Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

DISI - Via Sommarive 14 - 38123 Povo - Trento (Italy)
http://www.disi.unitn.it

A FACET-BASED METHODOLOGY FOR THE CONSTRUCTION OF A LARGE-SCALE GEO-SPATIAL ONTOLOGY

Fausto Giunchiglia, Biswanath Dutta,
Vincenzo Maltese, Feroz Farazi

October 2011

Technical Report # DISI-11-479

A facet-based methodology for the construction of a large-scale geo-spatial ontology [1]

Fausto Giunchiglia, Biswanath Dutta, Vincenzo Maltese, Feroz Farazi

DISI - Università di Trento, Trento, Italy

Abstract. We concentrate on geo-spatial ontologies. Our main contribution in this paper is a methodology and a minimal set of guiding principles, inspired by the faceted approach, as originally developed in library science, and a large-scale ontology for *Space* that we have constructed following the methodology proposed. The approach we propose, centered on the fundamental notions of *domain* and *facet*, guarantees the creation of high quality ontologies in terms of robustness, extensibility, reusability, compactness and flexibility. Taking into account the different aspects of *Space*, the ontology we have developed, and that we have obtained from the refinement and extension of some existing resources including GeoNames, WordNet and the Italian part of MultiWordNet, provides knowledge about places of the world, their classes, their attributes and the spatial relations between them. The construction procedure was largely automatic, with manual intervention for the critical parts. This has allowed us to obtain a very satisfactory quantitative and qualitative result.

Keywords: Geo-spatial ontologies, faceted ontologies, methodology, ontology integration, geo-catalogues, Space

1 Introduction

As an essential support to geo-spatial applications, there is a pressing need and growing interest in geo-spatial ontologies [9, 10]. We consider *Space* in accordance with what people commonly understand by this term, which includes the surface of the earth, the space inside it and the space outside it. It comprises the usual geographical classes, often known as features, like land formations (continents, islands, countries), water formations (oceans, seas, streams) and physiographical classes (desert, prairie, mountain). It also comprises the areas occupied by a population cluster (city, town, village) and buildings or other man-made structures (school, bank, mine). Thus, for geo-spatial ontology we mean an ontology including geo-spatial entities, their classes, their attributes and relations (such as *part-of, overlaps, near-to*) 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 in the *part-of* hierarchy) and, among its attributes, the corresponding latitude and longitude coordinates. In some contexts, tools

[1] This paper is a substantially revised and extended version of two papers. The first was entitled "GeoWordNet: a resource for geo-spatial applications" and was presented at the ESWC 2010 conference [4]; the second was entitled "A facet-based methodology for geo-spatial modeling" and was presented at the GEOS 2011 conference [32]. The ontology presented in this paper is an extension of GeoWordNet, a semantic and linguistic resource distributed as open source that can be freely downloaded from http://geowordnet.semanticmatching.org/.

which maintain this kind of information are also called semantic gazetteers (for instance in [17]) or semantic geo-catalogues [11].

Applications requiring the use of geo-spatial ontologies include semantic Geographic Information Systems [11, 27], semantic annotation (but also matching and discovery) of geo-spatial Web services [12, 13], geographic semantics-aware web mining [20] and Geographical Information Retrieval (GIR) [16, 18]. In particular, restricted to GIR, there are various competitions, for instance GeoCLEF[2], specifically for the evaluation of geographic search engines. In all such applications, ontologies are mainly used for word sense disambiguation [15], semantic (faceted) navigation [19], document indexing and query expansion [16, 18], but in general they can be used in all the contexts where ontologies are needed to foster interoperability.

Unfortunately, the current geographical standards, for instance the specifications provided by the Open Geospatial Consortium (OGC)[3], do not represent an effective solution to the interoperability problem. In fact, they specifically aim at syntactic agreement [7]. For example, if it is decided that the standard term to denote a harbour (defined 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 a different term, e.g. with *seaport.* Similarly, gazetteers do not represent a satisfactory solution. In fact, they are no more than yellow pages for place names and, consisting of ambiguous plain descriptions, they do not support logical inference [17]. As a response to this problem, some frameworks have been recently proposed to build and maintain geo-spatial ontologies (see for instance [19, 20, 27]), but to the best of our knowledge no comprehensive, sufficiently accurate and large enough ontologies are currently available.

WordNet[4], even if not specifically designed for this, is *de facto* used as knowledge base in many semantic applications (for instance in [2]). Unfortunately, its coverage in terms of geographic information is very limited [16], especially if compared to geographic gazetteers that usually contain millions of place names as well as fine-grained distinctions between classes, such as GeoNames[5]. 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, some recent attempts have been developed with the goal to integrate WordNet with geographical resources. Angioni et al. [14] propose a semi-automatic technique to integrate terms (classes and instances) from GEMET. Volz et al. [15] created a new ontology from the integration of WordNet with a limited set of classes and corresponding instances from GNS and GNIS[6]. The same resources are used by Buscardi et al. [16] 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 covered and accuracy.

Our main contribution to this problem is a methodology and a minimal set of guiding principles, based on the *faceted approach*, as originally developed in library sci-

[2] http://ir.shef.ac.uk/geoclef/
[3] http://www.opengeospatial.org/
[4] http://wordnet.princeton.edu/wordnet
[5] http://www.geonames.org
[6] http://earth-info.nga.mil/gns/html/index.html and http://geonames.usgs.gov respectively

ence [3], and a very large and accurate geo-spatial faceted ontology that we call *Space*, obtained from the refinement and extension of GeoNames, WordNet and the Italian part of MultiWordNet[7]. *Space* accounts for the relevant classes, entities, their relations and attributes and, because constructed following the principles of the faceted approach, it is of very high quality in terms of robustness, extensibility, reusability, compactness and flexibility [3, 30, 31].

The rest of the paper is organized as follows. Section 2 describes our data model for *Space* and introduces some basic terminology. Section 3 presents the methodology and the principles that we followed for the creation of the *Space* ontology. Sections 4 to 7 illustrate and provide examples concerning the application of the single steps of the methodology. Section 8 summarizes some of the difficulties that we had to deal with. Section 9 provides some details about the *Space* ontology. Section 10 concludes the paper by summarizing the work done and outlying the future work.

2 The Data Model for *Space*

The first step towards the creation of the *Space* ontology was the definition of the corresponding data model. With this purpose, we follow and adapt the *faceted approach* [3] that was proposed by the Indian librarian Ranganathan at the beginning of the last century. The faceted approach, used for decades and with profit in library science to organize knowledge at the purpose of classifying bibliographic material on the shelves, is centred on the fundamental notions of *domain* and *facet*.

A *domain* can be defined as *any area of knowledge or field of study that we are interested in or that we are communicating about*. Domains may include traditional fields of study (e.g. medicine, physics), applications of pure disciplines (e.g. engineering, agriculture), any aggregate of such fields (e.g. physical sciences, social sciences) or capture knowledge about our everyday lives (e.g. music, sport, recipes, tourism). In this paper our focus is on the domain *Space*. Notice that Space has always played a central role in all library classification systems [34, 35].

The domain under examination is decomposed into its basic constituents, each of them denoting a different *aspect of meaning*. Each of these components is a *facet*. For instance, in *Space* the facets may include bodies of water, geological formations and administrative divisions. More precisely, a *facet* is a *hierarchy of homogeneous terms describing an aspect of the domain, where each term in the hierarchy denotes a different concept*. In the original library science approach, since the purpose is to classify bibliographic material, each concept denotes a set of documents while links between concepts in the facet hierarchies denote subset relations. In our approach, since the purpose is to describe *Space* in terms of real world objects, each concept may denote a class, an entity, a relation or an attribute, while links denote a much richer set of relations. For instance, in the former case the term *river* denotes the set of all documents about rivers, while in the latter case it denotes the set of all real world rivers. Concepts inside a facet are arranged by *characteristics*, i.e. according to their distinctive properties. For instance, since both *river* and *brook* are *flowing bodies of water* (their characteristic) they are arranged in the same facet, i.e. *body of water*, and at the same

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

level of the facet hierarchy. When arranged together, siblings sharing the same characteristic form what in jargon is called an *array* of homogeneous terms.

We define *Space* as follows:

$$Space = \langle C, E, R, A \rangle$$

where C is a set of classes, E is a set of entities, R is a set of binary relations and A is a set of attributes. These sets correspond to what in the faceted approach are called *fundamental categories*. More in detail:

- **C**: Elements in C denote classes of real world objects
- **E**: Elements in E represent the instances of the classes in C
- **R**: The set R provides structure to *Space* by relating entities and classes. It includes the canonical *is-a* (between classes in C), *instance-of* (associating instances in E to classes in C) and *part-of* (between classes in C or between entities in E) relations and is extended with additional relations according to the purpose, scope and subject of the ontology. For instance, we may include *near to* and *far from*. Since they constitute the backbone of the facet hierarchies, *is-a*, *part-of* and *value-of* relations are said to be *hierarchical*. We assume *is-a* and *part-of* to be transitive. Since other relations allow elements from different facets to be connected, they are said to be *associative*.
- **A**: Elements in A denote qualitative/quantitative and descriptive attributes of the entities. We further differentiate between attribute names and attribute values. Each attribute name in A denotes a relation associating each entity to corresponding attribute values. With this purpose, we also define a *value-of* relation that associates each attribute name to the corresponding set of possible values.

Within each fundamental category, we organize *Space* in three levels:

- ***Formal language level***: it provides the terms used to denote the elements in C/E/R/A. We call them *formal terms* to indicate the fact that they are language independent and that they have a precise meaning and role in (logical) semantics. Each term in C denotes a class (e.g. *lake, river* and *city*). Each term in E denotes an entity (e.g. *Garda lake*). Each term in R represents the name of a relation (e.g. *direction*). Each term in A denotes either an attribute name (e.g. *depth*) or an attribute value (e.g. *deep*). Elements in C, R and A are arranged into facets using *is-a*, *part-of* and *value-of* relations.
- ***Knowledge level***: it codifies what is known about the entities in E in terms of attributes (e.g. *Garda lake* is *deep*), the relations between them (e.g. *Garda lake* is part of *Trento*) and with corresponding classes (e.g. *Garda lake* is an instance of *lake*). Terms in E are at the leaves of the facets and populate them. The knowledge level is codified using the formal language described in the item above and is, therefore, also language independent;
- ***Natural language level***: we define a natural language as a set of words (i.e. strings), that we also call *natural language terms*, such that words with same

meaning within each natural language are grouped together and mapped to the same formal term. This level can be instantiated to multiple languages (at the moment only to English and Italian);

Similarly to WordNet and following same terminology, words are disambiguated by providing their meaning, also called *sense*. The meaning of each word can be partially described by associating it a natural language description. For instance, *stream* can be defined as *"a natural body of running water flowing on or under the earth"*. Within a language, words with same meaning (synonymy) are grouped into a *synset*. For instance, since *stream* and *watercourse* have the same meaning in English, they are part of the same synset. Given that a word can have multiple meanings (homonymy), the same word can correspond to different senses and therefore belong to different synsets. For instance, the word *bank* may mean *"sloping land (especially the slope beside a body of water)"*, *"a building in which the business of banking transacted"* or *"a financial institution that accepts deposits and channels the money into lending activities"*. In our data model, within a language each synset is associated a set of words (the synonyms), a natural language description, a part of speech (noun, adjective, verb or adverb) and a corresponding formal term.

In *Space* we clearly separate the elements of C/R/A that provide the basic terminology, from those in E that provide the instantiation of *Space*. The data model we propose has a direct formalization in Description Logic (DL) [36]. In fact, classes correspond to concepts, entities to instances, relations and attributes to roles. The formal language level provides the TBox, while the knowledge level provides the ABox for *Space*. They correspond to what in our previous work we called the *background knowledge* [23], i.e. the a-priori knowledge which must exist to make semantics effective. Each facet corresponds to what in logics is called *logical theory* [5, 6] and to what in computer science is called *ontology*, or more precisely *lightweight ontology* [24], and plays a fundamental role in task automation (formal reasoning). The natural language level provides instead an interface to humans and can be exploited for instance in Natural Language Processing (NLP).

Fig. 1 provides a small fragment of the *Space* ontology following the proposed data model, where classes are represented with circles, entities with squares, relation names with hexagons, attribute names with trapezoids and attribute values with stars. Letters inside the nodes (capital letters for entities and small letters for classes, relations and attributes) denote formal terms, while corresponding natural language terms are provided as labels of the nodes. For sake of simplicity, synonyms are not given. Arrows denote relations between the elements in C/E/R/A; solid arrows represent those relations constituting the backbone of the facets (*is-a*, *part-of* and *value-of* relations) and which are part of the formal language level; dashed arrows represent *instance-of*, *part-of* and the other relations (*depth* in this case) which are part of the knowledge level. Here the hierarchies rooted in *body of water* and *populated place* are facets of entity classes and are subdivisions of *location*, the one rooted in *direction* is a facet of relations and the one rooted in *depth* is a facet of attributes.

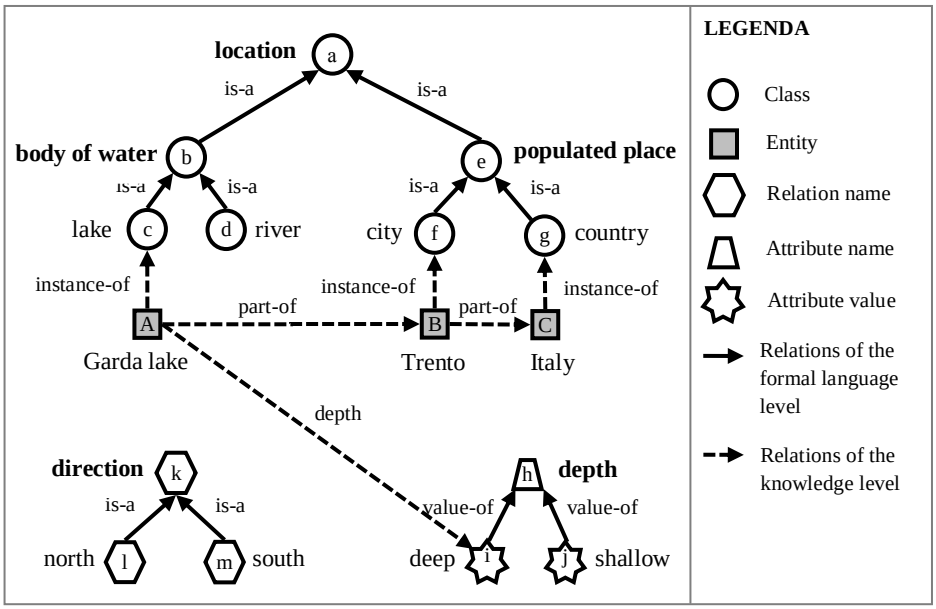

Fig. 1 - A small fragment of the *Space* ontology

3 The methodology

In this section we describe the main steps and the guiding principles that we follow for the construction of the *Space* ontology.

3.1 Steps in the process

The building process is organized in five subsequent phases: Identification of the terminology, Analysis, Synthesis, Standardization and Ordering. Let us describe them in turn.

Step 1: Identification of the terminology. It consists in collecting and classifying the natural language terms. In general, in the faceted approach this is mainly done by interviewing domain experts and by reading available literature about the domain under examination including *inter-alia* indexes, abstracts, glossaries, reference works. Analysis of query logs, when available, can be extremely valuable to determine user's interests. In our approach, each natural language term is analyzed and disambiguated by reconstructing the corresponding sense, by grouping those with same meaning into synsets, and by associating each synset to a formal term. Each formal term is then classified as a class, entity, relation or attribute (name or value).

Step 2: Analysis. The formal terms collected during the previous phase are analyzed per *genus et differentia*, i.e. in order to identify their commonalities and their differences. The main goal of the analysis is to identify as many characteristics as possible

of the real world entities represented by each of the terms. This allows being as fine grained as wanted in differentiating among them. For instance, for the term *river*, defined as *"a large natural stream of water (larger than a brook)"*, we can identify the following characteristics: a body of water; a flowing body of water; no fixed boundary; confined within a bed and stream banks; larger than a brook.

Step 3: Synthesis. With the synthesis, formal terms are arranged into facets. This is done by referring to their lexicalization in a language, i.e. to the corresponding English or Italian synsets and according to the characteristics identified with the previous phase. Following the principles described in the next section, the levels of the facet hierarchies are progressively formed by grouping terms into arrays by a common characteristic.

Step 4: Standardization. For each formal term in a facet, a standard (or preferred) term should be selected among the natural language terms associated to the corresponding synset. In the faceted approach this is usually done by identifying the term which is most commonly used in the domain and which minimizes the ambiguity. This is similar to the WordNet approach where words are ranked in the synset. The first word is the preferred one. For instance, the term *building* (defined as *"a structure that has a roof and walls and stands more or less"*) is more commonly used than the term *edifice*.

Step 5: Ordering. Formal terms in each array are ordered. There are many criteria one may follow, e.g., by chronological order, by spatial order, by increasing and decreasing quantity (for instance by size), by increasing complexity, by canonical order, by literary warrant and by alphabetical order. The sequencing criteria should be based upon the purpose, scope and subject of the ontology.

3.2 Guiding principles

We propose a minimal set of guiding principles for building facets:

1. ***Relevance***. The selection of the characteristics that are used to form the facets should reflect the purpose, scope and subject of the ontology. For example, while in the context of *Space* the characteristic *by populated cluster group* is appropriate to group villages, cities and towns, it is instead not suitable to classify state capitals, provincial capitals and national capitals. In fact, in the latter case the characteristic *by seat of government of a political entity* would be more realistic and appropriate. It is worthwhile also noting that the selection of the characteristics should be done carefully, as they cannot be changed unless there is a change in the purpose, scope and subject of the ontology.

2. ***Ascertainability***. Characteristics must be definite and verifiable. For example, the characteristic *flowing body of water* for rivers can be ascertained easily from the scientific literature and from the geo-scientists.

3. ***Permanence***. Each characteristic should reflect a permanent quality of an entity. For example, a *spring* (*"a natural flow of ground water"*) is always a flow-

ing body of water, thus the facet *flowing body of water* represents a permanent characteristic of *spring*.

4. ***Exhaustiveness***. Terms in each array should be totally exhaustive w.r.t. their respective common parent term in the facet hierarchy. For example, to classify the bodies of water based on the *water movement*, we need both *flowing body of water* and *stagnant body of water*. If we miss any of these two, the classification becomes incomplete.

5. ***Exclusiveness***. All the characteristics used to classify a term must be *mutually exclusive*, i.e. no two facets can overlap in content. For example, the bodies of water cannot be classified by both the characteristics *inland body of water* and *water movement*, as they would produce the same division for bodies of water such as lakes, rivers and ponds.

6. ***Context***. The position of a formal term in the ontology is a function of its meaning. This principle is particularly helpful to distinguish among homonyms. See for instance how we solve the ambiguity of the word *bank* in Section 8.

7. ***Currency***. The words chosen to denote formal terms should be those of current usage in the subject field. For example, in the context of transportation systems, *metro station* is more commonly used than *subway station.*

8. ***Reticence***. The words chosen to denote formal terms should not reflect any bias or prejudice (e.g. of gender, cultural, religious) or express any personal opinion of the person who develops the ontology. For example, it is not appropriate to use words like *devils places, criminal houses* to mean the jailhouses or any other type of correctional institutions.

9. ***Ordering***. The order of the facets and of the terms within each facet should reflect the purpose, scope and subject of the ontology. It should be applied consistently and should not be changed unless there is a change in the purpose, scope or subject of the ontology. Note that ordering carries semantics as it provides implicit relations between terms within an array. For example, the facet *populated place* may include *hamlet, village, town* and *city*. They are in ascending order according to population. This ordering clearly reflects that a *hamlet* is less populated than a *village*, that a *village* is less populated than a *town*, and so forth.

4 Identification of the terminology

The first step in the methodology consists in the selection of the resources that allow identifying the natural language terms representing the geo-spatial classes, the entities, the relations, the attributes and their disambiguation into formal terms. In the construction of *Space* this was done in four steps as follows:

- ***Step 2.1: Selection of the information sources.*** Possible sources of terminology were collected, evaluated in terms of quality and quantity of the information provided and the best candidates were selected. This step was manual.
- ***Step 2.2: Resource pre-processing.*** It consisted in (a) the extraction of the relevant natural language terms from each selected source, (b) the analysis and categorization of the terms into classes, entities, relations and attributes, (c) the disambiguation of the terms into senses, thus making explicit the meaning of each term and, in case of multiple terms with same meaning, grouping them into synsets. This step was manual, but in general it can be partially automated if the sources are sufficiently structured.
- ***Step 2.3: Mapping the resources.*** As preliminary step towards the integration, synsets identified with the previous step were mapped across sources. Among other things, this allowed duplicates to be identified. The mapping was manually produced and validated.
- ***Step 2.4: Integration of the resources.*** It consisted in using the mapping produced with the previous step to integrate the synsets extracted from the different sources. This step was fully automatic.

These steps are extensively described below.

4.1 Selection of the information sources

As main sources of natural language terminology, we selected WordNet 2.1 for the English and MultiWordNet[8] for the Italian language, respectively[9]. MultiWordNet is a multilingual lexical database including many languages such as Italian, Spanish, Romanian and Latin. Synsets for these languages are mapped to English synsets in WordNet.

Among the various sources of *Space* specific terminology, we particularly concentrated on geo-spatial gazetteers. In fact, these gazetteers contain huge quantities of locations and corresponding classes. They are sometimes organized in hierarchies, thus providing also relations between them, and offer attributes such as latitude and longitude coordinates. On the basis of quantity and quality criteria, we evaluated several candidates including Wikipedia[10], DBPedia[11], YAGO [8], GEMET[12] and the ADL gazetteer[13], but they are all limited in classes, entities, relations or attributes. GeoNames and TGN, instead, both met our requirements:

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

[9] These two resources were selected because of the importance that the English and Italian languages have respectively in the context of the Living Knowledge (http://livingknowledge-project.eu) and the Live Memories (http://www.livememories.org) projects we are involved in.

[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/

- ***Thesaurus of Geographical Names (TGN)***[14]. TGN is a poly-hierarchical (i.e. multiple parents are allowed) structured vocabulary containing 688 classes and around 1.1 million place names.
- ***GeoNames***. GeoNames provides 8 million place names in various languages amounting to 7 million unique places and corresponding attributes such as latitude, longitude, altitude and population. At the top level, the places are categorised into 9 broader categories, called feature classes, further divided into 663 classes, most of them associated with a natural language description. A special *null* class contains unclassified entities. In Table 1 they are given in detail.

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

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. Classes in GeoNames (version downloaded on March 2009)

Nevertheless, both TGN and GeoNames are pretty poor in relations. Since, understanding spatial relations is one of the fundamental features of Geographic Information Systems (GIS), we looked elsewhere for their identification. In particular, in producing our set of relations, we mainly followed the work by Arpinar et al. [26], Egenhofer and Herring [29], Egenhofer and Dupe [22] and Pullar and Egenhofer [25]. According to Egenhofer and Herring, spatial regions form a relational system comprising the relations between interiors, exteriors, and boundaries of two objects. Arpinar et al. suggest three major types of spatial relations: topological relations, cardinal direction and proximity relations. Egenhofer and Dupe propose topological and directional relations. According to them, topological relations have a leading role in qualitative spatial reasoning. Pullar and Egenhofer group spatial relations into direction relations (e.g. *north*, *northeast*), topological relations (e.g. *disjoint*), comparative or ordinal relations (e.g. *in*, *at*), distance relations (e.g. *far from*, *near to*) and fuzzy relations (e.g. *next to*, *close*). The spatial relations we propose include all these relations and some additional relations such as relative level (e.g. *above*, *below*), longitudinal (e.g. *in front*, *behind*), side-wise (e.g. *right*, *left*) and position in relation to border or

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

frontier (e.g. *adjacent*, *overlap*). In addition to spatial relations, we also consider some other kinds of relations, which can be treated as functional. For example, in the context of lakes, *primary inflow* and *primary outflow* are two important functional relations.

4.2 Resource pre-processing

With this step we extracted from GeoNames the natural language terms denoting the names of the classes, the names of the entities and the names of the attributes. Attribute values, being mostly quantitative, do not provide additional terminology. A part from the basic ones, the only relation explicitly provided in GeoNames is *neighbour* connecting each country with those of neighboring ones.

With the analysis, we mainly focused on the relations. In fact, since in GeoNames entities are neatly separated from classes with attributes directly associated to each entity, they could be easily identified. Conversely, with the only exception of *neighbour*, the kind of the relations is in general not explicitly provided. Relations between instances can be mapped to a generic *part-of* relation, including administrative and physical containment. The former connects administrative divisions, i.e. entities of classes such as country, province and district. The latter connects entities of classes such as lake, river and mountain to the corresponding administrative division. Relations between entities and classes correspond to *instance-of*. Since in GeoNames classes are provided in a flat list, no relations between classes are available.

With the disambiguation we created the senses by associating a natural language description (in English and Italian) to each natural language term found. Since we did not find cases of synonymy, each sense coincided with the synset.

Concerning the disambiguation of the classes, we found that out of the 663 classes in GeoNames, in 57 cases 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 associated to it. It was also observed that, even though 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.

Concerning the disambiguation of the entities, the names were directly extracted from the *name* and *alternative name* attributes in GeoNames, while the descriptions, in English and Italian, were automatically generated starting from the information provided by the *is-a* and *instance-of* relations. Several rules were used. For instance, one that we used for English is:

entity_name + " is " + *article* + " " + *class_name* + " in " + *parent_name* + "(" + *parent_class* + " in " + *country_name* + ")";

This allows for instance describing the *Garda Lake* as *"Garda Lake is a lake in Trento (Administrative division in Trentino Alto-Adige)"*.

The only relation found, *neighbour* was disambiguated as *"a nearby object of the same kind"*.

The disambiguation of the attributes led to the identification of 13 distinct attributes including *name*, *latitude*, *longitude* and *altitude*. Notice that we defined one single attribute representing *name* and *alternative name*, the latter codifying secondary names for the locations. In fact, we considered the value of the *name* attribute as standard term. These attributes are those provided for all the entities, while the other attributes are mainly provided for populated places and administrative divisions. The attributes extracted from GeoNames, with corresponding natural language description, are provided in Section 9.

4.3 Mapping the resources

As preliminary step towards their integration, we first mapped MultiWordNet to WordNet and then we mapped the synsets generated from GeoNames to those in WordNet. However, this was only done for the synsets of the classes, the attributes and the *neighbour* relation. In fact, the other relations in GeoNames correspond to the basic ones, while WordNet and MultiWordNet do not contain a significant number of entities. Note that the official number of entities in WordNet is 7671 [21], while we found out that 683 of them are common nouns instead. We identified the wrong ones by manually verifying those with no uppercased lemma. The wrong ones were converted into noun synsets, while the other 6988 were considered still entities.

The Italian part of MultiWordNet is strictly aligned with WordNet 1.6. Therefore, in order to align such information with WordNet 2.1, we first had to design an ad hoc procedure to map the two versions. This was done by first using an already existing mapping[15] between WordNet 1.6 and 2.0 and then by creating our own mapping between WordNet 2.0 and 2.1. This was achieved with the support of some heuristics, mainly based on the presence of same words, same part of speech and (almost) same gloss between the synsets. Notice that for adjectives and adverbs we had to directly compute the mapping between WordNet 1.6 and 2.1 since it was not available elsewhere. 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 English synsets have a corresponding synset in Italian.

In mapping GeoNames with WordNet, we distinguished the following cases:

- ***Case 1: there is an equivalent synset in WordNet.*** Two synsets were marked as equivalent if they denote the same meaning. We say that we have an *exact match* if the word in the GeoNames synset is also present in the WordNet synset. We say that there is a *partial match* if there is a corresponding synset in WordNet but the word in the GeoNames synset is not present in the WordNet synset. It is clear that the latter case is very difficult to detect with automatic tools. An example of the first case is *river*. An example of the second case is *leprosarium*. This term is not available in WordNet, but there is a synset for the equivalent term *lazaret*.

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

- ***Case 2: there is a more general synset in WordNet***. In case of mismatch, we looked for a more general synset according to the *is-a* (*hypernym*) relation. In this case the GeoNames synset was marked as more specific than the WordNet synset. Consider for instance the class *palm grove*, defined in GeoNames as "*a planting of palm trees*". There is no equivalent synset for it in WordNet, but the more general synset for *grove*, defined as "*garden consisting of a small cultivated wood without undergrowth*", is available in WordNet. In this case *palm grove* in GeoNames is marked as more specific than *grove* in WordNet.
- ***Case 3: there is a synset in WordNet that can be linked using part-of.*** We occasionally considered appropriate to associate synsets using the *part-of* (*part meronym*) relation instead of the *is-a* relation. In these cases, we explicitly marked the GeoNames synset as *part-of* the WordNet synset. 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*.

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 from those who were involved in the first phase of our work. This was done 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 natural language description for the senses determined during the first phase, suitability of the selected synsets in WordNet and suitability of assigned names for the plural forms. Section 8 provides 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. The result of our analysis is summarized in Table 2.

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 there is no equivalent synset in WordNet	357	53.84

Table 2. Main results of the GeoNames class analysis and their mapping to WordNet

4.4 Integration of the resources

Once the mapping was produced and validated, the next phase consisted in the integration of the three resources. This phase was fully automatic and consisted of the following steps[16]:

- ***Bootstrapping the knowledge base (KB).*** Following the data model described in Section 2, we created a KB allowing the storage of classes, entities, relations and attributes as well as their natural language lexicalization. We imported all the words and synsets from WordNet into the natural language level of our KB, instantiated for the English language. For each synset we then created a corresponding formal term in the formal language level. WordNet relations are also codified at this level. By importing words and synsets from MultiWordNet, we then instantiated the natural language level of our KB for the Italian language. Using the mapping between WordNet and MultiWordNet, we connected each Italian synset to the corresponding formal term in the formal language level.
- ***Concept Integration.*** By using the mapping between GeoNames and WordNet, we integrated GeoNames synsets with those in the KB. Here, by integration we mean the importing in the KB of the GeoNames synsets which do not have an exact or partial match with WordNet and are therefore not already present in the KB. For each missing synset, this was done by creating a corresponding English and Italian synset in the natural language level of the KB by specifying the word, the natural language description and the part of speech. We also created a corresponding formal term and the *is-a* or *part-of* relation necessary to connect it to the parent term. For the cases of partial match, we just added the missing word to the corresponding synset in the KB. For the cases of exact match, we just saved a reference to the synset in the KB for future use (see next step).
- ***Instance importing***. This step consisted in importing the entities contained in GeoNames into the KB. For each of the entities in GeoNames we created a new formal term denoting an entity in the knowledge part of our KB and, by means of instance-of relations, we related each of them to the formal term of the corresponding class previously created or identified as equivalent to an existing one. We also created *part-of* relations between such entities, according to the information provided in GeoNames. For instance, we codify the information that *Florence* is an instance of *city* and is part of the *Tuscany* region in *Italy*. Note that, the entities of the special *null* class were treated as instances of the generic class *location*.
- ***Attribute importing***. The attributes associated to each entity in GeoNames were imported as attributes and corresponding values (focusing on English and Italian names for the moment) in the knowledge part of our KB. This generated around 70 million attributes and corresponding values.

In Table 3 we report some statistics about the data we imported from WordNet and MultiWordNet. Excluding the 6988 entities and corresponding relations, WordNet

[16] GeoWordNet corresponds to the knowledge base obtained after these phases.

was completely imported. MultiWordNet, mainly due to the heuristics used to reconstruct the mapping with WordNet 2.1, was only partially imported. In particular, we imported 92.47% of the words, 94.28% of the senses and 94.30% of the synsets. We did not import the 318 (Italian) lexical and semantic relations provided.

WordNet 2.1	
Object	**Instances**
Synset	110,609
Relation	204,481
Word	147,252
Sense	192,620
Word exceptional form	4,728

MultiWordNet	
Object	**Instances**
Synset	36,448
Relation	-
Word	41,705
Sense	63,595
Word exceptional form	-

Table 3. Data imported from WordNet 2.1 and MultiWordNet

Table 4 shows the amount and kind of new relations that we created with the concept integration of GeoNames. Notice that for each relation we also created the corresponding inverse relations. Therefore, the actual number of relations is double the number shown in the table.

Objects involved	**Kind of relation**	**Quantity**
Relations between classes	is-a	327
	part-of	36
Relations between entities and classes	instance-of	6,907,417
Relations between entities	part-of	2,265,283

Table 4. Statistics about the number of relations created

5 Analysis

With the analysis, the terms collected and disambiguated during the previous phase were used as building blocks for the construction of the facets that constitute the *Space* ontology. For sake of simplicity, for the rest of the steps we focus only on the terms denoting classes.

The integration of the resources also helped us identifying the main sub-trees in our KB containing the necessary synsets representing geographical classes. In fact, with the integration, each of the synsets coming from GeoNames was hooked to one of the sub-trees rooted in:

- **location** - a point or extent in space
- **artifact, artefact** - a man-made object taken as a whole
- **body of water, water** - the part of the earth's surface covered with water (such as a river or lake or ocean); "they invaded our territorial waters"; "they were sitting by the water's edge"

- **geological formation, formation** - the geological features of the earth
- **land, ground, soil** - material in the top layer of the surface of the earth in which plants can grow (especially with reference to its quality or use); "the land had never been plowed"; "good agricultural soil"
- **land, dry land, earth, ground, solid ground, terra firma** - the solid part of the earth's surface; "the plane turned away from the sea and moved back over land"; "the earth shook for several minutes"; "he dropped the logs on the ground"

It is worthwhile to underline that not all the nodes in these sub-trees necessarily need to be part of *Space*. As a matter of fact, many of the descendants of *location* and *artifact* cannot be classified in our fundamental categories and therefore they were not included in *Space*. For instance, the following terms were discarded:

(Descendants of location)

- **there** - a location other than here; that place; "you can take it from there"
- **somewhere** - an indefinite or unknown location; "they moved to somewhere in Spain"
- **seat** - the location (metaphorically speaking) where something is based; "the brain is said to be the seat of reason"

(Descendants of artifact)

- **article** - one of a class of artifacts; "an article of clothing"
- **anachronism** - an artifact that belongs to another time
- **block** - a solid piece of something (usually having flat rectangular sides); "the pyramids were built with large stone blocks"

Terms denoting classes of real world entities were analyzed using their topological, geometric or geographical characteristics. We tried to be exhaustive in their determination. This leaves open the possibility to form a huge number of very fine grained groups. In order to illustrate the analysis process, consider the following list:

- **Mountain** - a land mass that projects well above its surroundings; higher than a hill
- **Hill** - a local and well-defined elevation of the land; "they loved to roam the hills of West Virginia"
- **Stream** - a natural body of running water flowing on or under the earth
- **River** - a large natural stream of water (larger than a brook); "the river was navigable for 50 miles"

Following the principles provided in Section 3.2, and in particular the *principle of relevance* and the *principle of ascertainability*, we can derive the following characteristics:

- **Mountain characteristics:**
  - the well-defined elevated land
  - formed by the geological formation (where geological formation is a natural phenomenon)
  - altitude in general >500m

- **Hill characteristics:**
  - the well-defined elevated land
  - formed by the geological formation, where geological formation is a natural phenomenon
  - altitude in general <500m

- **Stream characteristics:**
  - a body of water
  - a flowing body of water
  - no fixed boundary
  - confined within a bed and stream banks

- **River characteristics:**
  - a body of water
  - a flowing body of water
  - no fixed boundary
  - confined within a bed and stream banks
  - larger than a brook

6 Synthesis

Consider the list of characteristics selected with the analysis. The first characteristic of each of the terms above clearly suggests the distinction between two basic categories, the first consisting of *mountain* and *hill* and the second consisting of *stream* and *river*. Based upon those characteristics, two facets can be formed. They can be named *natural elevation* and *flowing body of water*, respectively. A further analysis of the characteristics suggested the creation of the more general facets *landform* and *body of water*, respectively.

The terms *mountain* and *hill* can be further differentiated *by size*. Note that, according to the *principle of relevance* and the *principle of permanence*, in this case size is a good distinguishing characteristic. In fact, it can be considered (almost) permanent in nature. Note that this is not true in general. For instance, it is not appropriate to distinguish animals by size because in this respect size is transitional in nature, i.e. their size rapidly changes over time. This is an example of what Aristotle called *accidental predicates* [28].

Note that *river* is a natural stream, and therefore a special kind of *stream*. In particular, this means that all the properties of stream are inherited by river (but not the vice versa). This is reflected in the facet hierarchy by putting *river* under *stream*. Based

upon the observations above we can build the following two facets, *body of water* and *landform*:

Body of water	**Landform**
Flowing body of water	Natural elevation
Stream	Mountain
River	Hill

An important property of facets is that they are *hospitable* (the interested reader can refer to [1] for a list of important properties of facets), i.e. they can be easily extended to accommodate additional terms as needed. Assume for instance that the new term *lake*, defined as *"a body of (usually fresh) water surrounded by land"*, is identified. By analysing it, we can derive the following characteristics:

- **Lake characteristics:**
  - a body of fresh water
  - fixed geographical boundary
  - a stagnant body of water

Going through the characteristics above, it should be quite easy to understand that *lake* cannot be put under the *flowing body of water*, even though it is a *body of water*. This implies that our classification is not good enough to classify all sorts of body of water, i.e. it is not exhaustive (*principle of exhaustiveness*). In order to include lakes, we need to extend the body of water facet with *stagnant body of water* in the same array of *flowing body of water*. This solves our problem.

In order to understand the importance of the *principle of exclusiveness*, assume to create in our classification the sub-classes *inland body of water*, *marine body of water*, *flowing body of water* and *stagnant body of water* and to put them in the same array under the main class *body of water*. Such categorization brings to confusion. In fact, lake can be now classified as both *inland body of water* and *stagnant body of water*. To avoid this confusion, the *principle of exclusiveness* plays an important role. According to this principle, all the characteristics used to classify a term must be mutually exclusive. So, we should not include all those four classes in the same array.

Similarly to lakes, we can extend the *natural elevation* facet in order to accommodate the term *valley* (defined as "*a long depression in the surface of the land that usually contains a river*"). Valley is a natural depression. So, in order to assign a place for *valley* inside this scheme, we have to create another sub-facet, namely, *natural depression*. Consider also that valleys are seen in both the oceanic areas (called *oceanic valleys*) and continental areas (called *valleys*). There is in general symmetry of real world entities in the continental and oceanic areas. For most of the continental entity classes there is a corresponding oceanic entity class with similar features but different name. So, in order to correctly classify the entities based upon the characteristic of their location, i.e. oceanic or continental, we should create the sub-facets *oceanic* and *continental* under the *natural elevation* and *natural depression* respectively as shown below. These additional facets make the classification of *landforms* exhaustive. See the appendix for an extended version of the *body of water* facet.

Body of water
- Flowing body of water
  - Stream
    - Brook
    - River
- Stagnant body of water
  - Pond
  - Lake

Landform
- Natural depression
  - Oceanic depression
    - Oceanic valley
    - Oceanic trough
  - Continental depression
    - Trough
    - Valley
- Natural elevation
  - Oceanic elevation
    - Seamount
    - Submarine hill
  - Continental elevation
    - Hill
    - Mountain

7 Standardization and ordering

Specifying different words for the same notion allows supporting semantic interoperability between systems using different terminology. Nevertheless, within each synset we selected a standard term among the synonyms. Following the *principle of currency*, for the synsets extracted from WordNet, we followed the order of the words in the corresponding synsets. Analogously, for the synsets created or enriched with the words from GeoNames we either kept the original terms - if found appropriate - or we changed them based on the study of some relevant scientific publications or standard vocabularies. For instance, we substituted *mountains* (from the feature class T, including land formations) with *mountain range* (as from Geology terminology), and *hill* (from the feature class U, including undersea entities) with *submarine hill* (as from Oceanography terminology).

In general it is good practice to avoid choosing the same standard term to denote two totally different concepts. However, in one case - for the word *bank* - we had to allow an exception:

- **bank** - sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- **bank** - a building in which the business of banking transacted; *"the bank is on the corner of Nassau and Witherspoon"*

In these extreme cases, it is the context that disambiguates their meaning (*principle of context*). The two meanings of bank were disambiguated as follows:

- **Landform >** Natural elevation > Continental elevation > Slope > Bank
- **Facility >** Business establishment > Bank

Given our purpose and scope, following the *principle of ordering* we ordered the classes based upon the *decreasing quantity* of the entities instantiating the class. Within each chain of terms, from the root to the leaves, we followed the same ordering preference. However, it is not always possible or appropriate to establish this order, especially when the classes do not share any characteristic. For example, we could not establish any order between *body of water* and *landform*. In such cases we preferred the *canonical order*, i.e. the order traditionally followed in library science. The final result, after ordering, was as follows:

Landform
- Natural elevation
  - Continental elevation
    - Mountain
    - Hill
  - Oceanic elevation
    - Seamount
    - Submarine hill
- Natural depression
  - Continental depression
    - Valley
    - Trough
  - Oceanic depression
    - Oceanic valley
    - Oceanic trough

Body of water
- Flowing body of water
  - Stream
    - River
    - Brook
- Stagnant body of water
  - Lake
  - Pond

8 Difficulties

The main difficulties we faced in the process described in the previous sections were mainly due to the different conceptualization in GeoNames and WordNet. Here we briefly describe them.

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 a building or a group of buildings. There are classes which are not buildings but that can be still 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 a place. On the one hand many buildings provide services. Buildings housing banks usually provide transaction services; buildings housing hospitals usually provide health care services; buildings housing libraries usually provide access to the catalogue and book consultation. On the other hand, there are also buildings (or generic constructions) that 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).

Plurals and Parenthesis. 92 class names in GeoNames are given in singular form, e.g., *populated place* and *vineyard*, as well as in plural form, e.g., *populated places* and *vineyards*. In addition, 99 class names are given 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 are used to denote single entities; plural forms indicate groups of entities; mixed forms are preferred when it is not easy to discriminate between the two previous cases.

The approach we followed was to avoid plurals, thus identifying for each plural or mixed form a more appropriate name. For instance, we substituted *lakes* with *lake chain* and *mountains* with *mountain range*.

Dealing with polysemy. 242 class names in GeoNames are polysemous, namely they have two or more similar or related meanings in WordNet. 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 of each class, if available, to each of the meanings of that class in WordNet. In 15 cases, 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*, 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 award 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 the description given in GeoNames and is therefore 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 analyzed 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 ("*a large and stately mansion*") and forth ("*official residence of an exalted person (as a sovereign) correspond to it*") senses for the term in WordNet look like possible candidates. Following the proposed approach, we found that *Buckingham Palace* is the only instance in common with the first sense whereas there are no instances in common with the fourth sense. Therefore, we chose the first sense.

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*. Conversely, we provided distinct names for them. 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*. Clearly, these terms were not just arbitrarily assigned. They were in fact collected from authentic literature on Geography, Oceanography and Geology (e.g., Encyclopaedia Britannica[17]).

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

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

as physical entities, i.e. having physical existence. However, when mapping the classes from GeoNames to WordNet, we observed that for 27 of them, 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 WordNet. In these cases we rather preferred to create a new synset in WordNet somewhere under *physical entity*. In the specific case, we created a new synset with the term *geo-political entity* defined as "*the geographical area controlled or managed by a political entity*" as more specific than *physical object*.

9 The *Space* ontology

Table 5 provides the total number of objects we identified for each C/E/R/A in the *Space* ontology. Note that for the relations we do not include the basic *is-a*, *part-of*, *instance-of* and *value-of* relations. Similarly, for the attributes we do not include the attribute values, but only the attribute names.

Objects	Quantity
Classes (C)	845
Entities (E)	6,907,417
Relations (R)	70
Attributes (A)	31

Table 5. Overall statistics of the *Space* ontology

The facets of entity classes we created are:

- **Region** – "a large indefinite location on the surface of the Earth"
- **Administrative division** – "a district defined for administrative purposes"
- **Populated place** – "a city, town, village, or other agglomeration of buildings where people live and work"
- **Facility** – "a building or any other man-made permanent structure that provides a particular service or is used for a particular industry"
- **Abandoned facility** – "abandoned or ruined building and other permanent man made structure which are no more functional"
- **Land** – "the solid part of the earth's surface"
- **Landform** – "the geological features of the earth"
- **Body of water** – "the part of the earth's surface covered with water (such as a river or lake or ocean)"
- **Agricultural land** – "a land relating to or used in or promoting agriculture or farming"
- **Wetland** – "a low area where the land is saturated with water"

Each of these top-level facets is further sub-divided into several sub-facets. For example, *facility* is sub-divided into *living accommodation*, *religious facility*, *education facility*, *research facility*, *education research facility*, *medical facility*, *transportation facility*, and so on. Similarly, *body of water* is further sub-divided primarily into the two sub-facets *flowing body of water* and *stagnant body of water*. In a similar way, *landform* is further subdivided into the two sub-facets *natural elevation* and *natural depression*. At lower levels all of them are further sub-divided into sub-sub-facets and so on. For example, *natural elevation* consists of *continental elevation* and *oceanic elevation*, while *natural depression* consists of *continental depression* and *oceanic depression*.

Some examples of facets of relations are reported in Table 6.

Direction	East South-east South South-west …
External spatial relation	Alongside Adjacent Near Neighbourhood ...
Sideways spatial relation	Right (right side) Centre-line Left Alongside …
Relative level	Above Below Up ...

Table 6. Examples of spatial relations

The attributes extracted from GeoNames are the following:

- **Name** - "a language unit by which a person or thing is known"
- **Latitude** - "the angular distance between an imaginary line around a heavenly body parallel to its equator and the equator itself"
- **Longitude** - "the angular distance between a point on any meridian and the prime meridian at Greenwich"
- **Altitude** - "elevation especially above sea level or above the earth's surface"
- **Total area** - "the sum of all land and water areas delimited by international boundaries and/or coastlines"

- **Population** - "the number of inhabitants (either the total number or the number of a particular race or class) in a given place (country or city etc.)"
- **Top level domain** - "one of the domains at the highest level in the hierarchical Domain Name System (DNS) of the Internet"
- **Domain name** - "strings of letters and numbers (separated by periods) that are used to name organizations and computers and addresses on the internet"
- **Natural language** - "a human written or spoken language used by a community"
- **Calling code** - "a number usually of 3 digits assigned to a telephone area as in the United States and Canada"
- **Country code** - "short alphabetic geographical codes developed to represent countries and dependent areas"
- **Code** - "a coding system used for transmitting messages requiring brevity or secrecy"
- **Time zone** - "any of the 24 regions of the globe (loosely divided by longitude) throughout which the same standard time is used"

We extended this set by defining some additional attributes, including for instance *depth* (e.g. of a lake), *climate* and *temperature*.

The ontology allows the 6,907,417 entities extracted from GeoNames to be indexed, browsed and exploited. Table 7 provides a fragment of the populated ontology.

Objects	Quantity
Mountain	279,573
Hill	158,072
Mountain range	19,578
Chain of hills	11,731
Submarine hills	78
Chain of submarine hills	12
Oceanic mountain	5
Oceanic mountain range	0

Table 7. A fragment of the populated scheme

In comparing it to the existing geo-spatial ontologies, our *Space* ontology turns out to be much richer in all its aspects. Just to provide a small glimpse, GeoNames and TGN count 663 and 688 classes respectively; while in our ontology we already have, at this stage, 845 classes. In fact, it is worthwhile to underline that, since hospitality is one of the significant features of facets, maintenance costs are kept low as it is always possible to extend it at the desired level of granularity. In this respect, we have been already working to further extend it. For instance, this is what has been done by importing classes and locations from the dataset of the Autonomous Province of Trento in Italy [33]. This allows a more and more accurate annotation, disambiguation, indexing and search on geographical resources.

10 Conclusions

Starting from the observation that ontologies are fundamental towards achieving semantic interoperability in a domain, and that many attempts have been already made in building geo-spatial ontologies, we have emphasized the need to follow a systematic approach, based on a well-founded methodology and guiding principles, to ensure high quality results. We have presented our methodology and guiding principles, mainly inspired by the faceted approach. By applying the methodology and by integrating data coming from GeoNames, WordNet and MultiWordNet we created a large-scale ontology for *Space* where the main components are the classes, the entities, their relations and attributes. The construction procedure we followed is largely automatic, with manual intervention for the critical parts. This allowed obtaining a very satisfactory quantitative and qualitative result. By comparing our ontology w.r.t. well-known geographical resources, we have shown that, in all its components, its coverage is much higher and its quality is much better (as well-established feature of the methodology followed).

As future work, we plan to further extend the coverage of our *Space* ontology. This will be achieved mainly from the analysis of the WordNet synsets that were not considered during the first phase of our work and by importing data from other sources.

11 Acknowledgements

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231126 LivingKnowledge: LivingKnowledge - Facts, Opinions and Bias in Time. Thanks to Gaia Trecarichi and Veronica Rizzi for the pleasant and fruitful discussions about geo-spatial issues. We also want to thank our colleagues Ilya Zaihrayeu and Marco Marasca for their contribution to the definition of the data structures and Abdelhakim Freihat for the importing of the Italian part of MultiWordNet.

References

1. F. Giunchiglia, B. Dutta, V. Maltese. Faceted Lightweight Ontologies. In "Conceptual Modeling: Foundations and Applications", A. Borgida, V. Chaudhri, P. Giorgini, Eric Yu (Eds.) LNCS 5600 Springer, 2009.
2. F. Giunchiglia, A. Autayeu, J. Pane. S-Match: an open source framework for matching lightweight ontologies. The Semantic Web journal, 2010.
3. S. R. Ranganathan. Prolegomena to library classification. Asia Publishing House (1967).
4. F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta. GeoWordNet: a resource for geo-spatial applications. In the Proc. of ESWC, 2010..
5. F. Giunchiglia, A. Villafiorita, T. Walsh. Theories of Abstraction. AI Communications, IOS Press, Vol 10, n.3/4, pp. 167-176, 1997.
6. F. Giunchiglia, T. Walsh. Abstract Theorem Proving. Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI'89), pp. 372-377, 1989.
7. W. Kuhn. Geospatial semantics: Why, of What, and How? Journal of Data Semantics (JoDS) III, pp. 1–24 (2005)

8. F. M. Suchanek, G. Kasneci, G. Weikum. YAGO: a core of semantic knowledge unifying WordNet and Wikipedia. In Proc. of the 16th WWW, pp. 697-706 (2007).
9. M. J. Egenhofer. Toward the Semantic GeoSpatial Web. In the 10th ACM Int. Symposium on Advances in Geographic Information Systems (ACM-GIS), pp. 1-4 (2002).
10. D. Kolas, M. Dean, J. Hebeler. Geospatial Semantic Web: architecture of ontologies. In Proc. of First Int. Conference on GeoSpatial Semantics (GeoS), pp. 183-194 (2005).
11. P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, F. Farazi. A semantic geo-catalogue implementation for a regional SDI. In Proc. of the INPSIRE Conference, 2010.
12. K. Janowicz, S. Schade, A. Bröring, C. Keßler, C. Stasch, P. Maue', Y. Diekhof. A transparent Semantic Enablement Layer for the Geospatial Web. In the Terra Cognita Workshop at ISWC (2009).
13. D. Roman, E. Klien, D. Skogan. SWING – A Semantic Web Service Framework for the Geospatial Domain. In the Terra Cognita Workshop (2006).
14. M. Angioni, R. Demontis, F. Tuveri. Enriching WordNet to Index and Retrieve Semantic Information. Proc. of 2nd Int. Conf. on Metadata and Semantics Research (2006).
15. R. Vorz, J. Kleb, W. Mueller. Towards ontology-based disambiguation of geographical identifiers. In Proc. of the 16th WWW Conference,(2007).
16. D. Buscardi, P. Rosso. Geo-wordnet: Automatic Georeferencing of wordnet. In Proc. of the 5th Int. Conference on Language Resources and Evaluation (LREC) (2008).
17. C. Keßler, K. Janowicz, M. Bishr. 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).
18. C. B. Jones, A.I. Adbelmoty, G. Fu. 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).
19. S. Auer, J. Lehmann, S. Hellman. LinkedGeoData - Adding a Spatial Dimension to the Web of Data. In Proc. of the 8th World Int. Semantic Web Conference (ISWC) (2009).
20. M. S. Chaves, M. J. Silva, B. Martins. A Geographic Knowledge Base for Semantic Web Applications. In Proc. of 20th Brazilian Symposium on Databases (SBBD) (2005).
21. G. A. Miller, F. Hristea. WordNet Nouns: classes and instances. Computational Linguistics, 32(1):1.3 (2006)
22. M. J. Egenhofer, M. P. Dube. Topological Relations from Metric Refinements. In Proc. of the 17th ACM SIGSPATIAL Int. Conference on Advances in GIS (2009).
23. F. Giunchiglia, P. Shvaiko, P. Yatskevich. Discovering Missing Background Knowledge in Onology Matching. In Proceedings of the 17th European Conference on Artificial Intelligence - ECAI 2006.
24. F. Giunchiglia, I. Zaihrayeu. Lightweight ontologies. In S. LNCS, editor, Encyclopedia of Database Systems, 2008.
25. D. Pullar, M. J. Egenhofer. Toward formal definitions of topological relations among spatial objects. In Proceedings of the 3rd International Symposium on Spatial Data Handling, 1988, Sydney, Australia, pp. 165–176.
26. I. B. Arpinar, A. Sheth, C. Ramakrishnan. Geospatial ontology development and semantic analytics. Handbook of Geographic Information Science, J. P. Wilson, A. S. Fotheringham (Eds.), Blackwell Pub., 2004.
27. A. I. Abdelmoty, P. Smart, C.B. Jones. Building Place Ontologies for the Semantic Web: issues and approaches. In Proc. of the 4th ACM workshop on GIR, 2007.
28. B. Smith, D. M. Mark. Ontology and geographic kinds. In Proc. of the International Symposium on Spatial Data Handling, Vancouver, Canada, 1998.
29. M. Egenhofer, J. Herring. Categorization binary topological relationships between regions, lines, and points in geographic databases. In "A Framework for the Definition of Topological Relationships and an Approach to Spatial Reasoning within this Framework", M. Egenhofer and J. Herring (Eds.), Santa Barbara, CA, 1991.

30. V. Broughton. The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings, 58(1/2) pp. 49-72, 2006.
31. L. Spiteri. A Simplified Model for Facet Analysis. Journal of Information and Library Science, Vol 23, pp. 1-30, 1998.
32. B. Dutta, F. Giunchiglia, V. Maltese. A facet-based methodology for geo-spatial modeling. In Proceedings of the GEOS, Vol. 6631, 2011.
33. F. Farazi, V. Maltese, F. Giunchiglia, A. Ivanyukovich. A faceted ontology for a semantic geo-catalogue. In Proceedings of the ESWC 2010.
34. Ranganathan, S. R. 1965. The Colon Classification. In the Rutgers Series on Systems for the Intellectual Organization of Information, S. Artandi (etd.), IV. Graduate School of Library Science, Rutgers University, New Brunswick, NJ.
35. G. Battacharyya, 1975. POPSI: its fundamentals and procedure based on a general theory of Subject Indexing Languages. Library Science with a Slant to Documentation, 16, 1, 1-34.
36. F. Baader, D. Calvanese, D. McGuinness, D. Nardi, P. F. Patel-Schneider, 2002. The Description Logic Handbook: Theory, Implementation and Applications. Cambridge University Press.

Appendix: The body of water facet

Body of water

- Ocean
- Sea
  - Bay
- Bight
- Gulf
- Inlet
  - Cove
- Flowing body of water
  - Stream
    - River
      - Lost river
    - Brook
      - Brooklet
      - Tidal brook
    - Headstream
    - Rivulet
    - Branch
      - Anabranch
      - Billabong
      - Distributory
      - Tributory
    - Canalized stream
    - Tidal stream
    - Intermittent stream
  - Channel
    - Watercourse
      - Abandoned watercourse
    - Navigation channel
    - Reach
    - Marine channel
    - Lake channel
    - Cutoff
  - Overfalls
  - Current
    - Whirlpool
  - Section of stream
    - Headwaters
    - Confluence
    - Stream mouth
      - Estuary
    - Midstream
    - Stream bend
  - Waterway
    - Ditch
    - Rapid
  - Spring
    - Hot spring
    - Geyser
    - Sulphur spring
  - Waterfall
    - Cataract
    - Cascade
- Stagnant body of water
  - Lake
    - Lagoon
    - Chain of lagoons
    - Salt lake
      - Intermittent salt lake
    - Chain of intermittent salt lakes
    - Chain of salt lakes
    - Underground lake
    - Intermittent lake
    - Chain of intermittent lakes
    - Glacial lake
    - Crater lake
    - Chain of crater lakes
    - Oxbow lake
      - Intermittent oxbow lake
  - Chain of lakes
  - Pond
    - Salt pond
      - Intermittent salt pond
    - Chain of salt ponds
    - Fishpond
    - Chain of fishponds
    - Horsepond
    - Mere
    - Millpond
  - Pool
    - Intermittent pool
      - Billabong
    - Mud puddle
    - Wallow