A Facet-Based Methodology for Geo-Spatial Modeling

Biswanath Dutta, Fausto Giunchiglia, and Vincenzo Maltese

DISI - Università di Trento, Trento, Italy

Abstract. Space, together with time, is one of the two fundamental dimensions of the universe of knowledge. Geo-spatial ontologies are essential for our shared understanding of the physical universe and to achieve semantic interoperability between people and between software agents. In this paper we propose a methodology and a minimal set of guiding principles, mainly inspired by the faceted approach, to produce high quality ontologies in terms of robustness, extensibility, reusability, compactness and flexibility. We demonstrate - with step by steps examples - that by applying the methodology and those principles we can model the space domain and produce a high quality facet-based large scale geo-spatial ontology comprising entities, entity classes, spatial relations and attributes.

Keywords: space domain, methodology, principle, theory, domain ontology, geo-spatial ontology.

1 Introduction

Space and time are the two fundamental dimensions of the universe of knowledge [12, 3]. The notion of space is essential to understand the physical universe. We consider space as is 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 concepts, often known as features, like land formations (continents, islands, countries), water formations (oceans, seas, streams) and physiographical concepts (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).

Spatial (geo-spatial) and temporal ontologies, because representing a shared understanding of a domain [10], are essential to achieve semantic interoperability between people and between applications. Equally important, the definition of entity types and corresponding properties has become a central issue in data exchange standards where a considerable part of the semantics of data may be carried by the categories that entities are assigned to [20]. As a matter of fact, current standards - for instance the specifications provided for the geographical domain by the Open Geospatial Consortium (OGC)[1] - do not represent an effective solution to the interoperability problem. In fact, they only aim at syntactic agreement [11] by fixing the standard terms and not allowing for variations on the terminology to be used.

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

C. Claramunt, S. Levashkin, and M. Bertolotto (Eds.): GeoS 2011, LNCS 6631, pp. 133–150, 2011.

Several frameworks have been proposed to build and maintain geo-spatial ontologies [13, 14, 15, 21], and we also recently proposed our multilingual ontology, called GeoWordNet, that overcomes their qualitative and quantitative limitations (as extensively described in [2]). However, to the best of our knowledge no systematic ways, i.e. based on a well founded methodology and guiding principles, for building geo-spatial ontologies have been proposed so far.

Our main contribution is a methodology and a minimal set of guiding principles aimed at modelling the spatial domain and at building the corresponding background knowledge taking into account the classes, the entities, their relations and properties. As explained across this paper, the domain knowledge is organized following the well founded *faceted approach* [3], borrowed from library and information science. Note that the methodology and the guiding principles we propose are not only applicable to the spatial domain, but across domains. In this approach, the analysis of the domain allows the identification of the basic classes of real world objects. They are arranged, per *genus et differentia* (i.e. by looking at their commonalities and their differences), to construct specific ontologies called *facets*, each of them codifying a different aspect of the domain at hand. This allows being much more rigorous in the definition of the domain and its parts, in its maintenance and use [1]. The intended use of this background knowledge is manifold. Identifying the domain specific terminology and corresponding entity names allows using them to annotate, index and search geographical resources as well as for word sense disambiguation.

The rest of the paper is organized as follows. In Section 2 we illustrate our methodology and the guiding principles we propose to model a domain. In Section 3, with some step by step examples, we highlight some of the issues we faced in building the space domain. In Section 4 we describe how we further organize the elements of the domain into three main categories: entity classes, relations and attributes. Section 5 provides some statistics about the space domain, as we modelled it so far. Section 6 concludes the paper and provides our future research directions.

2 The Methodology

Our methodology is mainly inspired by the *faceted approach* proposed by the Indian librarian Ranganathan [3] at the beginning of the last century. In this approach, the domain under examination is decomposed into its basic constituents, each of them denoting a different *aspect of meaning*. Each of these components is called a *facet*. More precisely, a facet is a hierarchy of homogeneous terms, where each term in the hierarchy denotes a primitive atomic concept, i.e. a primitive class of real world objects. In the next two sections we describe the main steps in the creation of the set of facets for a given domain and the guiding principles to be used.

2.1 Steps in the Process

The building process is organized into subsequent phases as follows:

- ***Step 1: Identification of the atomic concepts***. It consists in collecting the terms representing the relevant (according to the purpose) real world entities of the domain at hand. Each term denotes a class of objects. In general, this is mainly

done by interviewing domain experts and by reading available literature on that particular domain including *inter-alia* indexes, abstracts, glossaries, reference works. Analysis of query logs, when available, can be extremely valuable to determine user's interests. Each term is analyzed and disambiguated into an atomic concept. This can be approximated by associating a natural language definition to each of them. For instance, *river* can be defined as *"a large natural stream of water (larger than a brook)"* and represents the set of all real world rivers.

- ***Step 2: Analysis.*** The atomic concepts are analyzed per *genus et differentia*, i.e. in order to identify their commonalities and their differences. The main goal is to identify as many distinguishing properties - called *characteristics* - as possible of the real world entities represented by the concepts. This allows being as fine grained as wanted in differentiating among the concepts. For instance, for the concept *river* 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.*** The synthesis aims at arranging the atomic concepts into *facets* by characteristic. At each level of the hierarchy - each of them representing a different level of abstraction - similar concepts are grouped by a common characteristic. Concepts sharing the same characteristic form an *array* of homogeneous concepts. Concepts in each array can be further organized into sub-groups (or sub-facets) generating a new level in the hierarchy. Children are connected to their parent through a *genus-species* (is-a) or *whole-part* (part-of) relation. For instance, due to their commonalities we could place in the same array the concept *river* and the concept *brook*.

- ***Step 4: Standardization.*** Each atomic concept can be potentially denoted with different words. When more than one candidate is available, a standard (or preferred) term should be selected among the synonyms. 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[2] approach where terms are ranked in the synset and the first one is the preferred. For instance, the concept *pharynx*, defined as *"the passage to the stomach and lungs; in the front part of the neck below the chin and above the collarbone"*, can be denoted also with *throat*. However, *pharynx* is the one most commonly used by subject specialists in the medicine domain.

- ***Step 5: Ordering.*** Concepts 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 classification system.

[2] `http://wordnet.princeton.edu/`

For example, since the purpose of the medicine domain is to prevent and cure the diseases that can affect the human body, the facets in the domain can be, in order: body and its organs, diseases and treatments.

Following the steps above leads to the creation of a set of facets. They constitute a *faceted representation scheme* for the domain. A faceted representation scheme codifies the basic building blocks that can be used - at indexing, classification and searching time - to construct complex labels, called *subjects*. This is what in library science is called post-coordination, in contrast to pre-coordination, as it is pursued by classical enumerative approaches, where a totally new concept is added to the scheme each time a new subject has to be included. Pre-coordination clearly leads to an exponential explosion in the number of subjects, while in the faceted approach they are instead created by composing the atomic concepts from the facets. A faceted representation scheme corresponds to what in our previous work we call the *background knowledge* [4, 5], i.e. the a-priori knowledge which must exist to make semantics effective. Each facet corresponds to what in logics is called *logical theory* [23, 24] and to what in computer science is called *ontology*, or more precisely *lightweight ontology* [6].

2.2 Guiding Principles

In this section we propose a minimal set of guiding principles for building facet-based domain ontologies. These principles are derived from the canons postulated by Ranganathan in his work on prolegomena to library classification [3]. Originally he proposed a huge amount of canons and principles, with a lot of redundancy and complexities. However, instead of going into the technicalities of all of them, here we rather prefer to summarize them into a minimal set of basic principles to be followed:

1. ***Relevance.*** The selection of the characteristics to form the facets in the scheme from the atomic concepts should reflect the purpose, scope and subject of the classification system. For example, while the characteristic *by grade* looks appropriate to classify the universe of boys and girls in the context of the education domain, for sure it is not suitable to classify the universe of cows. In the latter case *by breed* would be more realistic. It is worthwhile also noting that the selection of characteristics should be done carefully, as they cannot be changed unless there is a change in the purpose, scope and subject of the classification system.

2. ***Ascertainability.*** Characteristics must be definite and ascertainable. 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 flowing body of water, thus the facet *flowing body of water* represents a permanent characteristic of spring.

4. ***Exhaustiveness.*** Classes in an array of classes and the sub-classes in an array of sub-classes should be totally exhaustive w.r.t. their respective common immediate universe. For example, to classify the universe of people *by gender*, we

need both *male* and *female*. If we miss any of these two, the classification becomes incomplete. Note that we are not pretending to achieve such exhaustiveness in advance. The identification of the classes is based on the known real world entities. It is always possible to extend the classification in the future.

5. ***Exclusiveness.*** All the characteristics used to classify an entity must be *mutually exclusive*, i.e. no two facets can overlap in content. For example, the universe of people cannot be classified by both the characteristics *age* and *date of birth*, as they produce the same divisions.

6. ***Context.*** The denotation of a term is determined by its position in a classification system. This principle is particularly helpful for distinguishing the homographs, i.e. same term but totally different meanings. See for instance how we solve the ambiguity of the term *bank* in Section 3.4.

7. ***Currency.*** The terms denoting the classes and sub-classes 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 terms used to denote the classes and sub-classes should not reflect any bias or prejudice (e.g. of gender, cultural, religious), or express any personal opinion of the person who develops the classification system. For example, it is not appropriate to use terms like *minor author* or *black man*.

9. ***Ordering.*** The order should reflect the purpose, scope and subject of the classification system. Also, the ordering of facets should be consistent and should not be changed unless there is a change in the purpose, scope or subject of the classification system. Note that ordering carries semantics as it provides implicit relations between coordinate (siblings) terms.

Following the principles guarantees the creation of high quality domain ontologies in terms of robustness, extensibility, reusability, compactness and flexibility [3, 25, 26].

3 The Space Domain

Following the steps and the principles described in the previous section, we created a faceted representation scheme for the space domain.

3.1 Identification of the Atomic Concepts

Similarly to any other domain, our first step was to collect the terms and to identify the corresponding concepts representing real world geographical entities. For instance, the term *lake* corresponds to the concept *"a body of (usually fresh) water surrounded by land"* (as it is defined in WordNet) and represents the set of all real world lakes. To collect such terms we mainly used GeoNames[3] and WordNet (version 2.1).

[3] `http://www.geonames.org`

We also occasionally used the Getty Thesaurus of Geographical Names (TGN)[4] and referred to domain specific scientific literature to solve ambiguous cases.

- ***GeoNames*** is one of the most famous geo-spatial databases. It includes over 8 millions of place names in multiple languages. It also provides corresponding properties such as latitude, longitude, altitude and population. At top level, the places are categorised into 9 feature classes, further divided into 663 sub-classes.

- **WordNet** is the Princeton lexical database for the English language. 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. Basically, 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*.

- **TGN** is a structured vocabulary for place names. Similarly to GeoNames it provides around 1.1 millions of place names and 688 feature classes. It includes administrative political (e.g., cities, nations) and physical (e.g., mountains, rivers) entities. It focuses on places particularly important for the study of art and architecture.

As a preliminary step, we mapped GeoNames feature classes with WordNet synsets. From their integration we created GeoWordNet, one of the biggest multi-lingual geo-spatial ontologies currently available and therefore particularly suitable to provide semantic support for spatial applications. A large subset of GeoWordNet is available as open source[5] in plain CSV and RDF formats. This mapping allowed, among other things, identifying the main subtrees of WordNet containing synsets representing geographical classes. These are 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"

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

[5] http://semanticmatching.org/download.html

It is worthwhile to underline that not all the nodes in these sub-trees necessarily need to be part of the space domain. As a matter of fact, most of the descendants of *location* and *artifact* do not fall under the space domain. For instance the following:

(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"

3.2 Analysis

The purpose of the analysis is to enlist the characteristics to be used to form the facets. In other words they are used to form the different levels of abstraction of the conceptual categories. Real world geographical 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 of atomic concepts.

In order to illustrate the analysis process, consider the following list of real world geographical entities and their corresponding glosses.

- **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 the previous section, it is not difficult to 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

3.3 Synthesis

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

The concepts *mountain* and *hill* can be further differentiated *by size*. Note that, according to the guiding principles, size is a good distinguishing characteristic for the space domain. 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* [16].

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 by putting *river* under *stream*.

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

Body of water
- Flowing body of water
  - Stream
    - River

Landform
- Natural elevation
  - Mountain
  - Hill

An important property of facets is that they are *hospitable* (the interested reader can refer to [1] for the list of the most important properties of facets), i.e. they can be easily extended to accommodate additional concepts as needed. Assume for instance that the new concept *lake* (*a body of (usually fresh) water surrounded by land*) is identified. By analyzing 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 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 kinds-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* in the same array level 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 an entity must be *mutually exclusive*. So, we should not include all those four sub-classes in the same array.

Similarly to lakes, we can extend the *natural elevation* facet in order to accommodate the concept *valley* (*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 that valleys are seen in both the oceanic areas (called *oceanic valley*) and continental areas (called *valley*). 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.

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

By applying more and more characteristics of division, the extension of the concepts decreases and the intension increases. For example, there are fewer kinds-of *lake* than kinds-of *stagnant body of water*. See the appendix for a complete example.

3.4 Standardization

For each concept a standard term was selected while all the others are still kept as synonyms. This allows variations supporting semantic interoperability between systems using different terminology. For the concepts extracted from WordNet, we followed the order of the words in the corresponding synsets. For the concepts extracted from GeoNames we either kept the original terms - if found appropriate - or we changed them based on the study of some scientific publications. For instance, we changed *mountains* (from the feature class T, including land formations) into *mountain range* (as from Geology terminology), and *hill* (from the feature class U, including undersea entities) into *submarine hill* (as from Oceanography terminology). Some other examples and the criteria we used can be found in [2]. For the remaining concepts we used standard vocabularies.

In general it is good practice to avoid choosing the same standard term to denote two totally different concepts within a domain. 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

3.5 Ordering

Given our purpose and scope, we ordered the classes based upon the *decreasing quantity* of the entities instantiating the class. Within each chain of concepts, 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 an 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

4 Elements of the Space Domain

The faceted representation scheme we created represents *classes* of real world geographical entities. To complete our model of the domain we also provide in this section the *relations* between them and their *attributes*. We consider classes, relations, and attributes as the three fundamental components, or categories, of any domain.

4.1 Entity Classes

This category contains the classes of the faceted representation scheme. It is the main means to determine what an object is. In other words, we can characterize each real world geographical entity by associating it to its entity class. The space domain consists of the following basic facets:

- **Region** - a large indefinite location on the surface of the Earth; "penguins inhabit the polar regions"
- **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; "the assembly plant is an enormous facility"

- **Abandoned facility** - abandoned or ruined building and other permanent man made structure which are no more functional
- **Land** - 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"*
- **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) "they invaded our territorial waters"; "they were sitting by the water's edge"

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 the sub-facets *continental elevation* and *oceanic elevation*, while *natural depression* consists of the sub-facets *continental depression* and *oceanic depression*.

4.2 Relations

The real world entities indeed exist in the real world and they occupy some region of space on the earth surface. It is quite natural to describe how objects are located in space in relation to other objects. Understanding spatial relations is one of the fundamental features of Geographic Information Systems (GIS). According to Egenhofer and Herring [19], spatial regions form a relational system comprising the relations between interiors, exteriors, and boundaries of two objects. Spatial relations play an important role for effective geographical knowledge discovery. Consider for instance the following queries:

- "R*etrieve all the secondary schools within 500 meters of the Dante railway station in Trento*"
- "*Find all the highways of the Trentino province adjacent to marine areas*".

Since people tend to express and understand spatial relations through natural language [8], we also expressed them accordingly. Arpinar et al. [8] suggest three major types of spatial relations: topological relations, cardinal direction and proximity relations. Egenhofer and Dupe [9] propose topological and directional relations. According to them, topological properties have a leading role in qualitative spatial reasoning. Pullar and Egenhofer in [7] 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, near) and fuzzy relations (e.g. next to, close).

The spatial relations we propose can be compared to the work in [7]. However, in addition to the standard direction, topological, ordinal, distance and fuzzy relations,

we extend them by including relative level (e.g. above, below), longitudinal (e.g. in front, behind), side-wise (e.g. right, left), position in relation to border or frontier (e.g. adjacent, overlap) and other similar relations. A partial list of the spatial relations we propose is reported in Table 1, organized in a faceted fashion.

Note that in addition to the spatial relations, we also consider some other kinds of relations, which can be treated as functional relations. For example, in the context of lakes, primary inflow and primary outflow are two important relations.

Table 1. Partial list of spatial relations

Direction	East South-east South South-west …
Internal spatial relation	Inside Central - Midpoint - Midplane - Concentric - Eccentric …
External spatial relation	Alongside Adjacent Near Neighbourhood ...
Position in relation to a border or frontier	Adjacent (touching) Overlap Opposite …
Longitudinal spatial relation	In front Mid-length (amidships) Behind In line Toward …
Sideways spatial relation	Right (right side) Centre-line Left Alongside Across …
Relative level	Above Below Up ...

4.3 Attributes

An attribute is an abstraction belonging to or a characteristic of an object. This is a construct through which objects or individuals can be distinguished. Attributes are therefore effective for Named Entity Recognition (NER) [18] and for efficient geographical information retrieval (GIR) [17]. For example, there are 14 locations called Rome in United States of America (USA), one in Italy (the capital city of Italy) and one in France. Using the latitude and longitude attributes stored in the background knowledge - for instance GeoWordNet - we can easily distinguish them.

Attributes are primarily *qualitative* and *quantitative* in nature. For example, we may mention depth (of a river), surface area (of a lake), length (of a highway) and altitude (of a hill). For each of these attributes, we may have both qualitative and quantitative values. We store the possible qualitative values in the background knowledge. This provides a controlled vocabulary for them. They are mostly *adjectives*. For example, for depth (of a river) the possible values are {wide, narrow}. Similarly, for altitude (of a hill) the possible values are {high, low}.

We also make use of *descriptive* attributes. They are used to describe, usually with a short natural language sentence, a specific aspect of an entity. Typical examples are the history (of a monument) or the architectural style (of a building) or any user defined tag.

5 Statistics

In this section we report some statistics about our space domain. Table 2 provides the total number of objects we identified. Note that for the relations we do not count the taxonomical *is-a* and *part-of* relations. Similarly, for the attributes we do not count the number of attribute values, but only the attribute names. As part of this work, the faceted representation scheme we developed has been aligned with GeoWordNet and it is used to classify its 6,907,417 locations. This provides a faceted infrastructure to index, browse and exploit GeoWordNet. We are further increasing this number by importing locations from other sources. For instance, with the SGC project in collaboration with the Autonomous Province of Trento in Italy, a dataset of 20,162 locations of the province has been analyzed and integrated with GeoWordNet [22]. Table 3 provides a fragment of the scheme populated with the locations from GeoWordNet.

Table 2. Statistics of the Space domain

Objects	Quantity
Entity classes	845
Relations	70
Attributes	35
Locations	6,907,417

In comparing our space domain with the existing reputed and popularly used geospatial ontologies, like GeoNames and TGN, our space domain is much richer in all its aspects. Just to provide a small glimpse, GeoNames and TGN count 663 and 688

classes respectively; while in our domain we have, at this stage, 845 classes. Our plan is in fact to further increase the coverage of our space domain, both in terms of entities, entity classes, arbitrary relations and attributes. This allows a more and more accurate annotation, disambiguation, indexing and search on geographical resources. It is worthwhile to underline that, since hospitality is one of the significant features of our representation scheme, we can extend the domain at any given point of time and at any extend of granularity as we want to be.

Table 3. A fragment of the populated scheme

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

6 Conclusion

Starting from the observation that ontologies are fundamental to achieve semantic interoperability in a domain, and that many attempts have been already made towards 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, used for several decades and currently in use with great success in the library and information science field. By applying the methodology we modelled the space domain as a faceted representation scheme where the main components are the entities, the entity classes, their relations and attributes. By comparing our result w.r.t. well known geographical resources, like GeoNames and TGN, we have shown how, in all its components, our coverage is much bigger and our quality (as a well established feature of the methodology followed) is much better.

As future work, we plan to further extend the coverage of our space domain, in terms of entities, entity classes, relations and attributes. This will be achieved mainly from the analysis of the WordNet concepts not considered during the first phase of our work and by importing entities from other sources.

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.

References

1. Giunchiglia, F., Dutta, B., Maltese, V.: Faceted Lightweight Ontologies. In: Borgida, A.T., Chaudhri, V.K., Giorgini, P., Yu, E.S. (eds.) Conceptual Modeling: Foundations and Applications. LNCS, vol. 5600, pp. 36–51. Springer, Heidelberg (2009)
2. Giunchiglia, F., Maltese, V., Farazi, F., Dutta, B.: GeoWordNet: A resource for geo-spatial applications. In: Aroyo, L., Antoniou, G., Hyvönen, E., ten Teije, A., Stuckenschmidt, H., Cabral, L., Tudorache, T. (eds.) ESWC 2010. LNCS, vol. 6088, pp. 121–136. Springer, Heidelberg (2010)
3. Ranganathan, S.R.: Prolegomena to library classification. Asia Publishing House (1967)
4. Giunchiglia, F., Shvaiko, P., Yatskevich, P.: Discovering Missing Background Knowledge in Onology Matching. In: Proceedings of the 17th European Conference on Artificial Intelligence - ECAI 2006 (2006)
5. Giunchiglia, F., Kharkevich, U., Zaihrayeu, I.: Concept Search: Semantics Enabled Syntactic Search. In: Semantic Search 2008 Workshop (SemSearch2008) at the 5th European Semantic Web Conference, ESWC (2008)
6. Giunchiglia, F., Zaihrayeu, I.: Lightweight ontologies. In: Ozsu, M.T., Liu, L. (eds.) Encyclopedia of Database Systems. Springer, Heidelberg (2008)
7. Pullar, D., Egenhofer, M.J.: Toward formal definitions of topological relations among spatial objects. In: Proceedings of the 3rd International Symposium on Spatial Data Handling, Sydney, Australia, pp. 165–176 (1988)
8. Arpinar, I.B., Sheth, A., Ramakrishnan, C.: Geospatial ontology development and semantic analytics. In: Wilson, J.P., Fotheringham, A.S. (eds.) Handbook of Geographic Information Science. Blackwell Pub., London (2004)
9. Egenhofer, M.J., Dube, M.P.: Topological relations from metric refinements. In: ACM GIS, Seattle, WA, USA (2009)
10. Gruber, T.R.: Toward Principles for the Design of Ontologies Used for Knowledge Sharing. International Journal of Human and Computer Studies 43(5/6), 907–928 (1995)
11. Kuhn, W.: Geospatial semantics: Why, of What, and How? Journal of Data Semantics (JoDS) III, 1–24 (2005)
12. Maltese, V., Giunchiglia, F., Denecke, K., Lewis, P., Wallner, C., Baldry, A., Madalli, D.: On the interdisciplinary foundations of diversity. In: At the first Living Web Workshop at ISWC 2009 (2009)
13. 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)
14. Auer, S., Lehmann, J., Hellmann, S.: LinkedGeoData: Adding a spatial dimension to the web of data. In: Bernstein, A., Karger, D.R., Heath, T., Feigenbaum, L., Maynard, D., Motta, E., Thirunarayan, K. (eds.) ISWC 2009. LNCS, vol. 5823, pp. 731–746. Springer, Heidelberg (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. Smith, B., Mark, D.M.: Ontology and geographic kinds. In: Proc. of the International Symposium on Spatial Data Handling, Vancouver, Canada (1998)
17. Jones, C.B., Abdelmoty, A.I., Fu, G.: Maintaining Ontologies for Geographical Information Retrieval on the Web. In: Chung, S., Schmidt, D.C. (eds.) CoopIS 2003, DOA 2003, and ODBASE 2003. LNCS, vol. 2888, pp. 934–951. Springer, Heidelberg (2003)
18. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. Knowledge Engineer. Review 18(1), 1–31 (2003)

19. Egenhofer, M., Herring, J.: Categorization binary topological relationships between regions, lines, and points in geographic databases. In: Egenhofer, M., Herring, J. (eds.) A Framework for the Definition of Topological Relationships and an Approach to Spatial Reasoning within this Framework, Santa Barbara, CA (1991)
20. Mark, D.M.: Toward a theoretical framework for geographic entity types. In: Frank, A.U., Campari, I. (eds.) COSIT 1993. LNCS, vol. 716, pp. 270–283. Springer, Heidelberg (1993)
21. Duce, S.: Towards an Ontology for Reef Islands. In: Proceedings of the 3rd International Conference on GeoSpatial Semantics (2009)
22. Farazi, F., Maltese, V., Giunchiglia, F., Ivanyukovich, A.: A semantic geographical catalogue for semantic search. DISI Technical report (2010)
23. Giunchiglia, F., Villafiorita, A., Walsh, T.: Theories of Abstraction. AI Communications 10(3/4), 167–176 (1997)
24. Giunchiglia, F., Walsh, T.: Abstract Theorem Proving. In: Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI 1989), pp. 372–377 (1989)
25. Broughton, V.: The need for a faceted classification as the basis of all methods of information retrieval. Aslib Proceedings 58(1/2), 49–72 (2006)
26. Spiteri, L.: A Simplified Model for Facet Analysis. Journal of Information and Library Science 23, 1–30 (1998)

Appendix: The Complete 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
  - ▪ 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

- 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
  - 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