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

DOMAINS AND CONTEXT: FIRST STEPS TOWARDS MANAGING DIVERSITY IN KNOWLEDGE

Fausto Giunchiglia, Vincenzo Maltese,
Biswanath Dutta

June 2011

Technical Report # DISI-11-468

Also in: Journal of Web Semantics, special issue on Reasoning with Context in the Semantic Web, April 2012.

Domains and context: first steps towards managing diversity in knowledge

Fausto Giunchiglia, Vincenzo Maltese, Biswanath Dutta

DISI - Università di Trento, Trento, Italy

Abstract. Despite the progress made, one of the main barriers towards the use of semantics is the lack of background knowledge. Dealing with this problem has turned out to be a very difficult task because on the one hand the background knowledge should be very large and virtually unbound and, on the other hand, it should be context sensitive and able to capture the diversity of the world, for instance in terms of language and knowledge. Our proposed solution consists in addressing the problem in three steps: (1) create an extensible diversity-aware knowledge base providing a continuously growing quantity of properly organized knowledge; (2) given the problem, build at run-time the proper context within which perform the reasoning; (3) solve the problem. Our work is based on two key ideas. The first is that of using domains, i.e. a general semantic-aware methodology and technique for structuring the background knowledge. The second is that of building the context of reasoning by a suitable combination of domains. Our goal in this paper is to introduce the overall approach, show how it can be applied to an important use case, i.e. the matching of classifications, and describe our first steps towards the construction of a large scale diversity-aware knowledge base.

Keywords. context; diversity; implicit assumptions; faceted approach; diversity-aware knowledge base; semantic matching

1. Introduction

Semantics is core in many knowledge management applications, such as natural language data and metadata understanding [20, 22, 23, 24], natural language driven image generation [54], abstract reasoning [55, 56], converting classifications into formal ontologies [7, 27, 28], automatic classification [25, 26], ontology matching [17, 18, 19] and semantic search [29]. However, despite the progress made, one of the main barriers towards the success of these applications is the lack of background knowledge. In fact, as underlined by several studies (see for instance [8, 9, 10, 11, 51]) without high quality and contextually relevant background knowledge it is impossible to achieve accurate enough results.

Dealing with this problem has turned out to be a very difficult task. In fact, on the one hand, in order to provide all the possible meanings of the words and how they are related to each other, the background knowledge should be very large and virtually unbound. On the other hand, the background knowledge should be context sensitive and able to capture the diversity of the world. The world is extremely diverse and diversity is visibly manifested in language, data and knowledge. The same real world object can be referred to with many different words in different communities and in different languages. For instance, it is widely known that in some Nordic circumpolar groups of people the notion of *snow* is denoted with hundreds of different words in the local language carrying very fine grained distinctions [1]. This phenomenon is often a function of the role and importance of the real world object in the life of a community. Conversely, the same word may denote different notions in different domains; for instance, *bug* as insect in entomology and *bug* as a failure or defect in a computer program in computer science. Space, time, individual goals, needs, competences, beliefs, culture, opinions and personal experience also play an important role in characterizing the meaning of a word. Diversity is an unavoidable and intrinsic property of the world and as such it cannot be avoided. At the same time, diversity is a local maximum since it aims at minimizing the effort and maximizing the gain [35].

Our approach is to take into account this diversity and exploit it to *make explicit the local semantics*, i.e. the meaning of words in a certain context, such that information becomes unambiguous to humans as well as to machines. Towards this goal a preliminary step is the creation of a diversity-aware knowledge base. This requires appropriate methodologies for its representation, construction and maintenance. With this purpose, we propose and adapt the *faceted approach*, a well-established methodology used in library science for the organization of knowledge in libraries [21]. In this paper, we describe the fundamental notions of *domain* and its components, called *facets*, which allow capturing diversity and, at the same time, allow for an incremental growth of the knowledge base.

The rest of the paper is organized as follows. In Section 2 we explain the main steps of our approach by taking semantic matching as an example. Semantic matching has been chosen because of its intrinsic importance, witnessed by the large amount of research and publications in this area and also because it was the main motivation which originally led us to the problem of managing diversity. In Section 3 we provide the definitions of domain and facet, we present the corresponding data model and describe their fundamental properties. In Section 4 we provide our definition of context and explain how to build and use it at run-time by selecting from the background knowledge the language and knowledge

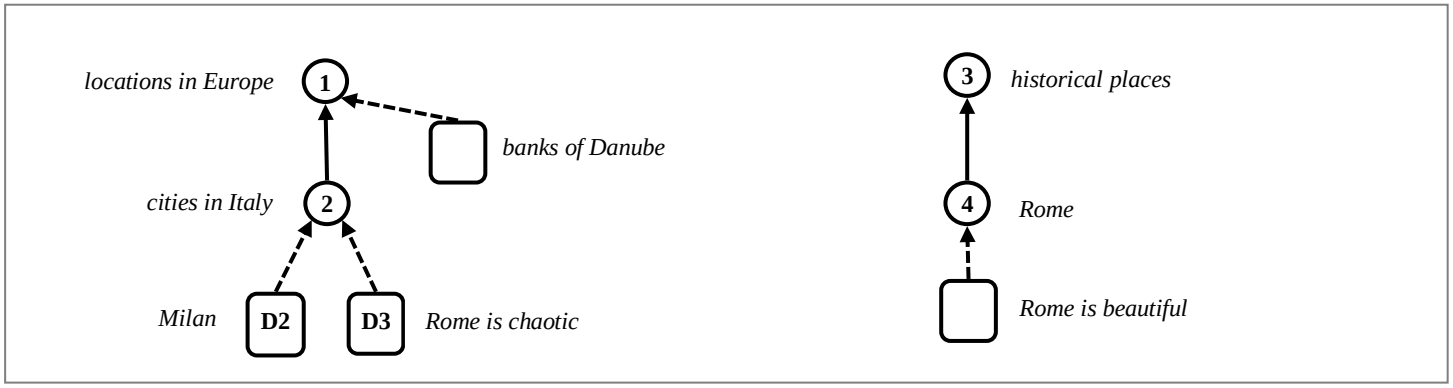

Fig. 1 – An example of two classifications

of the domains which are relevant to the problem. In Section 5 we describe and provide a brief description of the diversity-aware knowledge base that we have been developing. Section 6 focuses on the related work in terms of the notion of context (Section 6.1), methodologies for the construction and maintenance of domain knowledge (Section 6.2), existing knowledge bases and approaches followed for their construction (Section 6.3). Section 7 concludes the paper by summarizing the work done, listing the open problems and outlying the future work.

2. Diversity-aware semantic matching

Consider the example in Fig. 1. It represents two very simple classifications that, for instance, might have been created by two different persons. Round nodes represent categories while rectangles exemplify annotated documents. Solid arrows between nodes represent sub-category relations while dashed arrows denote the fact that a document is categorized into a certain category. Corresponding labels are also given attached to nodes. Initially, we do not know the circumstance in which they were created nor their precise purpose. As humans, we may understand that they were both built to categorize documents about places and, by tagging them, to eventually provide some opinions about those places. The identification of semantic correspondences between the nodes makes the two classifications interoperable [19], for instance, we may use the content of the node *Rome* to enrich the content of the node *cities in Italy*.

2.1. Semantic matching

Semantic matching is a technique used to identify semantic correspondences between two classifications. A good survey of tools developed for this purpose can be found in [16]. The system and algorithm proposed in [17, 18], that lead to the creation of the S-Match[1] tool and that we use across this paper, is organized into four macro steps as follows:

1. For all the labels in the two classifications compute the *concept at label*
2. For all the nodes in the two classifications compute the *concepts at node*
3. For all pairs of labels in the two classifications compute the semantic relations between the concepts at labels
4. For all pairs of nodes in the two classifications compute the relations between the concepts at node

With the first two steps S-Match converts the two classifications into *lightweight ontologies* [7]. Lightweight ontologies are tree-like formal ontologies where each label at node is translated into a propositional Description Logic (DL) formula codifying the meaning of the node and where each formula at node is subsumed by the formula of the node above. For instance, the meaning of the node *Rome* in the second classification in Fig. 1 is *Rome the historical place*. Since these classifications are used to classify documents, the meaning of each node has to be interpreted in terms of *classification semantics* [6]. This means that the extension of each concept is the set of documents about the label of the node and the arcs between nodes represent subset relations. For instance, the extension of the node *Rome* is the set of all documents about *Rome the historical place* which is a subset of all documents about *historical places*. Note that this is the semantics implicitly used in libraries where, according to the get-specific principle [25], each document is classified as deep as possible in a classification. Many types of commonly used ontologies (such as on-line catalogs, file systems, web directories and library classifications) can be translated into lightweight ontologies.

As explained in [7], with the first step the labels at nodes are taken in isolation. Using NLP techniques tuned for short phrases, such as those in [20, 24], their meaning is determined by constructing a corresponding formula (i.e. the *concept at label*). However, since the label alone does not provide enough clues for the disambiguation, all possible senses of the words have to be kept. For instance, the concept at label of

[1] S-Match is an open source tool freely downloadable from http://semanticmatching.org/

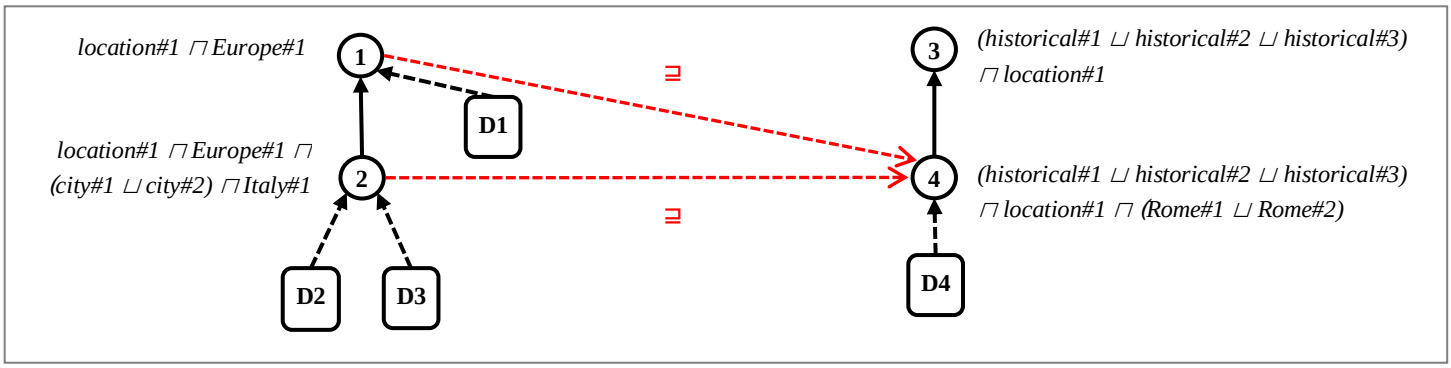

Fig. 2 – The two classifications in Fig. 1 converted into lightweight ontologies with the corresponding alignment

node 2 in Fig. 1 is (city#1 ⊔ city#2 ⊔ city#3) ⊓ Italy#1, where each atomic concept is taken from WordNet[2]:

- **city#1**: city, metropolis, urban center -- (a large and densely populated urban area; may include several independent administrative districts; "Ancient Troy was a great city")
- **city#2**: city -- (an incorporated administrative district established by state charter; "the city raised the tax rate")
- **city#3**: city, metropolis -- (people living in a large densely populated municipality; "the city voted for Republicans in 1994")
- **Italy#1**: Italy, Italian Republic, Italia -- (a republic in southern Europe on the Italian Peninsula; was the core of the Roman Republic and the Roman Empire between the 4th century BC and the 5th century AD)

With the second step, each formula is completed by taking into account the relative position of each node in the classification. This is done by taking the conjunction (⊓) of all the formulas along the path from the root to the node and by filtering out the senses which are not compatible each other, i.e. not related by relations in WordNet. We call this formula the *concept at node*. For instance, to determine the concept at node for node 2 in Fig. 1 we need to consider that for the words *location* and *Europe* the following meanings are provided in WordNet:

- **location#1:** location -- (a point or extent in space)
- **location#2:** placement, location, locating, position, positioning, emplacement -- (the act of putting something in a certain place)
- **location#3:** localization, localisation, location, locating, fix -- (a determination of the place where something is; "he got a good fix on the target")
- **location#4:** location -- (a workplace away from a studio at which some or all of a movie may be made; "they shot the film on location in Nevada")
- **Europe#1:** 1. (28) Europe -- (the 2nd smallest continent (actually a vast peninsula of Eurasia); the British use `Europe' to refer to all of the continent except the British Isles)
- **Europe#2:** European Union, EU, European Community, EC, European Economic Community, EEC, Common Market, Europe -- (an international organization of European countries formed after World War II to reduce trade barriers and increase cooperation among its members; "he took Britain into Europe")
- **Europe#3:** Europe -- (the nations of the European continent collectively; "the Marshall Plan helped Europe recover from World War II")

By further observing that in WordNet only the first and second meaning of *city* are related (through a chain of is-a relations) to the first meaning of *location*, and that the first meaning of *Europe* is related (through part-of) to the only sense available for *Italy*, while all the other senses are unrelated, after the sense filtering the concept at node of node 2 is computed as (location#1 ⊓ Europe#1) ⊓ ((city#1 ⊔ city#2) ⊓ Italy#1).

With the third step, the background knowledge is used to determine the semantic relations holding between all the atomic concepts appearing in the concepts at label in the two classifications. For, instance, it may contain the fact that city#1 ⊑ location#1. In other words, it allows constructing the local theory used to draw our conclusions. This step is fundamental to reduce the number of axioms to reason about when computing the semantic relations between nodes in the last step.

During the last step, the problem of matching the two classifications is decomposed into *n* x *m* node to node matching problems, where *n* and *m* are the sizes of the two classifications. For each pair of nodes, the problem of determining the semantic relation holding between them is reduced to an unsatisfiability problem using the local theory determined at the previous step. Fig. 2 shows the lightweight ontologies obtained from the classifications in Fig. 1 and the matching result, called alignment, between them.

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

Work such that the one described in [50], where S-Match is used to align two different vocabularies in the medicine domain using UMLS[3], prove that - when appropriate domain knowledge is used - precision and recall can be very high.

2.2. Improving matching by managing diversity

In approaching the semantic matching problem and all the problems dealing with semantics in general, we should consider that diversity emerges at least along three main dimensions:

- ***Diversity in natural language***: terms may denote classes (common nouns), entities (proper nouns), properties, qualities and other modifiers (adjectives and adverbs); different terms can be used to denote the same notion (synonymy), e.g. the term *location* in the first classification and the term *place* in the second; the same term may denote different things (polysemy), e.g. the term *bank* in the first classification may mean a sloping land or a financial institution. At the entity level, *Rome* the capital of Italy is also known as the *Eternal City*; there might be different places in the world (and in general different entities) called *Rome*;
- ***Diversity in formal language***: when disambiguated, each term corresponds to a concept written in some formal language. Different classifications, according to their specific scope and purpose, may use different formal languages. For instance, while for somebody it might be enough to distinguish between mountains and rivers, some others may need to further distinguish between mountains and hills, rivers, creeks and rivulets or even between oversea and undersea mountains.
- ***Diversity in knowledge***: at this level the relations between concepts are recognized. The amount of knowledge, in terms of axioms, necessary for a certain task is also a function of the local goals, culture, opinions and personal experience. For instance, while dogs are mainly perceived as pets, they are regularly served as food in China (culture); while someone may consider beautiful the city of Rome in Italy, somebody else may consider it too chaotic (opinion); somebody may consider climate change an urgent problem to be solved, while somebody else may even negate its existence (school of thought).

The ambiguity of natural language is a critical issue in the conversion of classifications into lightweight ontologies. In this respect, it is fundamental to identify resources providing the background knowledge relevant for the disambiguation [8, 9, 10, 11]. However, the meaning of the words and the context of use is almost always left implicit. This implicit knowledge, or *implicit assumptions* (as they have been called in [35]), is what allows their meaning to be determined. In other words, implicit assumptions constitute what is relevant and necessary to disambiguate and understand the labels. It is also quite intuitive and important to note that the amount of implicit knowledge is potentially infinite. As a consequence, it is quite never possible or desirable to completely determine them. A considerable portion of knowledge remains in the human minds [5].

The second problem we should consider is that by increasing the size of the background knowledge, reasoning can become arbitrarily complex. It is therefore fundamental to reduce the number of axioms that we use to reason about during step 4 to only those relevant to determine the semantic relations.

The solution we propose to address these problems can be summarized into three subsequent steps:

1. Develop an extensible diversity-aware knowledge base explicitly codifying the differences in (natural and formal) language and knowledge in multiple *domains* (this topic is discussed in Section 3).
2. Given the specific problem, build the corresponding *context* as a formal local theory by (2.1) determining from the knowledge base the implicit assumptions which are relevant to understand it and (2.2) building the corresponding context as a logical theory (this topic is discussed in Section 4). Here we use the word context with the meaning provided in [3].
3. Solve the problem in context (this step is standard).

Concerning the example in Fig. 1, we need to understand the (natural and formal) language and the knowledge about the *Space* domain, i.e. the terminology used to describe geographical classes (e.g. *location*, *city*, *bank*), places of the world (e.g. *Europe*, *Italy* and *Rome*) and relations between them (cities and banks are locations, Rome is a city). Moreover, we need some basic terminology, e.g. to express general properties (e.g. *historical*) or subjective qualities (e.g. *chaotic*, *beautiful*) that we assume to have same meaning in all domains. It is therefore fundamental to identify appropriate methodologies for the construction and formalization of domains as well as techniques to select at run-time the implicit assumptions which provide the context for the disambiguation as well as the knowledge necessary to solve the problem. To understand the role of domains and context in semantic applications, let us revisit the problem of matching the two classifications in Fig. 1. The three steps above can be mapped into the four steps of the semantic matching as shown in Table 1.

[3] http://www.nlm.nih.gov/research/umls/

Steps for a generic semantic task	Steps in matching
(1) create a diversity-aware knowledge base	
(2) given the problem, build the context	
(2.1) determine the implicit assumptions	1. For all the labels in the two classifications compute the *concept at label* 2. For all the nodes in the two classifications compute the *concepts at node*
(2.2) build the context	3. For all pairs of labels in the two classifications compute the semantic relations between the concept at labels
(3) use the context to solve the problem.	4. For all pairs of nodes in the two classifications compute the relations between the concepts at node

Table 1 – Mapping the semantic matching problem in the general three steps.

3. Domains and facets

The methodology we propose for the construction of domain knowledge is mainly inspired by the *faceted approach*, a well-established technique introduced by the Indian librarian Ranganathan [21] at the beginning of the last century and used with profit in library science for building classificatory structures from atomic concepts which are analyzed into macro-categories and combined by the application of what in jargon is called the system syntax [47]. The methodology is centered on the fundamental notions of *domain* and *facet* as described in the following.

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 provide a bird's eye view of the whole field of knowledge. Domains may include any conventional field of study (e.g., *library science*, *mathematics*, *physics*), applications of pure disciplines (e.g., *engineering*, *agriculture*), any aggregate of such fields (e.g., *physical sciences*, *social sciences*), and they may also capture knowledge about our everyday lives (e.g., *music*, *movie*, *sport*, *Space*, *Time*, *recipes*, *tourism*).

We define a domain as a 5-tuple

$$D = \langle id, FL, K, \{NL\}, \{F_{NL}\}\rangle$$

where:

1. *id* is a string denoting the name of the domain
2. FL is a 4-tuple <C, E, Q, q> where C is a set of classes, E is a set of entities, Q a set of qualities and q is a set of values. These sets are mutually disjoint.
3. K = {is-a, instance-of , part-of, value-of} is a set of binary relations, where
   a. is-a: C → C and Q → Q such that each element of C can be associated to zero or more elements of C and each element of Q can be associated to zero or more elements of Q
   b. instance-of: E → C;
   c. part-of: C → C and E → E such that each element of C can be associated to zero or more elements of C and each element of E can be associated to zero or more elements of E. By construction, we assume the part-of relation to be transitive;
   d. value-of: Q → q such that at each quality in Q associates one or more values in q.
4. Each NL in {NL} is a pair $\langle W_{NL}, S_{NL}\rangle$ where W_{NL} is a set of words in a natural language; $S_{NL} \subseteq P(W_{NL})$ is a subset of the power set of W_{NL} representing the set of all words in W_{NL} which are synonyms in the language and where each element of S_{NL} is called a *synset* (note that the same word may belong to different synsets). We refer to each NL as being a different *vocabulary*.
5. Each F_{NL}: $S_{NL} \rightarrow \{C \cup E \cup Q \cup q\}$ is a mapping function that for each synset $s \in S_{NL}$ in a natural language NL assigns an element in $C \cup E \cup Q \cup q$.

NL (for natural language), FL (for formal language) and K (for knowledge) are the three levels along which a domain is characterized and structured.

Consider the example of domain in Fig. 3. It provides a very simplified version of the *Space* domain. In Fig. 3 nodes denote elements of FL; circles represent classes in C; squares represent entities in E; trapezoids represent qualities in Q; stars represent values in q. Arrows denote relations in K; in particular, solid arrows represent is-a and instance-of relations; dashed arrows represent part-of relations between classes or between entities; dotted arrows represent value-of relations. FL and K are language independent. Each element in $C \cup E \cup Q \cup q$ is associated with a natural language synset from $S_{English}$. Synonyms are separated by a comma. Therefore, in this example we have:

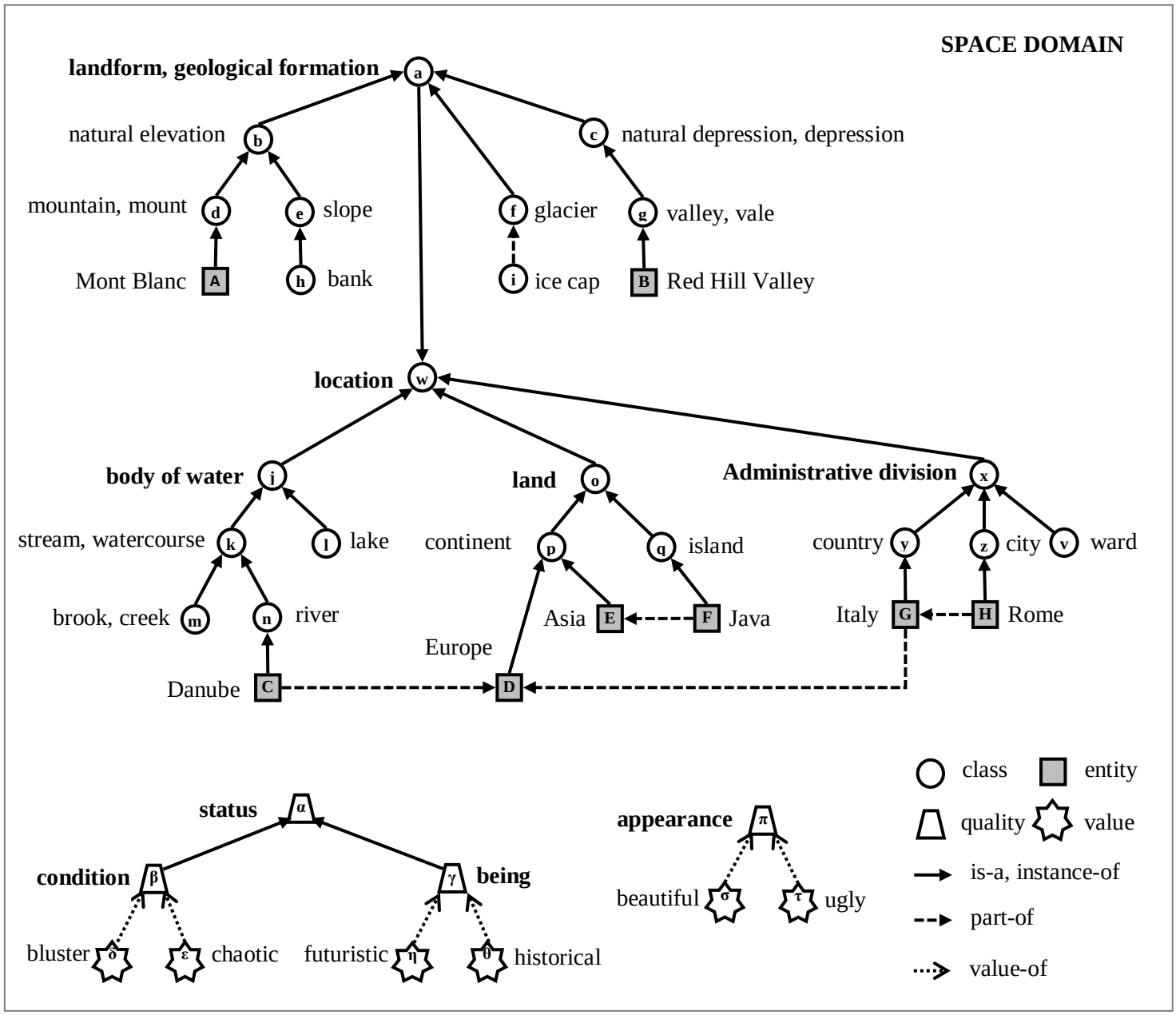

Fig. 3 – An example of domain (the Space domain simplified)

1. *id* = "Space"
2. In FL, C = {a, b, c, …, z}, E = {A, B, C, …, H}, Q = {α, β, γ, π} and q = {δ, ε, η, θ, σ, τ}
3. K includes the relations is-a(m,k), is-a(n,k),is-a(k,i), is-a(i,w), instance-of(C,n), part-of(i,f), part-of(C,D) and value-of(σ, π)
4. For NL = English, W_{NL} = {body of water, stream, watercourse, brook, creek, river, Danube, …} and S_{NL} = {{body of water},{stream, watercourse}, {brook, creek}, {river}, {Danube}, … }
5. The mapping function F_{NL} assigns for instance {body of water} to j, {stream, watercourse} to k and {Danube} to C

The main idea is that a domain provides the natural language terminology used to describe a set of classes and entities as well as their qualities and the relations holding between them in a restricted framework. Different natural languages, for instance English and Italian, can be mapped to the same classes, entities, qualities and values. In other words, while the classes, entities, qualities and values are language independent notions, the way to lexicalize them may vary from one language to another. Furthermore, the same notions might be expressed with different terms by different people speaking the same language. For instance, it is well known that in medicine the terminology used by non-professionals is very different from the one used by domain specialists [49]. For instance, the technical term *myocardial infarction* is commonly referred to as *heart attack*. This generates different vocabularies for the same natural language.

While domains provide a bird's eye view of the whole field of knowledge, facets provide a more detailed view of each of the components in a domain. In fact, a domain can be broken into

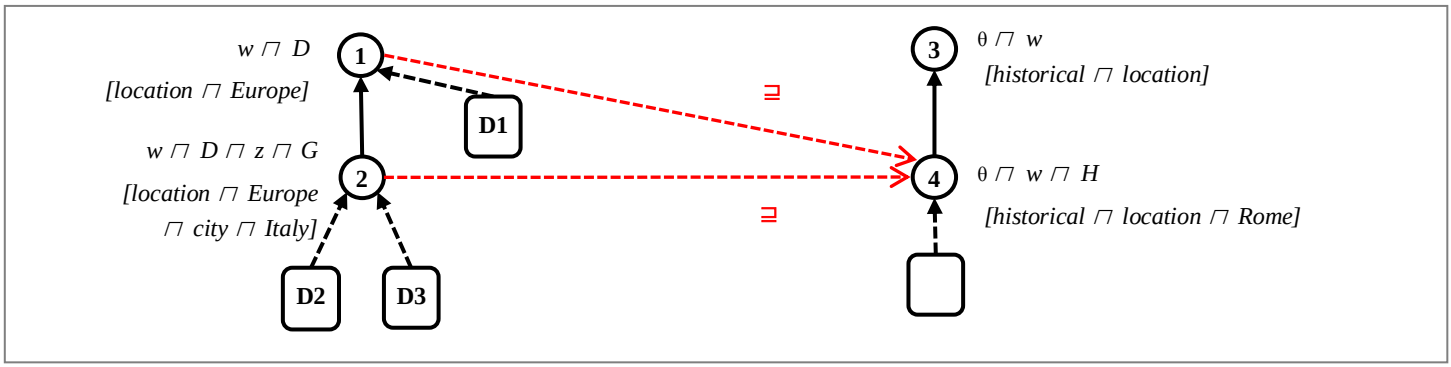

Fig. 4 – The lightweight ontologies constructed by using the Space domain

a number of facets each of them describing a specific aspect of the domain. For instance, in the *Medicine* domain we can distinguish among the *body parts*, the *diseases* that can affect them and the different *treatments* that can be taken to overcome or prevent them. Each of these aspects provides more detailed knowledge. A *facet* can be defined as *a hierarchy of homogeneous terms describing an aspect of the domain, where each term in the hierarchy denotes an atomic concept* [6]. An atomic concept is a class, an entity, a quality or value in the domain. Each hierarchy in Fig. 3, where the root is in bold, is a different facet in the *Space* domain. Each facet can be constituted by sub-facets. For instance, the *status* facet is further constituted by the facets *condition* and *being*.

Two important observations. The first is that domains are the fundamental way by which diversity is captured. In fact, a domain determines the terminology and its meaning (natural and formal language), codifies what is known (knowledge) and reflects the particular subjective view of the modeler (experience). For instance, according to the personal perception and purpose, the *Space* domain may or may not include buildings and man-made structures; the *food* domain may or may not include dogs according to the local customs. The second is that domains also allow scaling as they allow adding new knowledge at any time as needed. For instance, while initially local applications may require only knowledge of the *Space* domain, due to new scenarios, the *Time* and *food* domains might be needed and added. As proved by decades of research in library science, the use of the principles at the basis of the faceted approach guarantees the creation of better quality - in terms of robustness, extensibility, reusability, compactness and flexibility - and easier to maintain domain ontologies (see for instance [21, 30, 47, 48]).

4. Building the context

Following [35], we define a context as a 4-tuple

$$ctx = <id, L_c, K_c, IA>$$

where:

- *id* is an identifier for the context
- L_c is the local (formal) language
- K_c is the local knowledge
- *IA* is a set of implicit assumptions

In the case of semantic matching, implicit assumptions consist of a selection of the domains which are relevant to understanding the meaning of the words in a certain framework. Our baseline algorithm for domain recognition consists of parsing node labels and documents in classifications, linking them to the diversity-aware knowledge base and identifying the smallest set of domains in which words take a precise meaning. Consider again the example in Fig. 1. The analysis of the words appearing in the labels might reveal that the words *location*, *place*, *city* and *bank* (the root form of the words appearing in the labels) denote geographical classes, and that *Europe*, *Italy*, *Rome*, *Milan* and *Danube* are location names in the *Space* domain. Since most of the words assume a precise meaning if interpreted in the *Space* domain we can assume that it can provide most of the implicit assumptions.

The local context *ctx* is built by selecting from the domains the language and the knowledge which are strictly necessary to solve the problem. This corresponds to the third step in the matching and it is done on the basis of the concepts that were used in the formulas at labels. L_c is the set of all atomic concepts in the formulas at labels, while K_c is built by computing the strongest semantic relation holding between each of the concepts in L_c. Our approach is similar to the work described in [59] where the relevant knowledge is constructed by progressively expanding the set of axioms in the premises on the basis of the symbols occurring in the formula. Nevertheless, here the problem is easier given the lower complexity of reasoning (propositional). Moreover, the use of domains further mitigates the problem.

A context is therefore a logical theory over a certain language and domain of interpretation. More precisely, for the problem of matching classifications, the theory is a propositional DL theory. The FL and K of the selected domains are used as follows:

- Classes, entities, qualities and values from FL which are used in the formulas are codified as atomic concepts of the formal language L_c
- All the relations in K correspond to subsumption [6, 33]. [4] For all the concepts in L_c the semantic relations holding between them are selected or computed from K and added to K_c

For instance, given the domain knowledge in Fig. 3 and the lightweight ontologies in Fig. 2, to understand the meaning of node 2 and 4 and compute the strongest semantic relation holding between the two, reported in Fig. 4:

- L_c will include:
  w (location), z (city), D (Europe), G (Italy), H (Rome) and θ (historical).
- K_c will include the selected axioms:
  H ⊑ z (Rome is a city)
  H ⊑ G (Rome is part of Italy)
  G ⊑ D (Italy is part of Europe)
  and the inferred axioms:
  z ⊑ w (cities are locations)
  H ⊑ w (Rome is a location)
  H ⊑ D (Rome is part of Europe).

The semantics associated with the formal language are the classification semantics, therefore an interpretation function I: $L_c \rightarrow P(D)$ assigns each atomic concept in the formal language to a set of documents in D. For instance, the extension of the concept *city* will be the set of documents about real world cities, while the extension of the concept *beautiful* will be the set of documents about real world beautiful objects.

5. Creating a diversity-aware knowledge base

We have been developing a framework and a diversity-aware knowledge base currently covering an initial set of domains necessary for the kinds of scenarios we need to serve, but - in the spirit of the proposed approach - extensible according to the local scope, purpose, language and personal experience.

The expressive power of the representation language of our background knowledge is that of propositional DL with only conjunctions, no negations and no disjunctions. The expressive power we exploit is very low. Still, decades of work in library science and several studies conducted in the context of library classifications show that it is sufficient to describe their labels in terms of conjunctions of atomic concepts [20] representing intersections of sets of documents (see for instance Fig. 4). Furthermore, in an experiment we performed [58] the labels of the classifications considered turn out to have a simple translation into propositional DL with a few "local" disjunctions (around 1% of the overall number of logical connectives) and no negations.

In the following, we briefly describe how we have been progressively building our diversity-aware knowledge base.

5.1. Phase I: bootstrapping the knowledge base

We initially populated our knowledge base with a general terminology imported from WordNet 2.1 and the Italian section of MultiWordNet[5]. This essentially provided what is needed to bootstrap the natural language part, in English and Italian, respectively. For each synset in the two languages, a language-independent concept was created. If the same notion can be expressed in the two languages then corresponding synsets are linked to the same concept in the formal language part. Since MultiWordNet is aligned with the older WordNet 1.6 version, the mapping between the two languages was reconstructed by combining the existing mapping[6] between WordNet 1.6 and 2.0 with another one we created expressly between WordNet 2.0 and 2.1 using some heuristics. Notice that due to the partial coverage of the language in MultiWordNet and the well-known problem of gaps in languages (i.e. given a lexical unit in a language, it is not always possible to identify an equivalent lexical unit in another language) not all concepts have a corresponding synset in Italian. *Hypernym* (is-a) and transitive *part meronym* (part-of) (see [33] for a complete motivation for transitivity) relations were elected as semantic hierarchical relations in the knowledge part. All the other relations were defined as associative relations.

5.2. Phase II: building the Space and Time domains

To fill the knowledge part we need knowledge about specific domains. Given their pervasiveness [21] and the specific scenarios we need to serve (see the work in [34]) we started with *Space* and *Time*.

As first step for constructing the *Space* domain, we have collected several terms representing classes of real world geographical entities, i.e. the atomic concepts. For this we have consulted various sources including GeoNames[7], TGN[8], WordNet and some scientific literature about geography and its related areas. Following this, we have analyzed and synthesized the collected concepts according to the Analytico-Synthetic approach [21]. The purpose of the *analysis* is to enlist the characteristics of division to be used to form the facets. In other words, they are used to form the different levels of abstraction of the conceptual categories. The concepts are analyzed using the topological, geometric or

[4] Note that for the matching problem the value-of relations (described in Section 3, Fig. 3) are not used, but they play an important role in maintenance.

[5] http://multiwordnet.fbk.eu
[6] http://www.cse.unt.edu/~rada/downloads.html#wordnet
[7] http://www.geonames.org/
[8] http://www.getty.edu/research/tools/vocabularies/index.html

geographical characteristics of corresponding entities. We have tried to be exhaustive in their determination. The exhaustiveness (one of the principles we follow) leaves open the possibility to form a huge number of very fine grained groups of concepts. On the other hand, the purpose of *synthesis* is to arrange the concepts into facets by characteristics. 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 what in jargon is known as an *array* of homogeneous and mutually disjoint concepts.

For example, consider the concepts for *river* and *lake* in Fig. 3. The primary characteristic of *river* and *lake* is that both are bodies of water. Since they share the same characteristic and are disjoint, both of them are categorized in the same array under *body of water*. However, we further enlisted their characteristics as follows:

- *river* is a flowing body of water; has no fixed boundary; is confined within a bed and stream banks; is larger than a brook
- *lake* is a stagnant body of water and has fixed geographical boundary.

This sort of detailed list of concept characteristics not only helps in distinguishing them but also allows identifying the more general categories. In the complete facet, provided in [2], under the root concept *body of water*, we identified two broad categories, i.e. *stagnant body of water* and *flowing body of water*. Now for instance, if we need to include the new concept *pond* - with characteristics a stagnant body of water and smaller than a lake - into the system, we can easily extend the facet by adding it under *stagnant body of water*. This shows that the facets at the array level are exhaustive enough to accommodate new concepts.

This process led to the creation of a set of facets containing overall more than 1000 concepts (still increasing in size) and to the release of the open source geo-spatial ontology GeoWordNet[9] [32]. However, in that release we did not explicitly provide the facets. Conversely, the concepts and relations constituting them were rather merged with WordNet. This domain has been proven to be effective in real scenarios, for instance to enable semantic search on geo-catalogues [34]. See the appendix for a small snapshot of the *Space* domain that we have developed.

Similarly to *Space*, the *Time* domain was built by using WordNet and Wikipedia[10] as main sources and arranging identified concepts by common characteristics. For instance, *holidays* are grouped *by religion*. *Christian holydays* include *Easter* and *Christmas*; *Islamic holidays* include *Ramadan* and *Muharram*.

It is important to note that, unlike the original Analytico-Synthetic approach, inside the facets generated following our methodology we explicitly distinguish between classes, entities, qualities and values, and we explicitly provide the is-a, instance-of , part-of and value-of relations between them. In other words, the facets produced by the Analytico-Synthetic approach correspond to what in [6] we call *classification ontologies*, i.e. ontologies built for the purpose of classifying documents. Conversely, we produce *descriptive ontologies*, i.e. ontologies built to describe a domain. In [33] (modulo the value-of relations) an algorithm to automatically translate the latter into the former is provided.

5.3. Phase III: populate the knowledge base with entities

The next step was to import in the knowledge base around 7 million locations. This was done semi-automatically. Around 600,000 additional locations as well as 700,000 persons and 150,000 organizations are currently been imported from YAGO [37]. For this we are experimenting fully automatic techniques. Detailed statistics about the current size of the knowledge base can be found in Table 2. Overall, by including the attributes associated to the entities, it contains more than 80 million axioms.

Object	**Quantity**
Natural language part	
English synsets	110,609
English words	147,252
Italian synsets	33,356
Italian words	45,156
Formal language part	
Classes, qualities and values	110,609
Entities	~9.5 millions
Domains	2 (*Space* and *Time*)
Classes, qualities and values in the domains	>1000 (*Space*), >200 (*Time*)
Knowledge part	
is-a and part-of relations	204,481
instance-of relations	~9.5 millions

Table 2 – Detailed statistics about the current size of the knowledge base.

5.4. Next steps: building the Internet domains

Our long term goal is not to build the world knowledge, but to identify those domains which are more likely to play a role in everyday life and in particular on the Web. In the context of the Living Knowledge EU project[11], this has been identified as

[9] http://geowordnet.semanticmatching.org/
[10] http://www.wikipedia.org/
[11] http://livingknowledge-project.eu/

strategic towards enabling diversity aware applications for the Web. From a preliminary analysis on the query logs of the AOL search engine[12] conducted by our partners at the Indian Statistical Institute[13] in Bangalore, a prioritized list of around 350 domains was formed. On the very top of this list we find domains such as *Space*, *Time*, *food*, *sports*, *tourism*, *music*, *movie* and *software*. We refer to them as *Internet domains*.

Some of these domains are either finalized or under development. They are used and being evaluated in some concrete scenarios. In this regard, we have established several industrial collaborations. The first involves the SORA Institute[14], a company based in Austria specialized in statistical surveys conducted using media content analysis techniques. With them we have been developing the *political science* domain [60]. The domain is meant to be used for annotation and retrieval in a faceted search facility. The second collaboration involves Telecom Italia, a well-known telecommunication company based in Italy. We have been developing the *food* domain with them. The domain is being constructed by revising and extending an existing ontology used in the Piemonte project[15] where a mobile application allows users to run queries and navigate across entities in the domain (e.g. wines and their producers). A further collaboration with the Province of Trento aims at customizing the *Space* domain for local needs. Here a simplified version of the *Space* domain is used within the discovery service of a geo-catalogue to expand user queries [34].

6. Related work

6.1. The notion of context

Based on two different approaches, the first formal theories on context were proposed by McCarthy [13] and Giunchiglia [3].

According to McCarthy, contexts are a way to *partition* knowledge into a limited set of locally true axioms with common assumptions. This set of axioms should be at the right level of abstraction thus excluding irrelevant details in order to simplify local reasoning as much as possible. This is known as the generality principle [12]. In this setting, it is always possible to *lift* from the local context to a more general one by progressively making explicit the assumptions. This allows, among other things, *integrating* two or more contexts under the umbrella of a more general theory, thus assuming that a unique global schema can be always reconstructed. This process is called *relative decontextualization*. CYC is an example of knowledge base following this approach (see for instance [14]) as a way to partition huge quantities of common sense knowledge into smaller, easier to manage, sets of axioms.

According to Giunchiglia, context is a tool to specifically *localize* reasoning to a subset of facts known by an agent [15]. This is motivated by the intuition that reasoning is always local and always represents a partial approximate theory and subjective view of the world. Unlike McCarthy, in this view each context typically has its own language, grammar and theory thus leading to the maximum level of local autonomy. Moreover, the existence of a common global schema is not guaranteed. However, taking into account implicit assumptions, it might be possible to (partially) *relate compatible* axioms in distinct contexts [4]. These relations are the basis for interoperability.

6.2. Methodologies for the construction and maintenance of domain knowledge

In traditional libraries, fully faceted classification systems like the Colon Classification (CC) [45], the Bibliographic Classification (BC) [46] and partially faceted classifications system like the Universal Decimal Classification[16] (UDC) are very popular as knowledge organization systems. They have been used for several decades as knowledge organization tools in libraries for classifying and shelving books and documents in general.

We share with these systems the key notion that facets allow domain knowledge modeling by exploiting and making explicit the different aspects of knowledge within the domain. As facet based systems, they proved their usefulness and effectiveness in organizing and searching documents in conventional library systems [21, 30, 47, 48]. However, the major drawback of these systems lies in their structure. All these systems fail to make explicit the way the meaning (semantics) of subjects (what the document is about) is built starting from the semantics of their constituents. In fact, they only consider the syntactic form by which subjects are described in natural language (syntax). Consequently, they do not allow for a direct translation of their elements - terms and arcs in the facets - into a formal language, e.g. in form of DL axioms. They do not explicitly specify the taxonomical *is-a* and *instance-of* (genus/species) and mereological *part-of* (whole/part) relations between the classes thus limiting their applicability. However, given their purpose (classifying books) this is not strictly required [6, 33]. Though, making them explicit is a fundamental step towards automation and interoperability.

6.3. Knowledge bases and approaches followed for their construction

[12] http://search.aol.com/aol/webhome
[13] http://drtc.isibang.ac.in/DRTC/
[14] http://www.sora.at/
[15] http://www.piemonte.di.unito.it/
[16] http://www.udcc.org/

In the last thirty years many projects have aimed at constructing knowledge bases. DENDRAL [39] is widely considered the first expert system ever created embedding a knowledge base with domain specific knowledge (organic chemistry). We can divide knowledge bases into two main broad categories: (a) automatically built and (b) hand-crafted.

Among the projects aiming at automatic extraction of knowledge (mainly unary and binary predicates) from free-text we can mention for instance KnowItAll [40] and TextRunner [41]. However, since working in open scenarios is extremely difficult, these techniques typically achieve limited accuracy. For this reason, projects like DBPedia [42] and YAGO [36] that extract information from semi-structured knowledge sources (mainly Wikipedia infoboxes and categories) obtain more accurate results. In particular, while in general these systems lack explicit quality control systems and semantics, in YAGO this is achieved through an explicit quality control mechanism mainly based on a unique entity reference system (there cannot be two entities with the same name) and type checking routines on the domain and range of the predefined binary predicates. Moreover, in YAGO there is a precise knowledge representation model based on RDFS[17]. In its 2009 version[18], it contains around 2.5 million entities and 20 million facts.

Among hand-crafted resources it is worth mentioning CYC [37] that is a general-purpose common sense knowledge base containing around 2.2 million assertions and more than 250,000 terms about the real world. Its open source version OpenCYC contains 306,000 assertions and 47,000 terms. Organized according to the generality principle [12], the content of CYC is distributed into three levels from broader and abstract knowledge (the upper ontology) and widely used knowledge (the middle ontology) to domain specific knowledge (the lower ontology). Similarly to CYC, SUMO (Suggested Upper Merged Ontology) [57] is a free formal ontology of about 1,000 terms and 4,000 definitional statements. Its extension, called MILO (MId-Level Ontology), covers individual domains, comprising overall 21,000 terms mapped with WordNet and 73,000 axioms. Both SUMO and MILO are therefore quite small. Unlike our knowledge base, CYC, SUMO and their extensions are built by not targeting any particular range of reasoning tasks.

Neither in DBPedia nor in YAGO there is an explicit notion of domain. Everything is codified in terms of generic facts between entities (triples of the form source-relation-target). Notice that both in DBPedia and YAGO the entities include what we further differentiate into entities, classes, qualities and values. However, both have the disadvantage that their different released versions are not aligned, i.e. there is no direct way to map the same fact or entity in different versions. In CYC there is a notion of domain, but it is used only to partition knowledge into easier to manage components. Moreover, in CYC, too, there is a generic notion of entity.

Even if not specifically developed for supporting reasoning tasks, WordNet [38] - as demonstrated by the thousands of citations - is the most widely used linguistic resource nowadays. This is mainly due to the fact that it is manually constructed and exhibits a significant quality and size. For this reason it is also frequently adapted for semantic applications. However, one of its main drawbacks is that it is not tailored for any particular domain. Moreover, it is often considered too fine grained to be really useful in practice (see [43]). Multilingual extensions of WordNet include MultiWordNet and EuroWordNet[19].

Other valuable resources can be found in digital library communities, especially as regards domain specific knowledge encoded in informal or semi-formal knowledge organization systems such as subject headings and thesauri. For instance, about agriculture we can mention AGROVOC[20] and NALT[21]; about medicine the most widely known is UMLS. As already mentioned, in general their main drawback is the lack of an explicit semantics [52].

Hand-crafted resources are surely more accurate but difficult to construct and maintain. To alleviate this problem, some recent projects like Freebase [44] follow a collaborative approach by leveraging on volunteers to fill the knowledge base. Here the main focus is on named entities. Freebase however, does not make any effort to guarantee consistency in the use of the terminology and leaves its users *free* to independently define their axioms without enforcing effective mechanisms for duplicate detection or quality control.

Our knowledge base settles somewhere in between the two approaches. It is built by importing knowledge from existing resources such as GeoNames and YAGO, but a significant amount of manual work is done to guarantee high quality of the data (see for instance [32]). Moreover, domain knowledge is manually built by experts in library science following a precise methodology and guiding principles.

By comparing it with respect to pre-existing systems, our knowledge base has at least the following distinctive features, summarized in Table 3, that none of them has together:

- There is a clear split between natural language, formal language and knowledge
- There is an explicit definition of domain as a way to codify knowledge which is local to a community thus

[17] http://www.w3.org/RDF/
[18] http://www.mpi-inf.mpg.de/yago-naga/yago/downloads_yago.html
[19] http://www.illc.uva.nl/EuroWordNet
[20] http://aims.fao.org/website/AGROVOC-Thesaurus/sub
[21] http://agclass.nal.usda.gov/

Knowledge base	#entities	#facts	Domains	Distinction concepts/instances	Distinction NL/FL	Manually built	Framework included
YAGO	2.5 M	20 M	No	No	No	No	No
CYC	250,000	2.2 M	Yes	No	No	Yes	No
OpenCYC	47,000	306,000	Yes	No	No	Yes	No
SUMO	1,000	4,000	No	Yes	Yes	Yes	No
MILO	21,000	74,000	Yes	Yes	Yes	Yes	No
DBPedia	3.5 M	500 M	No	No	No	No	No
Freebase	22 M	unknown	No	Yes	No	Yes	No
Our knowledge base	10 M	80 M	Yes	Yes	Yes	Yes	Yes

Table 3 – Comparison of existing knowledge bases in terms of support to diversity

reflecting their specific purpose, needs, competences, beliefs and personal experience

- There is an explicit distinction between classes, entities, qualities and values
- It is completely modular, in the sense that it can be continuously extended with knowledge about new domains and new vocabularies
- Domain knowledge is created following a precise methodology and principles inspired by well-established library science methodologies and practices
- Domain knowledge is used to construct the context formalized (given the specific tasks we want to serve) as a propositional DL theory and therefore the complexity of reasoning is limited to propositional reasoning
- It does not only consist of a data repository, but it comes with a framework to support a precise set of basic semantic tasks including natural language understanding, automatic classification, semantic matching and semantic search by encoding knowledge in the most appropriate semantics according to the task at hand [33].

As we can see from the table, we can consider the combination of SUMO plus MILO as the closest in spirit to our approach.

However, as a drawback, in order to guarantee the high quality of the knowledge, its construction and maintenance requires a significant amount of manual work. In fact, building a domain may take several weeks of work by an expert familiar with the classical faceted approach and the novelties introduced with our methodology. For instance, bootstrapping the *Space* [2, 32] and *Time* domain - that, given their pervasiveness, are among the biggest ones - took around 6 and 2 man months, respectively. However, other domains should take much less. We plan to overcome this issue by adopting crowdsourcing techniques integrated with a certification pipeline based on ideas already exploited on ESP games [53]. Given the precise split that we enforce between concepts and instances, we plan to establish two pipelines: the first for experts at the purpose of defining the basic terminology of domains, in terms of classes, relations and qualities (the TBox); the second for generic users at the purpose of providing actual data for the entities (the ABox). The main reason for this distinction is that the first requires a higher level of expertise. At this purpose, in the context of the Living Knowledge project we already conducted some training activities with our partners at the Indian Statistical Institute where some library science students were asked to use our methodology for the construction of sample domains. Notice how the second pipeline will have to be able to manage a quantity of knowledge which is several orders of magnitude bigger that the first. When possible, given format and quality of the data, ready-made entities can be directly imported from existing sources. This is for instance what we did for the population of the *Space* domain [32, 34] and we are currently experimenting with YAGO.

7. Conclusions and future work

In this paper, by observing that the lack of background knowledge represents one of the main obstacles towards the success of semantics, we have stressed the necessity for a very large virtually unbound knowledge base able to capture the diversity of the world as well as to reduce the complexity of reasoning at run-time.

We have proposed the *faceted approach*, a well-established methodology centered on the fundamental notions of *domain* and *facet* and practiced with success in library science for the organization of knowledge in libraries, as an effective methodology for its construction and maintenance. The knowledge base that we have been developing can be seen as a proof of the applicability of the proposed approach. It is completely modular since at any moment it allows plugging an arbitrary number of domains and facets with corresponding classes, entities, qualities and values as well as vocabularies in different languages and for different communities. Its usefulness, in particular in the *Space* domain, has been already proven in real scenarios.

By taking semantic matching as an example, we have also shown how, in order to speed up reasoning, it is fundamental to select at run-time from the knowledge base - on the basis of the implicit assumptions - the relevant part of the domains for the construction of the appropriate context to be used to solve the semantic task. However, a general solution to the problem is still open and requires further research.

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. We want to express our gratitude to all the people working with us in the KnowDive group at University of Trento (http://disi.unitn.it/~knowdive/) for their contribution in the creation and population of the knowledge base and in particular to Ilya Zaihrayeu and Marco Marasca for their contribution to the definition of the data structures, Feroz Farazi for the data importing, Abdelhakim Freihat for importing the Italian MultiWordNet and Anand Pandey for the construction of the *Time* domain. We are also thankful to A.R.D. Prasad, D. P. Madalli and their research team at the Indian Statistical Institute in Bangalore for the research on *Internet domains* and their contribution in the development of domains.

References

1. Artic Climate Impact Assessment, Cambridge University Press, 2005, pp. 973.
2. B. Dutta, F. Giunchiglia, V. Maltese, A facet-based methodology for geo-spatial modelling, GEOS (2011).
3. F. Giunchiglia, Contextual reasoning, Epistemologica - Special Issue on I Linguaggi e le Macchine, 16 (1993), 345–364.
4. C. Ghidini, F. Giunchiglia, Local Model Semantics, or Contextual Reasoning = Locality + Compatibility, Artificial Intelligence, 127 2 (2001), 221–259.
5. L. Prusak, Knowledge in Organizations, Cap. 7: The tacit dimension by M. Polanyi, 1997.
6. 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).
7. F. Giunchiglia, M. Marchese, I. Zaihrayeu, Encoding Classifications into Lightweight Ontologies, Journal of Data Semantics 8 (2006), 57-81. Short version in: ESWC (2006).
8. F. Giunchiglia, P. Shvaiko, M. Yatskevich, Discovering missing background knowledge in ontology matching, European Conference on Artificial Intelligence ECAI (2006), 382–386.
9. B. Lauser, G. Johannsen, C. Caracciolo, J. Keizer, W. R. van Hage, P. Mayr, Comparing human and automatic thesaurus mapping approaches in the agricultural domain, Int'l Conf. on Dublin Core and Metadata Applications (2008).
10. P. Shvaiko, J. Euzenat. Ten Challenges for Ontology Matching, 7th Int. Conference on Ontologies, Databases, and Applications of Semantics, ODBASE, (2008).
11. Z. Aleksovski, W. ten Kate, F. van Harmelen, Using multiple ontologies as background knowledge in ontology matching, ESWC workshop on collective semantics (2008).
12. J. McCarthy, Generality in artificial intelligence, Communications of ACM 30 (1987), 1030–1035.
13. J. McCarthy, Notes on formalizing context, Bajcsy, R. (Ed.), Thirteenth International Joint Conference on Artificial Intelligence, IJCAI (1993), 555-560.
14. R. Guha, D. Lenat, Context dependence of representations in cyc, Colloque ICO (1993).
15. P. Bouquet, C. Ghidini, F. Giunchiglia, E. Blanzieri, Theories and uses of context in knowledge representation and reasoning. Journal of Pragmatics 35 3 (2003).
16. P. Shvaiko, J. Euzenat, Ontology Matching, Springer-Verlag, New York, Inc. Secaucus, NJ, USA (2007).
17. F. Giunchiglia, M. Yatskevich, E. Giunchiglia, Efficient semantic matching, European Semantic Web Conference ESWC (2005).
18. F.Giunchiglia, M. Yatskevich, P. Shvaiko, Semantic Matching: algorithms and implementation, Journal on Data Semantics IX (2007).
19. V. Maltese, F. Giunchiglia, A. Autayeu, Save up to 99% of your time in mapping validation, 9th International Conference on Ontologies, DataBases, and Applications of Semantics ODBASE (2010).
20. A. Autayeu, F. Giunchiglia, P. Andrews, Lightweight parsing of classifications into lightweight ontologies, ECDL (2010), 327-339
21. S. R. Ranganathan, Prolegomena to library classification, Asia Publishing House (1967)
22. N. E. Fuchs, K. Kaljurand, G. Schneider, Attempto controlled english meets the challenges of knowledge representation, reasoning, interoperability and user interfaces, FLAIRS Conference (2006), 664-669
23. R. Schwitter, M. Tilbrook, Lets talk in description logic via controlled natural language, LENLS (2006)
24. I. Zaihrayeu, L. Sun, F. Giunchiglia, W. Pan, Q. Ju, M. Chi, X. Huang, From web directories to ontologies: Natural language processing challenges, 6th International Semantic Web Conference ISWC (2007).

25. F. Giunchiglia, I. Zaihrayeu, U. Kharkevich, Formalizing the get-specific document classification algorithm, 11th European Conference on Research and Advanced Technology for Digital Libraries ECDL (2007).
26. F. Giunchiglia, M. Marchese, I. Zaihrayeu. Encoding classifications into lightweight ontologies. European Semantic Web Conference ESWC (2006).
27. P. Bouquet, L. Serafini, S. Zanobini, Semantic coordination: a new approach and an application, 2nd International Semantic Web Conference ISWC (2003).
28. B. Magnini, L. Serafini, M. Speranza, Making explicit the semantics hidden in schema models, Workshop on Human Language Technology for the Semantic Web and Web Services held at ISWC (2003).
29. F. Giunchiglia, U. Kharkevich, I. Zaihrayeu, Concept search, European Semantic Web Conference ESWC (2009).
30. V. Broughton, The need for a faceted classification as the basis of all methods of information retrieval, Aslib Proceedings 58 1/2 (2006), 49-72.
31. L. Spiteri, A Simplified Model for Facet Analysis, Journal of Information and Library Science 23 (1998), 1-30.
32. F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta, GeoWordNet: a resource for geo-spatial applications, Extended Semantic Web Conference ESWC (2010).
33. V. Maltese, F. Farazi, Towards the Integration of Knowledge Organization Systems with the Linked Data Cloud, UDC seminar (2011).
34. F. Farazi, V. Maltese, F. Giunchiglia, A. Ivanyukovich, A faceted ontology for a semantic geo-catalogue, Extended Semantic Web Conference ESWC (2011).
35. F. Giunchiglia, Managing Diversity in Knowledge, Invited Talk at the European Conference on Artificial Intelligence ECAI, Lecture Notes in Artificial Intelligence 2006.
36. F. M. Suchanek, G. Kasneci, G. Weikum, YAGO: A Large Ontology from Wikipedia and WordNet, Journal of Web Semantics (2011).
37. C. Matuszek, J. Cabral, M. Witbrock, J. DeOliveira, An introduction to the syntax and content of Cyc, AAAI Spring Symposium (2006).
38. C. Fellbaum, WordNet: An Electronic Lexical Database, MIT Press (1998).
39. B. G. Buchanan, J. Lederberg, The Heuristic DENDRAL program for explaining empirical data, Stanford University, technical report (1971).
40. O. Etzioni, M. J. Cafarella, D. Downey, S. Kok, A. Popescu, T. Shaked, S. Soderland, D. S. Weld, A. Yates, Web-scale information extraction in KnowItAll, WWW conference (2004).
41. M. Banko, M. J. Cafarella, S. Soderland, M. Broadhead, O. Etzioni, Open information extraction from the web, IJCAI conference (2007).
42. S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, C. Cyganiak, Z. Ives, DBpedia: A Nucleus for a Web of Open Data, 6th International Semantic Web Conference ISWC (2007).
43. R. Mihalcea, D. I. Moldovan, Automatic generation of a coarse grained wordnet, NAACL Workshop on WordNet and Other Lexical Resources (2001).
44. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, J. Taylor, Freebase: a collaboratively created graph database for structuring human knowledge, ACM SIGMOD international conference on Management of data (2008), 1247-1250.
45. S. R. Ranganathan, The Colon Classification, Rutgers Series on Systems for the Intellectual Organization of Information, S. Artandi (etd.), IV, Graduate School of Library Science, Rutgers University, New Brunswick, NJ, 1965.
46. J. Mills, V. Broughton, Bliss Bibliographic Classification, 2nd ed. London: Butterworth and Bowker-Saus, 1977.
47. V. Broughton, Building a Faceted Classification for the Humanities: Principles and Procedures, Journal of Documentation (2007).
48. L. Spiteri, A Simplified Model for Facet Analysis, Journal of Information and Library Science 23 (1998), 1-30.
49. A. Tse, D. Soergel, Procedures for mapping vocabularies from non-professional discourse: a case study: in 'consumer medical vocabulary', Annual Meeting of the American Society for Information (2003).
50. J. Shamdasani, T. Hauer, P. Bloodsworth, A. Branson, M. Odeh, R. McClatchey, Semantic Matching Using the UMLS, European Semantic Web Conference ESWC (2009).
51. B. Magnini, M. Speranza, C. Girardi, A semantic-based approach to interoperability of classification hierarchies: Evaluation of linguistic techniques, COLING (2004).
52. D. Soergel, B. Lauser, A. Liang, F. Fisseha, J. Keizer, S. Katz, Reengineering thesauri for new applications: the agrovoc example, Journal of Digital Information 4 (2004).
53. L. Von Ahn. Games With A Purpose, IEEE Computer Magazine (2006), 96-98.
54. G. Adorni, M. Di Manzo, F. Giunchiglia, Natural Language Driven Image Generation, 10th International Conference on Computational Linguistics COLING (1984), 495-500.
55. F. Giunchiglia, T. Walsh, Abstract Theorem Proving, 11th International Joint Conference on Artificial Intelligence IJCAI (1989), 372-377.
56. F. Giunchiglia, A. Villafiorita. T. Walsh, Theories of Abstraction, AI Communications 10 3/4 (1997), 167-176.
57. A. Pease, G. Sutcliffe, N. Siegel, S. Trac, Large theory reasoning with SUMO at CASC, AI Communications, 23 2-3 (2010) 137–144.
58. F. Giunchiglia, I. Zaihrayeu, F. Farazi, Converting classifications into owl ontologies, Artificial Intelligence and Simulation of Behaviour Convention Workshop on Matching and Meaning (2009).
59. K. Hoder, A. Voronkov, Sine qua non for large theory reasoning, International Conference on Automated Deduction (2011).
60. D. P. Madalli, A.R.D. Prasad, Analytico synthetic approach for handling knowledge diversity in media content analysis, UDC seminar (2011).

Appendix: A snapshot of the Space Domain

Geological formation

- Natural elevation
  - Continental elevation
    - Mountain
    - Hill
    - …
  - Oceanic elevation
    - Seamount
    - Oceanic rise
    - …
  - Icecap ridge
  - Icecap dome
- Natural depression
  - Continental depression
    - Valley
    - Pass
    - Cliff
    - …
  - Oceanic depression
    - Oceanic valley
    - Submarine pass
    - Oceanic hole
    - …
  - Icecap depression
- Sill
- Cordillera
- Asphalt lake
- Cave
- Shore
  - Seashore
    - Littoral
    - Seaside
    - …
- Beach
  - Plague
- Delta
- Foreshore
- Ice mass
  - Icecap
  - Continental glacier
    - Polar glacier
- Submarine fracture zone
- Submarine plain
  - Submarine flat
    - Oceanic platform

Body of water

- Ocean
- Sea
  - Bay
- Bight
- Gulf
- Inlet
  - Cove
- Flowing body of water
  - Stream
    - River
      - Lost river
    - Brook
      - Brooklet
      - Tidal brook
    - …
  - Channel
    - Watercourse
      - Abandoned watercourse
    - Navigation channel
    - Marine channel
    - …
  - Waterfall
    - Cataract
    - Cascade
  - …
- Stagnant body of water
  - Lake
    - Lagoon
    - Glacial lake
    - Crater lake
    - …
  - Chain of lakes
  - Pond
    - Fishpond
    - Millpond
    - …
  - …

Populated place

- City
- Town
- Village
- …