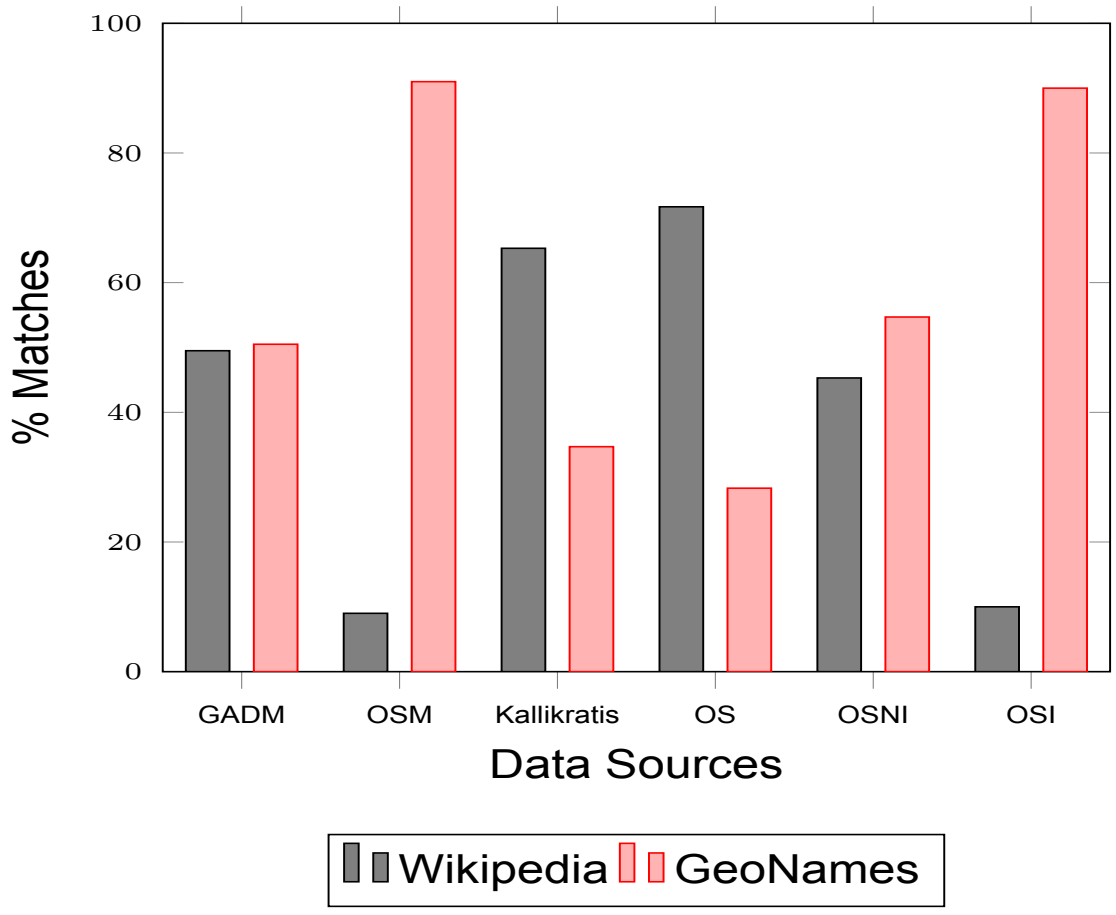

Figure 11: The comparison between Wikipedia and GeoNames

4.3 Layout of the extended knowledge graph

In this section we present the layout of the extended knowledge graph. The data of the extended knowledge graph is stored in N-TRIPLES files. For GADM, we have two files for each administrative level. The first file contains extended entities and the second one contains the new entities. For OpenStreetMap, we have generated a file for each feature

class that contains the extended entities of YAGO. For each official source, we have created two different files. Similalry to GADM, the first file contains the entities of YAGO that are extended, while the second file contains new entities that were created in order to represent the unmatched entities of the official data sources. Moreover, for each data source we provide a file that contains the links that were produced by the matching phase.

4.4 Summary

In this chapter we present the methodology that we followed in order to extend YAGO. We show in detail the results of matching phase between YAGO and each data source. Moreover, we give examples of entities of YAGO that are extended and also examples of new entities that we had to create in order to represent the entities of the data sources, that were not matched, in the extended knowledge graph. Last, we give the layout of the extended knowledge graph.

5. DEMONSTRATION OF THE EXTENDED KNOWLEDGE GRAPH

In this chapter we use the tools Strabon [15] and Sextant [21] to demonstrate the geospatial extension of YAGO. Strabon is an RDF triple store that supports both stSPARQL [14] and GeoSPARQL [24]. Sextant is a tool that visualizes geospatial linked data. In this work, we store the extended knowledge graph of YAGO in a Strabon endpoint. Afterwards, we issue queries to Sextant, that communicates with our endpoint, in order to visualize the results of the queries.

5.1 Municipality of Athens, GADM

In this example we demonstrate the information extracted from GADM about the municipality of Athens.

Listing 1: The SPARQL query that requests the geometry of the municipality of Athens.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX yago: <http://yago-knowledge.org/resource/>
PREFIX extr: <http://kr.di.uoa.gr/yago-extension/resource/>
PREFIX exto: <hhttp://kr.di.uoa.gr/yago-extension/ontology/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX strdf: <http://strdf.di.uoa.gr/ontology#>
SELECT ?athWKT
WHERE{
  yago:geoentity_Dimos_Athens_8133876 geo:hasGeometry ?athGeo .
  ?athGeo geo:asWKT ?athWKT .
}
```

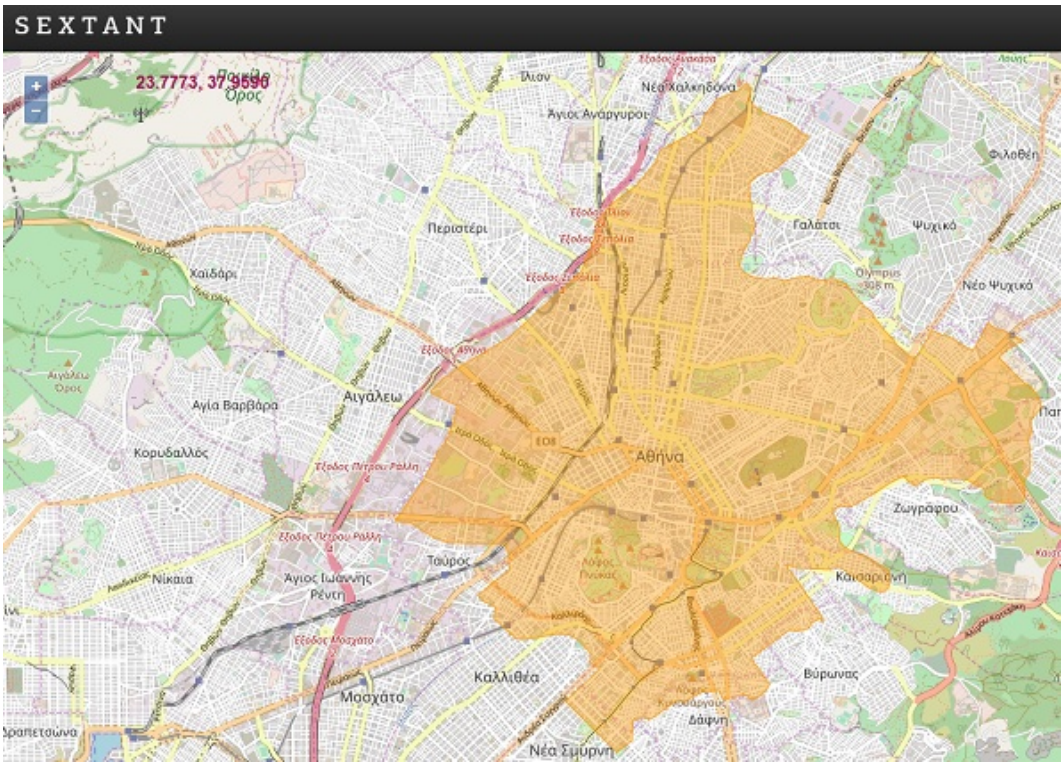

Figure 12: The municipality of Athens.

5.2 Unitary Authorities of Wales, GADM

This time we want to display the unitary authority of Wales that has the largest area. To achieve that, we are going to use additional information provided by GADM (i.e., type of administrative unit) and also stSPARQL, since it allows us to calculate the area of geometries. The unitary authority that is returned, is Powys.

Listing 2: The SPARQL query that requests the unitary authority of Wales, that has the largest area.

```
SELECT ?res ?resWKT (strdf:area(?resWKT) AS ?area)
WHERE{
  yago:Wales geo:hasGeometry ?wGeom .
  ?wGeom geo:asWKT ?wWKT .
  ?res rdf:type exto:UnitaryAuthority .
  ?res geo:hasGeometry ?resGeom .
  ?resGeom geo:asWKT ?resWKT .
  FILTER(geof:sfWithin(?resWKT, ?wWKT))
}
ORDER BY DESC(?area)
LIMIT 1
```

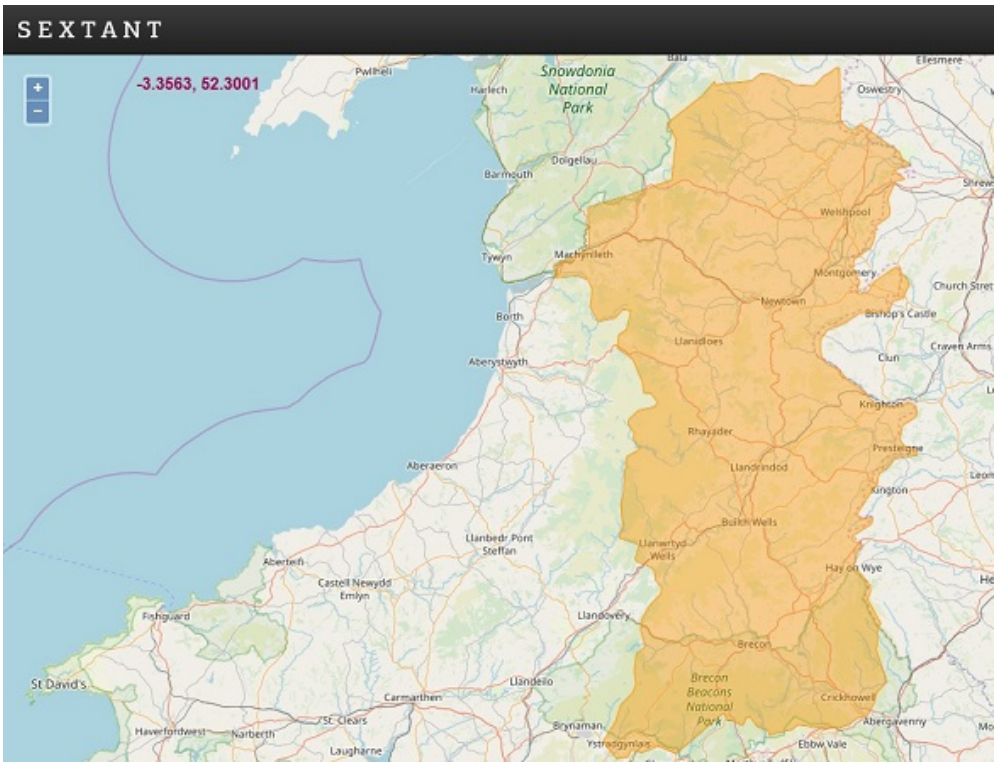

Figure 13: The largest unitary authority of Wales, Powys.

5.3 Forests, OpenStreetMap

In this section we want to visualize the largest forest in the knowledge graph. Geometries for the forests are extracted from OpenStreetMap, hence the result will be an extended entity. The forest we are looking for is the Hiawatha National Forest in Michigan.

Listing 3: The SPARQL query that requests the forest, that has the largest area.

```
SELECT ?res ?resWKT (strdf:area(?resWKT) AS ?area)
WHERE{
  ?res exto:hasOSM_FClass "forest" .
  ?res geo:hasGeometry ?resGeom .
  ?resGeom geo:asWKT ?resWKT .
}
ORDER BY DESC(?area)
LIMIT 1
```

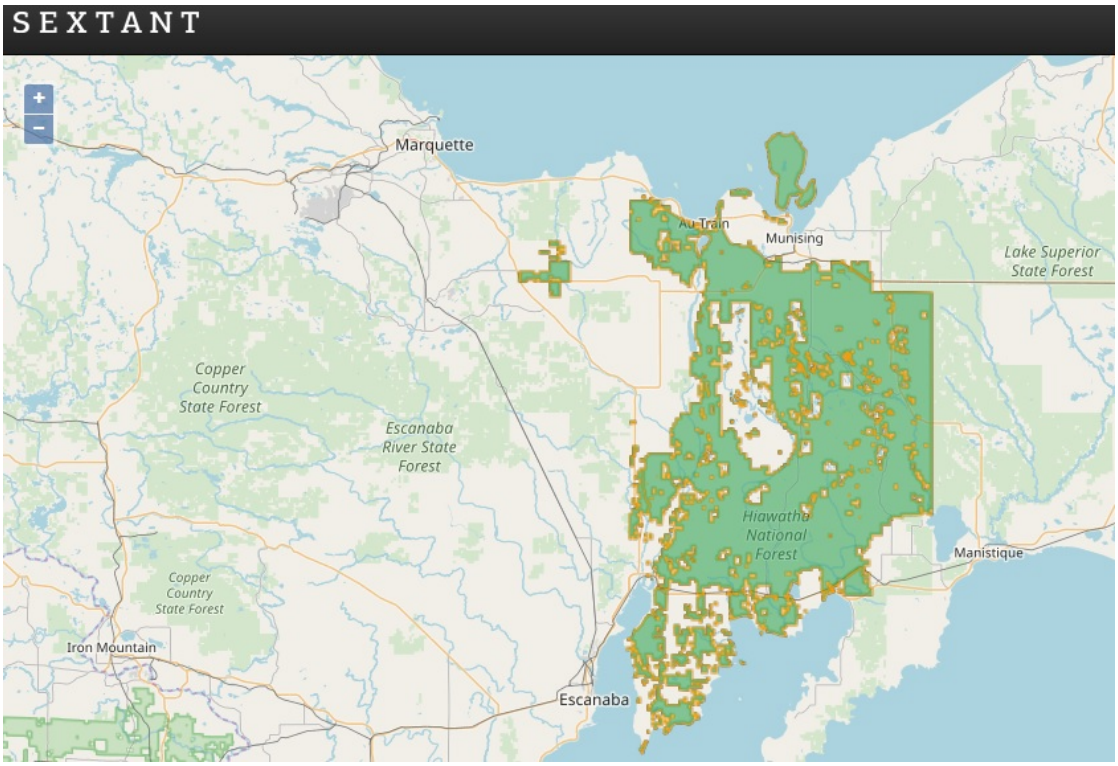

Figure 14: The largest forest in the extended knowledge graph, Hiawatha National Forest.

5.4 Water bodies and forests in Saarland, GADM and OpenStreetMap

Since we used multiple data sources to extend YAGO, we can combine their data. For example we can find the water bodies and forests that are within Saarland, a state of Germany. The geometries of water bodies and forests are provided by OSM, whereas the boundaries of Saarland by GADM. In the following query we get the water bodies in Saarland (Listing 5). By replacing *water* with *forest* we can obtain the forests. We write seperate queries in order to visualize water bodies and forests differently.

Listing 4: The SPARQL query that requests the water bodies in Saarland.

```
SELECT ?res ?resWKT
WHERE{
  yago:Saarland geo:hasGeometry ?sGeom .
  ?sGeom geo:asWKT ?sWKT .
  ?res exto:hasOSM_FClass "water" .
  ?res geo:hasGeometry ?resGeom .
  ?resGeom geo:asWKT ?resWKT .
  FILTER(geof:sfWithin(?resWKT, ?sWKT))
}
```

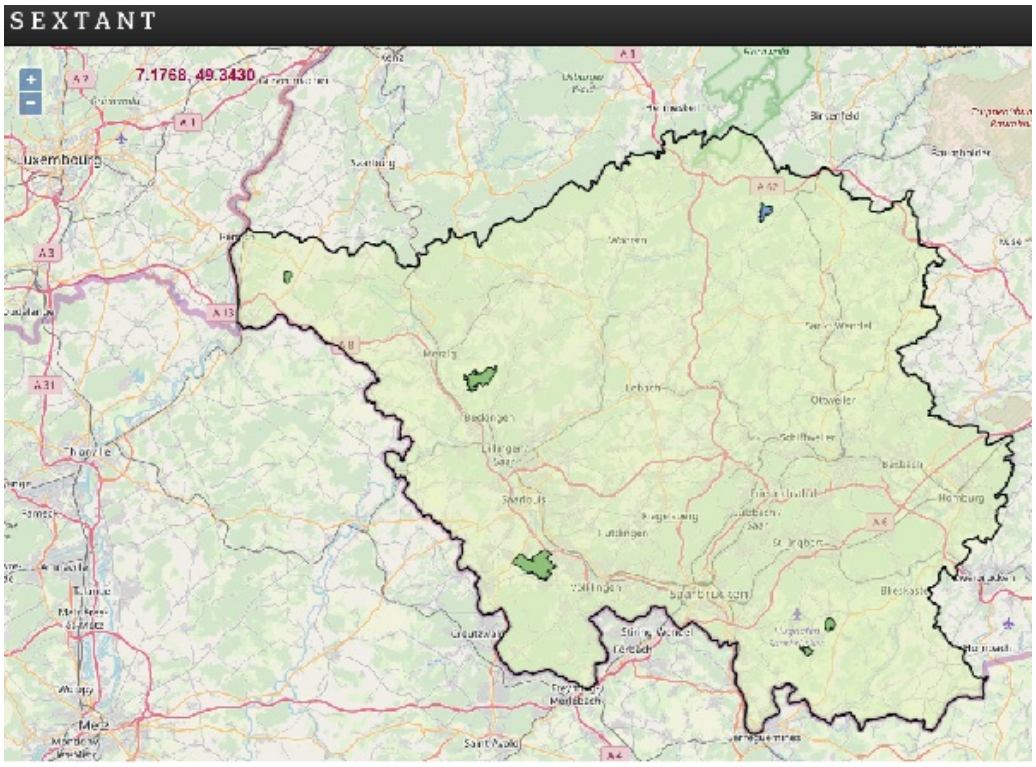

Figure 15: Water bodies and forests in Saarland.

5.5 Districts and district wards, Ordnance Survey

In this section we will use the description of each entity (i.e., administrative division they belong to) provided by Ordnance Survey to retrieve three districts that contain the most district wards.

Listing 5: The SPARQL query that requests the districts with the most district wards in the United Kingdom.

```
SELECT ?dis ?disWKT (COUNT(?dw) AS ?n_dw)
WHERE{
  ?dis exto:hasOS_Description "District" .
  ?dis geo:hasGeometry ?disGeom .
  ?disGeom geo:asWKT ?disWKT .
  ?dw exto:hasOS_Description "District Ward" .
  ?dw geo:hasGeometry ?dwGeom .
  ?dwGeom geo:asWKT ?dwWKT .
  FILTER(geof:sfWithin(?dwWKT, ?disWKT))
}
GROUP BY ?dis ?disWKT
ORDER BY DESC(?n_dw)
LIMIT 3
```

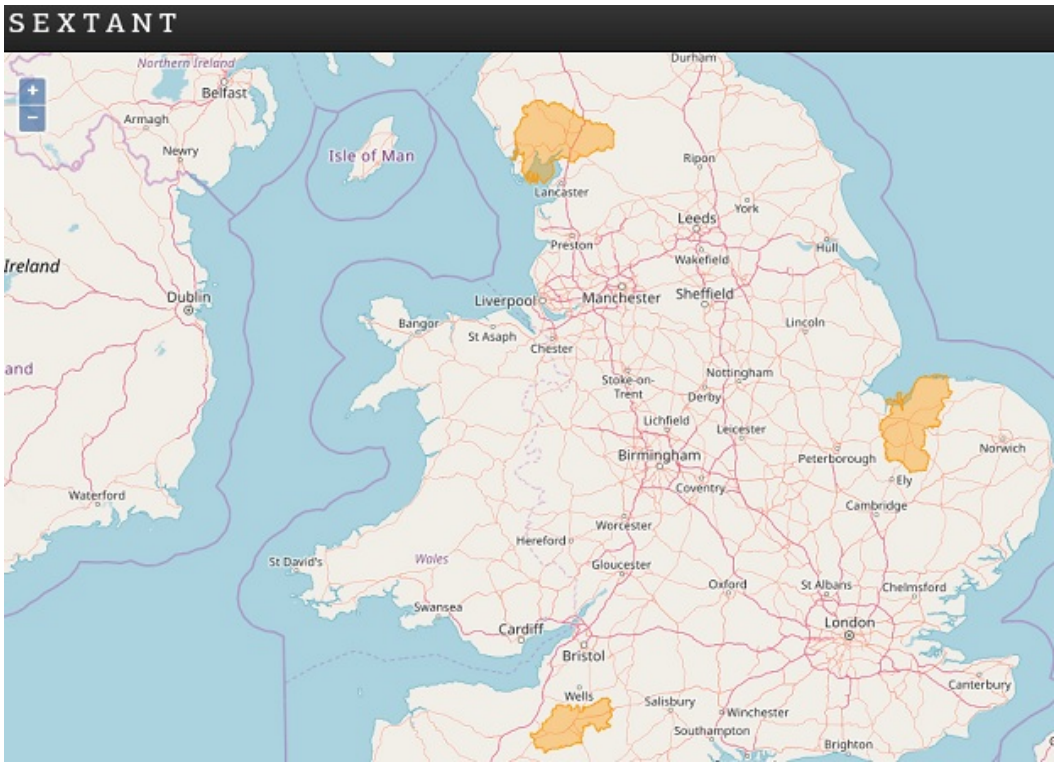

Figure 16: The districts with the most district wards in the United Kingdom.

5.6 Comparison between GADM and Kallikratis

The extended knowledge graph contains geospatial infomration about administrative units from multiple sources. Our extension can be used in order to compare the information provided by GADM and the official datasets, like Kallikratis. In this example we will use Listing 1 once more. This time *geoentity_Dimos_Athens_8133876* has also the information extracted from Kallikratis. The darker area in the following figure is the area where the geometries overlap with eachother.

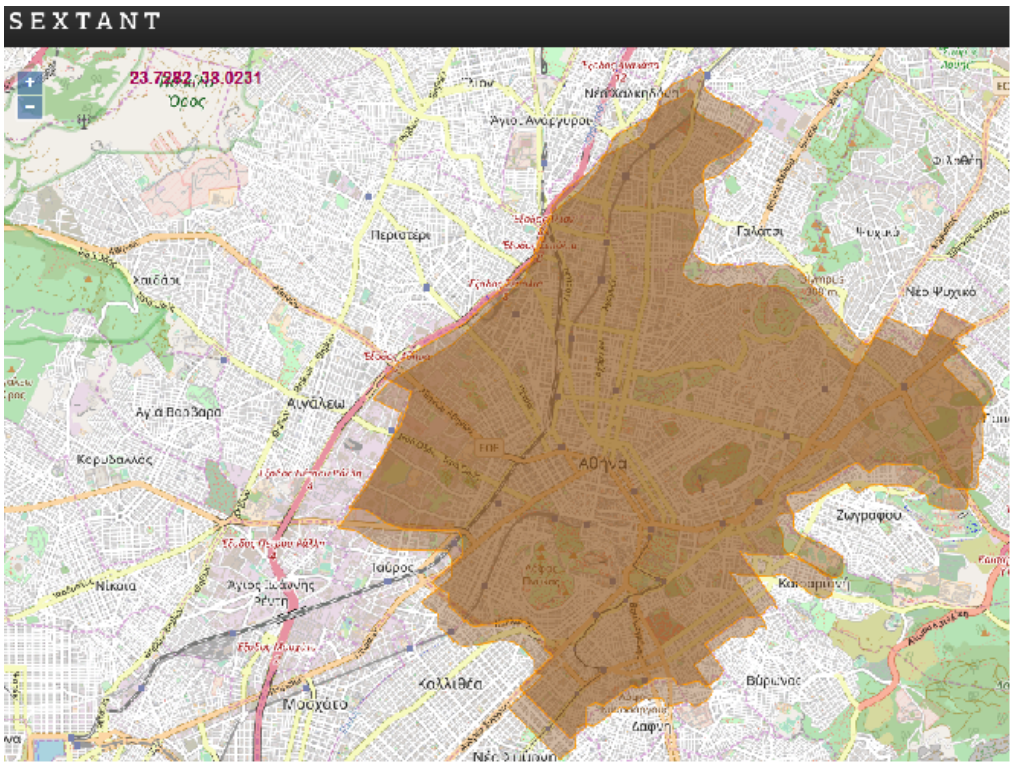

Figure 17: Comparison of the geometries provided by GADM and Kallikratis for the municipality of Athens.

5.7 Summary

In this chapter we demonstrate the extension of YAGO. We use the tools Strabon and Sextant in order to query and visualize the extended knowledge graph of YAGO. The extended knowledge graph can be used in order to combine information from multiple data sources (e.g., GADM and OSM) as well as to compare the information of administrative units that is retrieved from different sources (e.g., GADM and OS).

6. CONCLUSIONS AND FUTURE WORK

In this thesis we presented how we extended the YAGO knowledge graph with geospatial information. In order to achieve that we used multiple data sources. These data sources include the well known projects GADM and OpenStreetMap as well as datasets from official sources, like Ordnance Survey and Ordnance Survey Ireland. The results of this work show that our methodology produced matches between YAGO and each data source with almost perfect precision. The knowledge graph is extended with geometries that follow the standards of the Open Geospatial Consortium, which means that it can be queried using the query languages GeoSPARQL and stSPARQL.

For the purposes of this thesis, we extensively studied multiple data sources and we came to the following conclusions. The official datasets provide complete and up-to date information for the administrative divisions of the respective countries. In the data provided by YAGO and GADM we found inconsistencies about the administrative units. For instance, YAGO contains Greek administrative units that are no longer valid. Moreover, YAGO and GADM have administrative units of the United Kingdom, that according to the official datasets belong to different administrative divisions, in the same level. On the other hand, in YAGO, GADM and OpenStreetMap the information about each country follows the same schema making it easier to use. Since the official datasets of each country have different schemas, we had to treat each dataset differently. In addition, users of official datasets have to be familiar with the administrative divisions and language of the respective countries, in order to make proper use of their data.

Our future work will focus on extending YAGO with temporal information. Administrative divisions and units change over time. In order to capture these changes we plan to add temporal information about the time an administrative unit was created and the time it seized to be valid. The first use case is going to be Greece, since we are more experienced with its administrative divisions. Afterwards we plan to add such information about other countries as well. In this work we focused on specific feature classes of OpenStreetMap. In the future we plan to extend YAGO with information from more classes of OSM (e.g., mountains, forts, etc.). Apart from that, we also plan to use alternative sources of OpenStreetMap data (e.g., PlanetOSM). Moreover, we plan to create topological relations between the geometries of the administrative units that are part of the extended knowledge graph. Last but not least, part of our future work is the development of a question answering system over geographical knowledge graphs, that is going to be based on the work of Punjani et al. [25].

ABBREVIATIONS - ACRONYMS

RDF	Resource Description Framework
RDFS	RDF Schema
SPARQL	SPARQL Protocol and RDF Query Language
OSM	OpenStreetMap
OS	Ordnance Survey
OSNI	Ordnance Survey Northern Ireland
OSi	Ordnance Survey Ireland
RSS	Really Simple Sindication
UK	United Kingdom

REFERENCES

[1] Elise Acheson, Stefano De Sabbata, and Ross S. Purves. A quantitative analysis of global gazetteers: Patterns of coverage for common feature types. *Computers, Environment and Urban Systems*, 64:309–320, 2017.

[2] Dirk Ahlers. Assessment of the accuracy of geonames gazetteer data. In Christopher B. Jones and Ross Purves, editors, *Proceedings of the 7th Workshop on Geographic Information Retrieval, GIR 2013, 5th November, 2013, Orlando, Florida, USA*, pages 74–81. ACM, 2013.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary G. Ives. Dbpedia: A nucleus for a web of open data. In Karl Aberer, Key-Sun Choi, Natasha Fridman Noy, Dean Allemang, Kyung-Il Lee, Lyndon J. B. Nixon, Jennifer Golbeck, Peter Mika, Diana Maynard, Riichiro Mizoguchi, Guus Schreiber, and Philippe Cudré-Mauroux, editors, *The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11-15, 2007.*, volume 4825 of *Lecture Notes in Computer Science*, pages 722–735. Springer, 2007.

[4] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web. *Scientific american*, 284(5):34–43, 2001.

[5] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[6] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In Maria Fox and David Poole, editors, *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA, July 11-15, 2010*. AAAI Press, 2010.

[7] Christophe Debruyne, Alan Meehan, Éamonn Clinton, Lorraine McNerney, Atul Nautiyal, Peter Lavin, and Declan O'Sullivan. Ireland? s authoritative geospatial linked data. In *International Semantic Web Conference*, pages 66–74. Springer, 2017.

[8] Xin Dong, Evgeniy Gabrilovich, Geremy Heitz, Wilko Horn, Ni Lao, Kevin Murphy, Thomas Strohmann, Shaohua Sun, and Wei Zhang. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. In Sofus A. Macskassy, Claudia Perlich, Jure Leskovec, Wei Wang, and Rayid Ghani, editors, *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 601–610. ACM, 2014.

[9] Oren Etzioni, Michael J. Cafarella, Doug Downey, Ana-Maria Popescu, Tal Shaked, Stephen Soderland, Daniel S. Weld, and Alexander Yates. Unsupervised named-entity extraction from the web: An experimental study. *Artif. Intell.*, 165(1):91–134, 2005.

[10] Rolf Grütter, Ross S. Purves, and Lukas Wotruba. Evaluating topological queries in linked data using dbpedia and geonames in switzerland and scotland. *Trans. GIS*, 21(1):114–133, 2017.

[11] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. YAGO2: exploring and querying world knowledge in time, space, context, and many languages. In Sadagopan Srinivasan, Krithi Ramamritham, Arun Kumar, M. P. Ravindra, Elisa Bertino, and Ravi Kumar, editors, *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011 (Companion Volume)*, pages 229–232. ACM, 2011.

[12] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: A spatially and temporally enhanced knowledge base from wikipedia. *Artif. Intell.*, 194:28–61, 2013.

[13] Ehsan Kamalloo and Davood Rafiei. A coherent unsupervised model for toponym resolution. In Pierre-Antoine Champin, Fabien L. Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1287–1296. ACM, 2018.

[14] Manolis Koubarakis and Kostis Kyzirakos. Modeling and querying metadata in the semantic sensor web: The model strdf and the query language stsparql. In Lora Aroyo, Grigoris Antoniou, Eero Hyvönen, Annette ten Teije, Heiner Stuckenschmidt, Liliana Cabral, and Tania Tudorache, editors, *The Semantic Web: Research and Applications, 7th Extended Semantic Web Conference, ESWC 2010, Heraklion, Crete, Greece, May 30 - June 3, 2010, Proceedings, Part I*, volume 6088 of *Lecture Notes in Computer Science*, pages 425–439. Springer, 2010.

[15] Kostis Kyzirakos, Manos Karpathiotakis, and Manolis Koubarakis. Strabon: A semantic geospatial DBMS. In Philippe Cudré-Mauroux, Jeff Heflin, Evren Sirin, Tania Tudorache, Jérôme Euzenat, Man-

fred Hauswirth, Josiane Xavier Parreira, Jim Hendler, Guus Schreiber, Abraham Bernstein, and Eva Blomqvist, editors, *The Semantic Web - ISWC 2012 - 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I*, volume 7649 of *Lecture Notes in Computer Science*, pages 295–311. Springer, 2012.

[16] Douglas B Lenat. Cyc: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

[17] Vladimir I Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966.

[18] Farzaneh Mahdisoltani, Joanna Biega, and Fabian M. Suchanek. YAGO3: A knowledge base from multilingual wikipedias. In *CIDR 2015, Seventh Biennial Conference on Innovative Data Systems Research, Asilomar, CA, USA, January 4-7, 2015, Online Proceedings*. www.cidrdb.org, 2015.

[19] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41, 1995.

[20] Roberto Navigli and Simone Paolo Ponzetto. Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artif. Intell.*, 193:217–250, 2012.

[21] Charalampos Nikolaou, Kallirroi Dogani, Kostis Kyzirakos, and Manolis Koubarakis. Sextant: Browsing and mapping the ocean of linked geospatial data. In Philipp Cimiano, Miriam Fernández, Vanessa López, Stefan Schlobach, and Johanna Völker, editors, *The Semantic Web: ESWC 2013 Satellite Events - ESWC 2013 Satellite Events, Montpellier, France, May 26-30, 2013, Revised Selected Papers*, volume 7955 of *Lecture Notes in Computer Science*, pages 209–213. Springer, 2013.

[22] George Papadakis, Leonidas Tsekouras, Emmanouil Thanos, George Giannakopoulos, Themis Palpanas, and Manolis Koubarakis. Jedai: The force behind entity resolution. In Eva Blomqvist, Katja Hose, Heiko Paulheim, Agnieszka Lawrynowicz, Fabio Ciravegna, and Olaf Hartig, editors, *The Semantic Web: ESWC 2017 Satellite Events - ESWC 2017 Satellite Events, Portorož, Slovenia, May 28 - June 1, 2017, Revised Selected Papers*, volume 10577 of *Lecture Notes in Computer Science*, pages 161–166. Springer, 2017.

[23] George Papadakis, Leonidas Tsekouras, Emmanouil Thanos, George Giannakopoulos, Themis Palpanas, and Manolis Koubarakis. The return of jedai: End-to-end entity resolution for structured and semi-structured data. *PVLDB*, 11(12):1950–1953, 2018.

[24] Matthew Perry and John Herring. Ogc geosparql-a geographic query language for rdf data. *OGC implementation standard*, 2012.

[25] D. Punjani, K. Singh, A. Both, M. Koubarakis, I. Angelidis, K. Bereta, T. Beris, D. Bilidas, T. Ioannidis, N. Karalis, C. Lange, D. Pantazi, C. Papaloukas, and G. Stamoulis. Template-based question answering over linked geospatial data. In *Proceedings of the 12th Workshop on Geographic Information Retrieval*, GIR'18, pages 7:1–7:10, New York, NY, USA, 2018. ACM.

[26] Hanan Samet, Jagan Sankaranarayanan, Michael D. Lieberman, Marco D. Adelfio, Brendan C. Fruin, Jack M. Lotkowski, Daniele Panozzo, Jon Sperling, and Benjamin E. Teitler. Reading news with maps by exploiting spatial synonyms. *Commun. ACM*, 57(10):64–77, 2014.

[27] Panayiotis Smeros and Manolis Koubarakis. Discovering spatial and temporal links among RDF data. In Sören Auer, Tim Berners-Lee, Christian Bizer, and Tom Heath, editors, *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with 25th International World Wide Web Conference (WWW 2016)*, volume 1593 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2016.

[28] Claus Stadler, Jens Lehmann, Konrad Höffner, and Sören Auer. Linkedgeodata: A core for a web of spatial open data. *Semantic Web*, 3(4):333–354, 2012.

[29] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: a core of semantic knowledge. In Carey L. Williamson, Mary Ellen Zurko, Peter F. Patel-Schneider, and Prashant J. Shenoy, editors, *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, May 8-12, 2007*, pages 697–706. ACM, 2007.

[30] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. YAGO: A large ontology from wikipedia and wordnet. *J. Web Sem.*, 6(3):203–217, 2008.

[31] Julius Volz, Christian Bizer, Martin Gaedke, and Georgi Kobilarov. Discovering and maintaining links on the web of data. In Abraham Bernstein, David R. Karger, Tom Heath, Lee Feigenbaum, Diana Maynard, Enrico Motta, and Krishnaprasad Thirunarayan, editors, *The Semantic Web - ISWC 2009, 8th International Semantic Web Conference, ISWC 2009, Chantilly, VA, USA, October 25-29, 2009. Proceedings*, volume 5823 of *Lecture Notes in Computer Science*, pages 650–665. Springer, 2009.

[32] Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Commun. ACM*, 57(10):78–85, September 2014.

[33] Mark Wick and Bernard Vatant. The geonames geographical database. *Available from World Wide Web: http://geonames.org*, 2012.

[34] Eman M. G. Younis, Christopher B. Jones, Vlad Tanasescu, and Alia I. Abdelmoty. Hybrid geo-spatial query methods on the semantic web with a spatially-enhanced index of dbpedia. In Ningchuan Xiao,

Mei-Po Kwan, Michael F. Goodchild, and Shashi Shekhar, editors, *Geographic Information Science - 7th International Conference, GIScience 2012, Columbus, OH, USA, September 18-21, 2012. Proceedings*, volume 7478 of *Lecture Notes in Computer Science*, pages 340–353. Springer, 2012.

[35] Ce Zhang, Christopher Ré, Michael J. Cafarella, Jaeho Shin, Feiran Wang, and Sen Wu. Deepdive: declarative knowledge base construction. *Commun. ACM*, 60(5):93–102, 2017.