GeoWordNet: a resource for geo-spatial applications

Fausto Giunchiglia, Vincenzo Maltese, Feroz Farazi, Biswanath Dutta

DISI - Università di Trento, Trento, Italy

Abstract. Geo-spatial ontologies provide knowledge about places in the world and spatial relations between them. They are fundamental in order to build semantic information retrieval systems and to achieve semantic interoperability in geo-spatial applications. In this paper we present GeoWordNet, a semantic resource we created from the full integration of GeoNames, other high quality resources and WordNet. The methodology we followed was largely automatic, with manual checks when needed. This allowed us accomplishing at the same time a never reached before accuracy level and a very satisfactory quantitative result, both in terms of concepts and geographical entities.

Keywords: Geo-spatial ontologies, WordNet

1 Introduction

As part of the effort to achieve semantic interoperability in the Web, there is a pressing need and growing interest in geo-spatial ontologies, aiming at the so called geo-spatial semantic Web [2, 3]. For geo-spatial ontology we mean an ontology including geo-spatial entities (optionally associated with some properties/metadata), geographic classes (also called features) and topological relations [17] (such as *part-of, overlaps, near*) between them. For instance, a geo-spatial ontology can provide the information that *Florence* (the entity) is a *city* (its class) in *Italy* (its ancestor) and, among other information, the corresponding latitude and longitude coordinates. In some contexts, tools which maintain this kind of information are also called semantic gazetteers (for instance in [12]) or semantic geo-catalogs [4].

Geo-spatial ontologies are of fundamental importance in many applications, such as (among others) semantic Geographic Information Systems [4, 5], semantic annotation (but also matching and discovery) of geo-spatial Web services [6, 7], geographic semantics-aware web mining [15] and Geographical Information Retrieval (GIR) [10, 13]. In particular, restricted to GIR, there are various competitions, for instance GeoCLEF[1], specifically for the evaluation of geographic search engines. In all such applications, ontologies are mainly used for word sense disambiguation [9], semantic (faceted) navigation [14], document indexing and query expansion [10, 13], but in general they can be used in all the contexts where semantic interoperability is an issue.

Unfortunately, the current geographical standards, for instance the specifications provided by the Open Geospatial Consortium (OGC)[2], do not represent an effective solution to the interoperability problem. In fact, they specifically aim at syntactic agreement [11]. For example, if it is decided that the standard term to denote a har-

[1] http://ir.shef.ac.uk/geoclef/

[2] http://www.opengeospatial.org/

bour (defined in WordNet as "*a sheltered port where ships can take on or discharge cargo*") is *harbour*, they will fail in applications where the same concept is denoted with *seaport.* Similarly, current gazetteers do not represent a satisfactory solution. In fact, they are no more than just yellow pages for place names and, consisting of ambiguous plain descriptions, they do not support logical inference [12]. As a response to this problem, some frameworks have been recently proposed to build and maintain geo-spatial ontologies [5, 14, 15], but to the best of our knowledge no comprehensive, sufficiently accurate and large enough ontologies are currently available.

WordNet[3], even if not specifically designed for this, is de facto used as knowledge base in many semantic applications (for instance in [18, 19, 20]). Unfortunately, its coverage of geographic information is very limited [10], especially if compared to geographic gazetteers that usually contain millions of place names. In addition, WordNet does not provide latitude and longitude coordinates as well as other relevant information which is of fundamental importance in geo-spatial applications.

To overcome these limitations, there have been some recent attempts to integrate WordNet with geographical resources. Angioni et al. [8] propose a semi-automatic technique to integrate terms (classes and instances) from GEMET. Volz et al. [9] created a new ontology from the integration of WordNet with a limited set of classes and corresponding instances from GNS and GNIS[4]. The same resources are used by Buscardi et al. [10] to enrich 2,012 WordNet synsets with latitude and longitude coordinates. Unfortunately, all the above mentioned approaches are very limited in the number of terms (classes and instances) covered and accuracy. In particular, the problem in accuracy is mainly due to the semi-automatic approaches used.

Our main contribution to this problem is the creation of the GeoWordNet semantic and linguistic resource obtained from the integration of GeoNames[5] with WordNet plus the Italian section of MultiWordNet[6]. The methodology we followed is largely automatic, with manual intervention for the critical parts, thus accomplishing at the same time a never reached before accuracy and a very satisfactory quantitative result. We first created a multilingual knowledge base in which we imported WordNet and MultiWordNet. Then, for each place in GeoNames we automatically extracted metadata such as latitude and longitude coordinates, altitude, alternative names (available in multiple languages) and the spatial relations between them and integrated them in the knowledge base. This was achieved by first identifying those classes in GeoNames for which there existed already a corresponding synset in WordNet and then by enriching WordNet (i.e. the knowledge base) with new synsets for the uncovered classes. The new synsets were then connected to the most appropriate synset through *hypernym* (is-a) or *part meronym* (pat-of) relations. Synsets for individual places were then automatically created as instances of the previously identified or created synsets. The last step consisted in the importing of corresponding metadata.

The rest of the paper is organized as follows. In Section 2 we briefly describe the overall process followed for the construction of GeoWordNet. Individual phases are extensively described in Sections 3-6. Some interesting critical issues faced during the

[3] http://wordnet.princeton.edu/wordnet

[4] http://earth-info.nga.mil/gns/html/index.html and http://geonames.usgs.gov respectively

[5] http://www.geonames.org

[6] http://multiwordnet.fbk.eu

process are presented in Section 7. Section 8 presents some final statistics. Section 9 concludes the paper and outlines future work.

2 Creating GeoWordNet

Being our main goal to improve the geo-spatial search experience of end users and to support semantic interoperability in geo-spatial applications, we enriched WordNet with a huge number of geo-spatial concepts, entities and relations between them. We posed particular attention not only to the quantity, but also to the quality of the information being integrated. Towards this goal we organized the process in four phases (see Fig. 1), described in the next four sections:

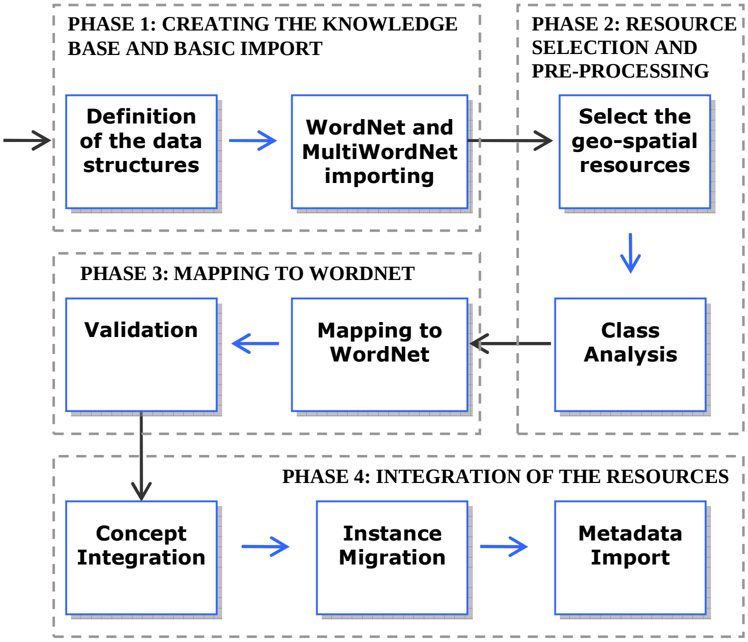
PHASE 1: CREATING THE KNOWLEDGE BASE AND BASIC IMPORT
Definition of the data structures
WordNet and MultiWordNet importing
PHASE 2: RESOURCE SELECTION AND PRE-PROCESSING
Select the geo-spatial resources
Class Analysis
PHASE 3: MAPPING TO WORDNET
Mapping to WordNet
Validation
PHASE 4: INTEGRATION OF THE RESOURCES
Concept Integration
Instance Migration
Metadata Import

Fig. 1. A global view of the phases of the GeoWordNet creation process

- ***PHASE 1: Creating the knowledge base and basic import.*** It consisted of the definition of some suitable data structures to store knowledge in multiple languages coming from different sources such as WordNet and MultiWordNet.
- ***PHASE 2: Resource selection and pre-processing.*** It consisted of the selection of the most appropriate resources of geo-spatial terms, analysis of the classes and entities contained, and creation of the corresponding concepts.
- ***PHASE 3: Mapping to WordNet.*** Concepts created with the previous step were mapped with those in WordNet. The mapping produced was manually validated.

- ***PHASE 4: Integration of the resources.*** It consisted of the full integration of the geo-spatial concepts with WordNet (including spatial relations between them), migration of the instances of such concepts (the places) and of the creation of the corresponding metadata (properties).

3 Creating the knowledge base and basic import

Our knowledge base is organized into four distinct parts:

- ***Linguistic part:*** it contains terms, synsets and lexical relations between them. This part is instantiated in multiple languages (e.g., English and Italian);
- ***Ontological part:*** it stores concepts and semantic hierarchical (e.g., *is-a*, *part-of*) and associative relations (e.g., *similar-to, cause-of*) between them. This section is language independent;
- ***Domain knowledge:*** concepts are organized into facet hierarchies [24] codifying knowledge about a specific domain. This section is also language independent;
- ***Entity part***: it contains the instances of the concepts contained in the ontological part and their attributes (possibly different according to their kind);

We initially populated the data structures with information taken from WordNet 2.1 and the Italian section of MultiWordNet. This is mainly motivated by the importance that the English and Italian languages have respectively in the context of the Living Knowledge[7] and the Live Memories[8] projects.

WordNet is a large lexical database for the English language, developed at the Cognitive Science Laboratory at Princeton University. WordNet groups words of different part of speech (nouns, verbs, adjectives and adverbs) into sets of cognitive synonyms, called synsets, each expressing a distinct concept. In other words, each synset groups all the words with same meaning or sense. Synsets are interlinked by means of conceptual-semantic and lexical relations. Typical semantic relations are *hypernym* (is-a) and *part meronym* (part-of). An example of lexical relation is *Participle of verb*. The structure of WordNet makes it a useful tool for computational linguistics and natural language processing and it is also frequently used in semantic applications. We imported all the words, synsets and lexical relations between them in the linguistic part of our knowledge base, instantiated for the English language. For each synset we then created a language independent concept in the ontological part. Semantic hierarchical and associative relations are codified at this level. We decided to do not import WordNet instances for two main reasons. First, they are not a significant number and no attributes are provided for them. In fact, the total number of entities in WordNet is 7671 [16]. Second, we plan to import huge quantities of entities and corresponding metadata from other resources, starting from GeoNames.

MultiWordNet is a multilingual lexical database including many languages such as Italian, Spanish, Romanian and Latin. The Italian part is strictly aligned with WordNet 1.6. Therefore, in order to align such information with those already imported by

[7] http://livingknowledge-project.eu
[8] http://www.livememories.org

WordNet 2.1, we first had to design an ad hoc procedure to map the two versions. This has been done by first using an already existing mapping[9] between WordNet 1.6 and 2.0 and then – using some heuristics - creating our own mapping between WordNet 2.0 and 2.1. Notice that for adjectives and adverbs we had to directly compute the mapping between WordNet 1.6 and 2.1 since not available elsewhere. We then instantiated the linguistic part of our knowledge base for the Italian language by importing words and synsets and - using the mapping – we connected each synset to the corresponding concept in the ontological part. Notice that due to the partial coverage of the language in MultiWordNet and the well known problem of gaps in languages (i.e. given a lexical unit in a language, it is not always possible to identify an equivalent lexical unit in another language) not all concepts have a corresponding synset in Italian. Detailed statistics are provided in Section 8.

4 Resource selection and pre-processing

Unfortunately, WordNet has quite limited coverage in geo-spatial information and lacks of latitude and longitude coordinates [10]. Therefore, it is essential to look elsewhere if we want an adequate amount of geographical information.

3.1 Selecting the geo-spatial resources

In order to enrich WordNet with the desired information, the first step in the process was the selection of one or more suitable sources of geo-spatial terms. In principle, there are various ways to collect such terms. For example, this can be done by extracting them from texts on the geo-spatial literature, by analysing the millions or billions of user queries stored in the query logs of existing search engines, by analyzing geo-spatial glossaries, or by selecting them from existing geo-spatial gazetteers. We chose the latter approach. In fact, geo-spatial gazetteers already contain high quality and huge quantities of readymade and usable names of geo-spatial classes (features or feature types) as well as corresponding instances (places), sometimes organized in hierarchies, thus providing also (spatial) relations between them. Last but not least, we especially looked to those providing latitude and longitude coordinates. On the basis of quantity and quality criteria, we evaluated several candidates including Wikipedia[10], YAGO [1], DBPedia[11], GEMET[12] and the ADL gazetteer[13], but they are limited either in locations, classes, relations or metadata. GeoNames and TGN, instead, both met our requirements:

- ***Thesaurus of Geographical Names (TGN)***[14]. TGN is a poly-hierarchical (i.e. multiple parents are allowed) structured vocabulary for place names. It also pro-

[9] http://www.cse.unt.edu/~rada/downloads.html#wordnet

[10] http://www.wikipedia.org/

[11] http://dbpedia.org/About

[12] http://www.eionet.europa.eu/gemet/about

[13] http://www.alexandria.ucsb.edu/gazetteer/

[14] http://www.getty.edu/research/conducting_research/vocabularies/tgn

vides alternative names, feature types and geographic (approximate) coordinates. It includes administrative political (e.g., cities, nations) and physical (e.g., mountains, rivers) entities. The temporal coverage of TGN ranges from prehistory to the present (some historical nations and empires are also included). It currently contains around 1.1 million names and 646 feature types, focusing on places particularly important for the study of art and architecture.

- ***GeoNames***. GeoNames is perhaps the most famous geo-spatial database. It includes geographical data such as place names in various languages, latitude, longitude, altitude and population collected from several data sources. Latitude and longitude coordinates are stored according to the WGS84 (World Geodetic System 1984) standard. It currently contains over 8 millions geographical names for around 7 millions unique places. At top level, the places are categorised into 9 broader categories (called feature classes), further divided into 663 sub-classes or features, most of them with a natural language description. In Table 1 they are given in detail. GeoNames provides an interface which allows users to manually edit and add new names. The data is available free of charge through a number of web services. The database is also available for free download under a creative commons attribution license.

Feature Class	**Description**	**Number of classes**
A	Administrative divisions of a country. It also represents states, regions, political entities and zones	16
H	Water bodies, e.g., ocean, sea, river, lake, stream, etc.	137
L	Parks, areas, etc.	49
P	Populated places, e.g., capitals, cities, towns, small towns, villages, etc.	11
R	Roads and railroads	23
S	Spots, buildings and farms	242
T	Mountains, hills, rocks, valleys, deserts, etc.	97
U	Undersea areas	71
V	Forests, heaths, vineyards, groves, etc.	17

Table 1. Feature classes and sub-classes in GeoNames

We used GeoNames as the main source. Being a thesaurus, TGN is instead used for consultation in order to better disambiguate GeoNames classes and relations.

3.2 Class analysis

This step is motivated by the following two objectives: (i) to make explicit the semantic of each class name, thus disambiguating each of them to a single concept, and (ii) to categorise and organise the semantically related concepts in a subsumption hierarchy. Notice in particular that relations in GeoNames are only implicitly provided (i.e. the kind is not explicitly mentioned). Relations between instances can be mainly mapped to a generic *part meronym* (part-of) relation, including administrative and physical containment. Relations between classes and instances can be mapped to *in-*

stance hyponym (instance-of) relation; no relations between classes are explicitly provided (i.e., the classes are provided in a flat list).

(i) We found that out of the 663 classes in GeoNames, for 57 of them no definition is provided at all. For these names we tried to understand the exact intended meaning, most of the time by considering the context of the term used, i.e. the corresponding feature class, and the instances (the places) associated to it. It was also observed that, even though the definitions are provided for the remaining terms, in some cases they are either ambiguous or not clear enough. Consider for instance the class *astronomical station*. GeoNames defines it as "*a point on the earth whose position has been determined by observations of celestial bodies*". Conversely, we decided that a more appropriate definition is "*a station from which celestial bodies and events can be observed*" and therefore we substituted it.

(ii) Once each of the 663 class names were refined and disambiguated to a single concept, following basic principles from Library Science we started categorising those semantically related concepts based upon their similar and dissimilar characteristics [22] and organised them in a hierarchical order. The result was a set of unconnected hierarchies. In choosing the characteristics, geo-spatial aspects were considered. Consider for instance the class *intermittent pond*. One may treat it as "a type of pond" and one may prefer to treat it as "a kind of intermittent thing". The former one is motivated by "geographical" feature. While, the latter one is motivated by its "temporal" aspect. Both views are correct from the classification point of view, but their correctness in a context is highly dependable on the purpose of the classification. In our case we chose the former one.

5 Mapping to WordNet

Concepts identified with the first phase were mapped - mainly manually with the help of some automatic discovery facilities - to WordNet synsets. We first tried to identify those concepts having an exact match with a synset in WordNet. At this purpose, it is clear that a syntactic match is not enough to judge about its existence. We rather worked at the conceptual level. For exact match at conceptual level we mean that a corresponding word for the class name exists in WordNet, and exactly one synset denotes the same meaning. For an easier identification of such synsets, we started from those concepts first which were more generic in nature according to the categorisation we did in the previous step. Consider for instance the following hierarchy:

- **valley** ("*a long depression in the surface of the land that usually contains a river*")
  - **ravine** ("*a deep narrow steep-sided valley (especially one formed by running water)")*
    - **canyon** ("*a ravine formed by a river in an area with little rainfall*")
    - **gorge** ("*a deep ravine (usually with a river running through it)*")
  - **hanging valley** ("*a valley the floor of which is notably higher than the valley or shore to which it leads; most common in areas that have been glaciated*")

We first looked in WordNet for a suitable synset for the concept *valley* and then we proceeded with the concept *ravine*, visiting the whole tree top-down. This order allowed restricting the search in WordNet to those synsets that are more specific than the previous one. In this way we found 306 exact correspondences with the GeoNames classes. In case of mismatch, we created a new synset in WordNet and identified the most appropriate synset denoting a more generic meaning for the class name. In other words we identified a suitable parent (according to the *hypernym* relation) for it. We faced several different situations and solved them accordingly. Due to space limitation, we present here only some remarkable examples:

- ***A more generic synset exists and no synset is available for the term***. Consider the class *palm grove,* defined in GeoNames as "*a planting of palm trees*". This concept is not available in WordNet, but the more generic synset for *grove* ("*garden consisting of a small cultivated wood without undergrowth*") is available. In this case we created a new synset for *palm grove* in WordNet and linked it with *grove* using a *hypernym* relation.

- ***A more generic synset exists but a synset is available for the term***. Consider the class *water tank.* GeoNames defines the term as "*a contained pool or tank of water at, below, or above ground level*", while WordNet defines it as "*a tank that holds the water used to flush a toilet*". WordNet does not provide any other sense for this term. It is clear that these two definitions are not equivalent. However, both definitions are more specific than *tank*, defined in WordNet as "*a large (usually metallic) vessel for holding gases or liquids*". In this situation we created a new sense for the term *water tank*. We positioned it as a sibling of the already existing one, by connecting it to *tank* using the *hypernym* relation.

- ***Linking synsets using the part meronym***. We occasionally considered appropriate to introduce some *part meronym* relations instead of the *hypernym* relation. For instance, an *icecap depression* (defined in GeoNames as "*a comparatively depressed area on an icecap*") is a part of an *icecap* (defined in GeoNames as "*a dome-shaped mass of glacial ice covering an area of mountain summits or other high lands; smaller than an ice street*") and not something more specific. A similar discourse can be done for *canal bend* and *section of canal* which are both parts of *canal.*

- ***Missing words in an existing synset***. It is interesting to note that in few cases we found that, even though the candidate term is not available, there is a synset denoting the same meaning in WordNet. In other words, the synset contains synonyms for the candidate term. It is clear that such cases are very difficult to detect just using automatic tools. One such example is the term *leprosarium.* This term is not available in WordNet, but there is a synset for the equivalent term *lazaret*. In these cases we added the GeoNames term to the corresponding WordNet synset. Another example is *metro station*, added in the synset for *subway station*.

- ***Multiple synset candidates***. The most subtle case is perhaps when the candidate concept has close match with multiple synsets in WordNet. This is due to the well known polysemy problem (see for instance [23]), namely very fine grained distinctions are provided. The solutions we adopted are described in Section 7.

To assess the quality of the mapping produced, a validation work was carried out by some experts in Library Science, particularly skilled in knowledge organization. The experts were different to those who were involved in the first phase of our work. This in order to assure that the validation work was not influenced by any unexpected external factor or bias. In order to carry out the validation work, the validators had to look at factors like the soundness of the description for the concepts (determined during the first phase), suitability of the selected synsets in WordNet, suitability of assigned names for the plural forms of concepts, and so on (see Section 7 for a list and corresponding description of the most interesting issues). In case of disagreement we iterated on the previous steps till all the conflicting cases were solved.

6 Integration of the resources

Once the mapping has been produced and validated, the next phase consisted in the integration of the two resources. This phase is fully automatic and consisted of the following three steps:

- ***Concept Integration***. We integrated GeoNames classes with WordNet (previously imported into the knowledge base). Here, by integration we mean the integration of the concepts built during the first phase (along with their description) which were not found in WordNet during the second phase, together with the *hypernym* and *part meronym* relations necessary to connect them to the existing concept network. For each new concept we created a corresponding English synset[15] by specifying the word, which is the name of the class, the gloss, which is the description of the class, and the part of speech, which is always noun. For the cases in which a synset already existed, but it did not contain the name of the class, we just added it to the list of words of the synset. For the classes having an exact match with WordNet, we just saved a reference to the existing concept/synset for future use (see next steps).

- ***Instance migration***. This step is about importing the locations contained in GeoNames into the knowledge base. Notice that in WordNet, the specific *instance hypernym* relation is used to link a synset denoting an entity to the synset denoting the corresponding class (or classes). We rather created a new object in the entity part of our knowledge base, clearly distinguishing between concepts and instances. We created a new object for each of the about 7 millions entities in GeoNames and related each of them to the concept of the corresponding class previously identified or created. We also created *part meronym* relations between such entities, according to the information provided in GeoNames. For instance, we codify the information that *Florence* is part of the *Tuscany* region in *Italy*.

[15] We will also create corresponding Italian synsets in the near future.

- ***Metadata Importing***. Locations in GeoNames are equipped with some metadata including the place name, alternative names in multiple languages (the specific languages can be identified), latitude, longitude, altitude and population. For instance, for the Italian city *Florence* the alternative names which are provided are *Firence*, *Firenze*, *Florencia*, *Florencija*, *Florens*, *Florenz*, *Florència*, *Flórens*; latitude is 43.7666667; longitude is 11.25; average altitude is 87 meters; population is 371,517 habitants. We attached all such information to the corresponding object (focusing on English and Italian names for the moment) created for the geographical entity in the entity part of the knowledge base.

7 Critical issues

This section describes the main issues we faced during the present work and the solutions we adopted for them. Due to the space limitation, only few of the issues, those considered particularly important and interesting, are described.

7.1 Facility: the service vs. function approach

The term *facility* is a key term in GeoNames. Being generic, a quite considerable amount of more specific classes are present in GeoNames. A mistake in the analysis of this term would have major consequences. In WordNet there are 5 different noun senses for the term, most of them focusing more on the notion of “service”, rather than on the notion of “function”:

- **facility**, installation (a building or place that provides a particular service or is used for a particular industry) *"the assembly plant is an enormous facility"*
- adeptness, adroitness, deftness, **facility**, quickness (skillful performance or ability without difficulty) *"his quick adeptness was a product of good design"; "he was famous for his facility as an archer"*
- **facility**, readiness (a natural effortlessness) *"they conversed with great facility"; "a happy readiness of conversation"--Jane Austen*
- **facility** (something designed and created to serve a particular function and to afford a particular convenience or service) *"catering facilities"; "toilet facilities"; "educational facilities"*
- **facility** (a service that an organization or a piece of equipment offers you) *"a cell phone with internet facility"*

On the other hand, the description of the term provided in GeoNames (“*a building or buildings housing a center, institute, foundation, hospital, prison, mission, courthouse, etc.*”) is rather generic and incomplete as includes only building or group of buildings. There are classes which are not buildings but still they can be treated as facilities, e.g., farms and parks. This is in line with the first sense in WordNet, where a facility can be a building or even a place. On one side many buildings provide services. Building housing banks usually provide transaction services; building housing hospitals usually provide health care services; building housing libraries usually provide access to the catalogue and book consultation. However, there are also buildings (or generic constructions) which do not provide any service, but are rather intended to have a function. For instance, houses are used for living purposes, while roads, streets and bridges have a transportation function (but no specific service is provided).

We decided to adhere to the WordNet vision and clearly distinguish between buildings and places providing a service (placed under the first sense) and those having just a (specific or generic) function (placed under the forth sense).

7.2 Plurals and Parenthesis

92 classes in GeoNames are present both in singular form, e.g., *populated place* and *vineyard*, and in plural form, e.g., *populated places* and *vineyards*. Furthermore, 99 classes are represented as a mixed singular-plural form, e.g., *arbour(s)*, *marsh(es)* and *distributary(-ies)*, sometimes in conjunction with the singular or plural form also.

From our analysis, singular forms represent single entities; plural forms indicate groups of entities; mixed forms are used when it is not easy to distinguish between the two previous cases. The approach we followed is to avoid plurals, identifying for each plural or mixed form a corresponding, more appropriate, name. For instance, we substituted *lakes* with *lake chain* and *mountains* with *mountain range*.

7.3 Dealing with polysemy

242 class names in GeoNames are polysemous, namely they have two or more similar, or related, meanings. It is not always easy to understand the correct meaning meant, especially in the cases in which no description is provided.

To find out the right concept, we compared the description, if available, of a class to each of the meanings of that class in WordNet. In some cases (15), we found out that a part of the description matches with one sense and another part of the description matches with another sense. Examples of such classes are *university*, *library* and *market*. During disambiguation such situations were overcome by comparing related terms in WordNet, for instance the ancestors, with the GeoNames feature class.

To be more concrete consider the following example for the term *university*. University is defined in GeoNames as: "*an institution for higher learning with teaching and research facilities constituting a graduate school and professional schools that award master's degrees and doctorates and an undergraduate division that awards bachelor's degrees*". It can be then summarized to be an institution for higher learning including teaching and research facilities that awards degrees. The term university has three meanings in WordNet:

- **university** (the body of faculty and students at a university)
- **university** (establishment where a seat of higher learning is housed, including administrative and living quarters as well as facilities for research and teaching)
- **university** (a large and diverse institution of higher learning created to educate for life and for a profession and to grant degrees)

The first meaning has little connection with GeoNames description and is excluded. The second meaning is relevant as it describes a university as an establishment for higher learning which also facilitates research and teaching. The third meaning is also relevant as it describes that it is a large institution of higher learning to educate for life and to grant degrees. To better disambiguate between the two remaining candidate meanings we then compared the hypernym hierarchy of the two synsets with the feature class provided for the term in GeoNames. The third meaning is a descendant of *social group*. The second meaning is a descendant of *construction*, which is closer to the feature class S (spots, building and farms). As a consequence, we finally selected the second meaning.

When such kind of analysis was not enough to disambiguate, we selected the instances from all close matched senses of WordNet and looked for their co-occurrence with the instances in GeoNames. In case of a match at instance level, we chose the corresponding sense. For example, consider the candidate term *palace*. GeoNames defines it as "*a large stately house, often a royal or presidential residence*". The first and forth senses for the term in WordNet look like possible candidates. They define it as "*a large and stately mansion*" and "*official residence of an exalted person (as a sovereign) correspond to it*" respectively. Following the proposed approach, we found that *Buckingham Palace* is the only instance in common with the first sense whereas no instances in common at all were found with the fourth sense. Therefore, we chose the first sense.

7.4 Unique name provision

In GeoNames, the same name is occasionally used to denote different concepts in different feature classes. This is particularly frequent for the classes under the feature class T - which denotes mountains, hills, rocks - and U - which denotes undersea entities. Some examples are *hill*, *mountain*, *levee* and *bench*. However, when feasible, it is always preferable to provide unique names to each semantically individual concept. And this is what we did, namely we identified a unique name to each concept. For the above examples, we distinguished between *hill* and *submarine hill*, between *mountain* and *seamount*, between *levee* and *submarine levee*, and between *bench* and *oceanic bench*. Such terms are not just arbitrarily assigned. They are rather collected from authentic literature available on Geography, Oceanography and Geology (e.g., Encyclopaedia Britannica[16]).

7.5 Physical vs Abstract entities

It is important to note that, since GeoNames always provides latitude and longitude coordinates for the entities, all of them must be seen as physical entities, that is having physical existence. However, when mapping the concepts from GeoNames to WordNet, we observed that for 27 of such concepts, WordNet only provides abstract senses, namely they are categorized as descendant of *abstract entity*. For example, for the concept *political entity* ("*a unit with political responsibilities*") WordNet provides a single synset at distance 6 from *abstract entity*. It is clear that, it would be incorrect to associate a geo-political entity, say *India*, under the abstract concept provided by

[16] http://www.britannica.com/

WordNet. In these cases we rather preferred to create a new synset in WordNet somewhere under *physical entity*. In the specific case, we created the new synset with the term *geo-political entity* defined as "*the geographical area controlled or managed by a political entity*" and connected it, through *hypernym* relation, to *physical object*.

8 Statistics

In this section we provide some interesting statistics regarding the imported resources as well as the constructed resource, GeoWordNet. In Table 2 we report statistical data about what we imported from WordNet 2.1 and the Italian MultiWordNet. WordNet was completely imported into the knowledge base. MultiWordNet, mainly due to the heuristics used to reconstruct the mapping with WordNet 2.1, was only partially imported. In particular, we imported all words, 88.4% of the senses and 86.3% of the synsets. We did not import the 318 (Italian) lexical and semantic relations provided.

WordNet 2.1	
Object	**Instances**
Synset	117,597
Relation	354,057
Word	147,252
Sense	207,019
Word exceptional form	4,728

MultiWordNet	
Object	**Instances**
Synset	33,156
Relation	-
Word	45,156
Sense	59,656
Word exceptional form	-

Table 2. Statistical data for WordNet 2.1 and MultiWordNet

Statistics about GeoNames (as from the version downloaded on 15th March 2009) are reported in Table 3. In particular, it shows the number of alternative names in multiple languages, names explicitly marked as preferred, and number of natural languages covered (those having an ISO 639 code).

GeoNames	
Object	**Instances**
Location	6,903,975
Alternative name	855,341
Preferred name	92,289
Natural language	230

Table 3. Statistical data for GeoNames

We analyzed the 663 GeoNames classes and their descriptions and compared them with those in WordNet. The result of our analysis is summarized in Table 4.

Table 5 shows the amount and kind of relations we created. Notice that for each relation we also created the corresponding inverse relations. Therefore, the provided numbers must be doubled (726 relations between classes, 13,807,950 relations between instances and classes, and 4,357,904 relations between instances).

GeoNames Classes	Instances	%
Which have a description in GeoNames	606	91.40
Which have no description in GeoNames	57	8.60
For which we provided or changed the description	92	13.88
For which we found a corresponding synset in WordNet	306	46.15
For which only one noun synset is available in WordNet	160	24.13
For which multiple noun synsets are available in WordNet	242	36.50
For which one part of the description matches with one synset and another part of the description matches with another synset	15	2.26
For which the description does not match with any of the synsets	38	5.73
For which we had to create a new synset in WordNet	357	53.84

Table 4. Main results of the GeoNames class analysis

Objects involved	Kind of relation	Quantity
Relations between classes	Hypernym	327
	Part meronym	36
Relations between instances and classes	Instance hypernym	6,903,975
Relations between instances	Part meronym	2,178,952

Table 5. Statistics about the number of relations created

9 Conclusions and future work

In this paper we presented GeoWordNet, a semantic and linguistic resource we created from the full integration of GeoNames with WordNet and the Italian portion of MultiWordNet. The methodology we followed is largely automatic, with manual intervention for the critical parts. This allowed obtaining a very satisfactory quantitative and qualitative result. By providing information about places in the world and proprieties like latitude and longitude coordinates, GeoWordNet supports interoperability in geo-spatial applications.

GeoWordNet is only the first step towards the creation of a huge and high quality knowledge base that we call the Universal Knowledge. The future work will mainly include the integration of other geo-spatial resources (like TGN) as well as concepts and instances from other domains (including people, organizations, events) and thus the instantiation of the domain part following the faceted approach (see for instance [24, 22]).

Acknowledgments: This work has been partially funded by the FP7 Living Knowledge FET IP European Project. Thanks to Gaia Trecarichi and Veronica Rizzi for the pleasant and fruitful discussions about geo-spatial issues. We also want to thank Ilya Zaihrayeu and Marco Marasca for their contribution to the definition of the data structures and Abdelhakim Freihat for the importing of the Italian MultiWordNet.

References

1. Suchanek, F. M., Kasneci, G., Weikum, G.: YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In Proc. of the 16th WWW, pp. 697-706 (2007).

2. Egenhofer, M. J.: Toward the Semantic GeoSpatial Web. In the 10th ACM Int. Symposium on Advances in Geographic Information Systems (ACM-GIS), pp. 1-4 (2002).
3. Kolas, D., Dean, M. , Hebeler, J.: Geospatial Semantic Web: architecture of ontologies. In Proc. of First Int. Conference on GeoSpatial Semantics (GeoS), pp. 183-194 (2005).
4. Shvaiko, P., Vaccari, L., Trecarichi G.: Semantic geo-catalog: a scenario and requirements. Poster at the 4th Ontology Matching Workshop (OM) at the ISWC (2009).
5. Abdelmoty, A.I., Smart, P., Jones, C.B.: Building Place Ontologies for the Semantic Web: issues and approaches. In Proc. of the 4th ACM workshop on GIR (2007).
6. Janowicz, K., Schade, S., Bröring, A., Keßler, C., Stasch, C., Maue', P., Diekhof, T.: A transparent Semantic Enablement Layer for the Geospatial Web. In the Terra Cognita Workshop at ISWC (2009).
7. Roman, D., Klien, E., Skogan, D.: SWING – A Semantic Web Service Framework for the Geospatial Domain. In the Terra Cognita Workshop (2006).
8. Angioni, M., Demontis, R., Tuveri, F.: Enriching WordNet to Index and Retrieve Semantic Information. In Proc. of 2nd Int. Conf. on Metadata and Semantics Research (2006).
9. Volz, R., Kleb, J., Mueller, W.: Towards ontology-based disambiguation of geographical identifiers. In Proc. of the 16th WWW Conference,(2007).
10. Buscardi, D., Rosso, P.: Geo-wordnet: Automatic Georeferencing of wordnet. In Proc. of the 5th Int. Conference on Language Resources and Evaluation (LREC) (2008).
11. Kuhl, W.: Geospatial semantics: Why, of What, and How? Journal of Data Semantics (JoDS) III, pp. 1–24 (2005)
12. Keßler, C., Janowicz, K., Bishr, M.: An agenda for the Next Generation Gazetteer: Geographic Information Contribution and Retrieval. In the Int. Conference on Advances in Geographic Information Systems (ACM SIGSPATIAL GIS) (2009).
13. Jones, C. B., Adbelmoty, A.I., Fu, G.: Maintaining Ontologies for Geographical Information Retrieval on the Web. In Proc. of On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE, Lecture Notes in Computer Science (2003).
14. Auer, S., Lehmann, J., Hellman, S.: LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In Proc. of the 8th World Int. Semantic Web Conference (ISWC) (2009).
15. Chaves, M. S., Silva, M. J., Martins, B.: A Geographic Knowledge Base for Semantic Web Applications. In Proc. of 20th Brazilian Symposium on Databases (SBBD) (2005).
16. Miller, G. A., Hristea, F.: WordNet Nouns: classes and instances. Computational Linguistics, 32(1):1.3 (2006)
17. Egenhofer, M. J., Dube, M. P.: Topological Relations from Metric Refinements. In Proc. of the 17th ACM SIGSPATIAL Int. Conference on Advances in GIS (2009).
18. Giunchiglia, F., Zaihrayeu, I.: Lightweight Ontologies. The Encyclopedia of Database Systems (2007)
19. Giunchiglia, F., Yatskevich, M., Shvaiko, P.: Semantic Matching: algorithms and implementation. Journal on Data Semantics, IX (2007).
20. Giunchiglia, F., Maltese, V., Autayeu, A.: Computing minimal mappings. In Proc. of the 4th Ontology Matching Workshop at the ISWC (2009).
21. Hill, L. L., Frew, J., Qi Zheng: Geographic names: the implementation of a gazetteer in a georeferenced digital library. D-Lib Magazine, 5 (1) (1999).
22. Ranganathan, S. R.: Prolegomena to library classification. Asia Publishing House (1967).
23. Mihalcea, R., Moldovan, D.I.: EZ.Wordnet: principles for automatic generation of a coarse grained WordNet. In Proc. of FLAIRS (2001).
24. Giunchiglia, F., Dutta, B., Maltese, V.: Faceted Lightweight Ontologies. In "Conceptual Modeling: Foundations and Applications", A. Borgida, V. Chaudhri, P. Giorgini, Eric Yu (Eds.) LNCS 5600 Springer (2009).