Department of
Information Engineering
and Computer Science
DISI

UNIVERSITY
OF TRENTO - Italy

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

A FACETED ONTOLOGY FOR A SEMANTIC GEO-CATALOGUE

Feroz Farazi, Vincenzo Maltese,
Fausto Giunchiglia and
Alexander Ivanyukovich

December 2010

Technical Report # DISI-10-061

Also: in proceedings of the 8th Extended Semantic Web Conference (ESWC) 2011

A faceted ontology for a semantic geo-catalogue

Feroz Farazi[1], Vincenzo Maltese[1], Fausto Giunchiglia[1], Alexander Ivanyukovich[2]

[1] DISI - Università di Trento, Trento, Italy
[2] Trient Consulting Group S.r.l., Trento, Italy

Abstract. Geo-spatial applications need to provide powerful search capabilities to support users in their daily activities. However, discovery services are often limited by only syntactically matching user terminology to metadata describing geographical resources. We report our work on the implementation of a geographical catalogue, and corresponding semantic extension, for the spatial data infrastructure (SDI) of the Autonomous Province of Trento (PAT) in Italy. We focus in particular to the semantic extension which is based on the adoption of the S-Match semantic matching tool and on the use of a faceted ontology codifying geographical domain specific knowledge. We finally report our experience in the integration of the faceted ontology with the multi-lingual geo-spatial ontology GeoWordNet.

Keywords: Semantic geo-catalogues, faceted ontologies, ontology integration, entity matching

1 Introduction

Geo-spatial applications need to provide powerful search capabilities to support users in their daily activities. This is specifically underlined by the INSPIRE[1] directive and regulations [15, 16] that establish minimum criteria for the *discovery services* to support search within the INSPIRE metadata elements. However, discovery services are often limited by only syntactically matching user terminology to metadata describing geographical resources [1]. This weakness has been identified as one of the key issues for the future of the INSPIRE implementation [11, 17, 18, 19].

As a matter of fact, current geographical standards only aim at syntactic agreement [23]. For example, if it is decided that the standard term to denote a harbour (defined in WordNet as "*a sheltered port where ships can take on or discharge cargo*") is *harbour*, they will fail in applications where the same concept is denoted with *seaport.* As part of the solution, domain specific geo-spatial ontologies need do be adopted. In [14] we reviewed some of the existing frameworks supporting the creation and maintenance of geo-spatial ontologies and proposed GeoWordNet - a multi-lingual geo-spatial ontology providing knowledge about geographic classes (features), geo-spatial entities (locations), entities' metadata and part-of relations between them - as one of the best candidates, both in terms of quantity and quality of the information provided, to provide semantic support to the spatial applications.

The purpose of the Semantic Geo-Catalogue (SGC) project [20] - promoted by the Autonomous Province of Trento (PAT) in Italy with the collaboration of Informatica

[1] http://inspire.jrc.ec.europa.eu/

Trentina, Trient Consulting Group and the University of Trento - was to develop a semantic geo-catalogue as an extension of the existing geo-portal of the PAT. It was conceived to support everyday activities of the employees of the PAT. The main requirement was to allow users to submit queries such as *Bodies of water in Trento*, run them on top of the available geographical resources metadata and get results also for more specific features such as *rivers* and *lakes*. This is clearly not possible without semantic support. As reported in [12], other technological requirements directly coming from the INSPIRE directives included (a) *performance* - send one metadata record within 3s. (this includes, in our case, the time required for the semantic expansion of the query); (b) *availability* - service up by 99% of the time; (c) *capacity* - 30 simultaneous service requests within 1s.

In this paper we report our work on the implementation of the semantic geographical catalogue for the SDI of the PAT. In particular, we focus on the semantic extension of its discovery service. The semantic extension is based on the adoption of the S-Match[2] semantic matching tool [4] and on the use of a specifically designed faceted ontology [2] codifying the necessary domain knowledge about geography and including *inter-alia* the administrative divisions (e.g., municipalities, villages), the bodies of water (e.g., lakes, rivers) and the land formations (e.g., mountains, hills) of the PAT. Before querying the geo-resources, user queries are expanded by S-Match with domain specific terms taken from the faceted ontology. In order to increase the domain coverage, we integrated the faceted ontology with GeoWordNet.

The rest of the paper is organized as follows. Section 2 describes the overall system architecture and focuses on the semantic extension in particular. Section 3 describes the dataset containing the locations within the PAT and how we cleaned it. Sections 4, 5 and 6 provide details about the construction of the faceted ontology, its population and integration with GeoWordNet, respectively. The latter step allows supporting multiple languages (English and Italian), enlarging the background ontology and increasing the coverage of locations and corresponding metadata such as latitude and longitude coordinates. Finally Section 7 concludes the paper by summarizing the main findings and the lessons learned.

2 The architecture

As described in [1], the overall architecture is constituted by the front-end, business logic and back-end layers as from the standard three-tier paradigm. The geo-catalogue is one of the services of the existing geo-cartographic portal[3] of the PAT. It has been implemented by adapting available open-source tool[4] conforming to the INSPIRE directive and by taking into account the rules enforced at the national level. Following the best practices for the integration of the third-party software into the BEA ALUI framework[5] (the current engine of the geo-portal), external services are brought together using a portlet[6]-based scheme, where GeoNetwork is used as a back-end. Fig. 1

[2] S-Match is open source and can be downloaded from http://sourceforge.net/projects/s-match/
[3] http://www.territorio.provincia.tn.it/
[4] GeoNetwork OpenSource, http://geonetwork-opensource.org
[5] http://download.oracle.com/docs/cd/E13174_01/alui/
[6] http://jcp.org/en/jsr/detail?id=168

provides an integrated view of the system architecture. At the front-end, the functionalities are realized as three portlets for:

1. ***metadata management***, including harvesting, search and catalogue navigation functionalities;
2. ***user/group management***, to administer access control on the geo-portal;
3. ***system configuration***, which corresponds to the functionalities of the GeoNetwork's Administrator Survival Tool (GAST) tool of GeoNetwork.

These functionalities are mapped *1-to-1* to the back-end services of GeoNetwork. Notice that external applications, such as ESRI ArcCatalog, can also access the back-end services of GeoNetwork.

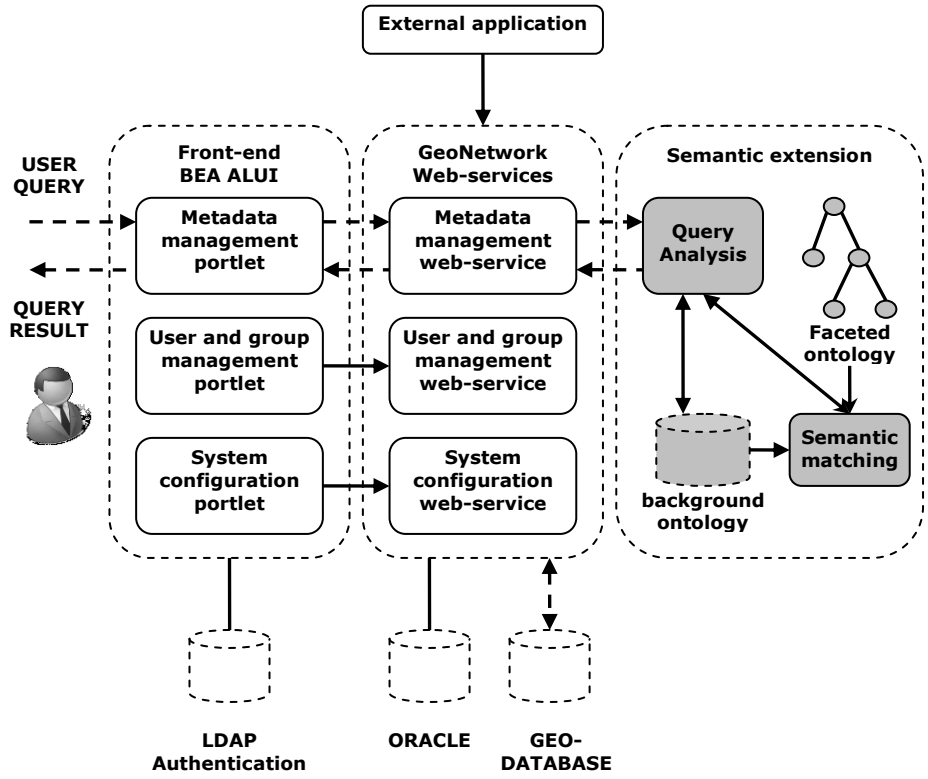

Fig. 1. The overall system architecture

The GeoNetwork catalogue search function was extended by providing ***semantic query processing*** support. In particular, the analysis of the available work in the field, such as [6, 7, 8, 9, 10, 11], summarized in [12], brought to the selection of the S-Match *semantic matching operator* as the best candidate to provide the semantic extension of the geo-catalogue. Given two graph-like structures (e.g., XML schemas) a

semantic matching operator identifies the pairs of nodes in the two structures that are semantically similar (equivalent, less or more specific), where the notion of semantic similarity is both at the node level and at the structure level [13, 21, 22]. For instance, it can identify that two nodes labelled *stream* and *watercourse* are semantically equivalent because the two terms are synonyms in English. This allows similar information to be identified that would be more difficult to find using traditional information retrieval approaches.

Initially designed as a standalone application, S-Match was integrated with GeoNetwork. As explained in [1], this was done through a wrapper that provides web services to be invoked by GeoNetwork. This approach mitigates risks of failure in experimental code while still following strict uptime requirements of the production system. Another advantage of this approach is the possibility to reuse this service in other applications with similar needs.

In order to work properly, S-Match needs domain specific knowledge. Providing this knowledge is the main contribution of this paper. Knowledge about the geographical domain is codified into a faceted ontology [1]. A faceted ontology is an ontology composed of several subtrees, each one codifying a different aspect of the domain. In our case, it codifies the knowledge about geography and includes (among others) the administrative divisions (e.g., municipalities, villages), the bodies of water (e.g., lakes, rivers) and the land formations (e.g., mountains, hills) of the PAT.

The flow of information, starting from the user query to the query result, is represented with arrows in Fig. 1. Once the user enters a natural language query (which can be seen as a classification composed by a single node), the query analysis component translates it into a formal language according to the knowledge codified in the background ontology[7]. The formal representation of the query is then given as input to the semantic matching component that matches it against the faceted ontology, thus expanding the query with domain specific terms. The expanded query is then used by the metadata management component to query GeoNetwork and finally access the maps in the geo-database.

At the moment the system supports queries in Italian through their translation in English, uses S-Match to expand feature classes and translates them back to Italian. For instance, in the query *Bodies of water in Trento* only *Bodies of water* would be expanded. Future work includes extended support for Italian and the semantic expansion of the entities such as *Trento* into its (administrative and topological) parts.

3 Data extraction and filtering

The first step towards the construction (Section 4) and population (Section 5) of the faceted ontology was to analyze the data provided by the PAT, extract the main geographical classes and corresponding locations and filter out noisy data. The picture below summarizes the main phases, described in detail in the next paragraphs.

[7] S-Match uses WordNet by default but it can be easily substituted programmatically, for instance by plugging GeoWordNet at its place.

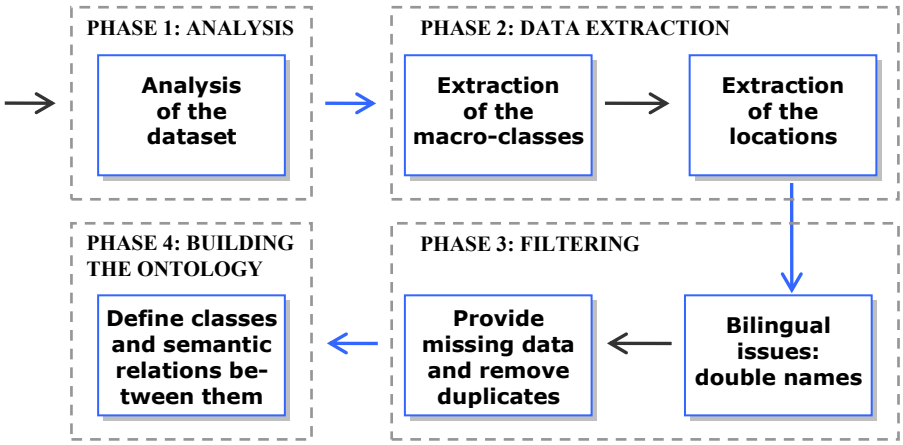

Fig. 2. A global view of the phases for the dataset processing

3.1 The dataset of the Autonomous Province of Trento

The data are available in MS Excel files (Table 1), and are gathered from the PAT administration. The *features* file contains information about the main 45 geographical classes; the *ammcom* file contains 256 municipalities; the *localita* file contains 1,507 wards and ward parts, that we generically call populated places; the *toponimi* file contains 18,480 generic locations (including *inter-alia* villages, mountains, lakes, and rivers). *Comune*, *frazione* and *località popolata* are the Italian class names for municipality, ward and populated place respectively.

FILE NAME	DESCRIPTION
features.xls	It provides the name of the feature classes.
ammcom.xls	It provides the name, id, latitude and longitude of the municipalities.
localita.xls	It provides the name, id, latitude and longitude of the wards and ward parts (that we map to populated places). It also provides the id of the municipality a given ward or ward part belongs to.
toponimi.xls	It provides the name, id, class, latitude and longitude of the locations. It also provides the ids of the ward or ward part and municipality a given generic location belongs to.

Table 1. The names and descriptions of the files containing PAT data

With the construction of the faceted ontology we identified a suitable name for the rest of the Italian class names from the analysis of the PAT geographical classes in the *features* file. In fact, they are very generic as they are meant to contain several, but similar, kinds of locations. For instance, there is a class that includes springs, waterfalls and other similar entities.

3.2 Extracting the macro-classes

We retrieved the main PAT classes, that we call **macro-classes** (as they group different types of locations), from the *features* file. In this file each class is associated an id (e.g., P110) and an Italian name (e.g., *Monti principali*).

CODE	ENGLISH NAME	ITALIAN NAME
E000	province	provincia
E010	municipality	comune
E020	ward	frazione
E021	populated place	località popolata

Table 2. Names of the administrative classes

We did not process the macro-class with id P310 (*Regioni limitrofe*) as it represents locations in the neighbouring region of Trento (out of the scope of our interest) and P472 (*Indicatori geografici*) as it represents geographic codes. Notice that names of the macro-classes needed to be refined as they are too generic and represent many kinds of locations grouped together. As this file lacks classes for the provinces, municipalities, wards and populated places, we created them as shown in Table 2.

3.3 Extracting the locations

We imported all the locations into a temporary database by organizing them into the part-of hierarchy *province* > *municipality* > *ward* > *populated place* (and other location kinds) as follows:

- ***The province level***. We created an entity representing the Province of Trento. This entity is not explicitly defined in the dataset but it is clearly the root of the hierarchy. We assigned the following names to it: *Provincia Autonoma di Trento*, *Provincia di Trento* and *Trento*. It was assigned to the *province* class.
- ***The municipality level***. Municipalities were extracted from the *ammcom* file. We created an entity for each municipality and a part-of relation between each municipality and the province. They were assigned to the *municipality* class.
- ***The ward and populated place level***. Wards and populated places (sections of wards) were extracted from the *localita* file. Here each ward is connected to the corresponding municipality and each populated place to the corresponding ward by specific internal codes. For each ward and populated place we created a corresponding entity. Using the internal codes, each ward was connected to the corresponding municipality and each populated place to the corresponding ward. They were assigned to the class *ward* or *populated place* accordingly.
- ***All other locations***. All other (non administrative) locations were extracted from the *toponimi* file. Here each of them is connected either to a municipality, a ward

or a populated place by specific internal codes. Using the internal codes, we connected them accordingly. A few of them are not connected to any place and therefore we directly connected them to the province. Each location was temporary assigned to the corresponding macro-class.

Locations are provided with latitude and longitude coordinates in Cartesian WGS84 (World Geodetic System 1984) format, a standard coordinate reference system mainly used in cartography, geodesy and navigation to represent geographical coordinates on the Earth[8]. Since in GeoWordNet we store coordinates in WGS84 decimal format, for compatibility we converted them accordingly.

3.4 Double names: bilingual issues

Locations are provided with a name and possibly some alternative names. A few names are double names, e.g., *Cresta di Siusi Cresta de Sousc*. The first (*Cresta di Siusi*) is in Italian and the second (*Cresta de Sousc*) is in Ladin. Ladin is a language spoken in a small part of Trentino and other Alpine regions. The combination of the two names is the official name of the location in the PAT.

In order to identify these cases, the PAT provided an extra text file for each municipality containing individual Italian and Ladin version of the names. In the temporary database, we put the Italian and Ladin names as alternative names. These extra files also contain additional name variants, which are also treated as alternative names. In the end, we found 53 additional Italian names, 53 Ladin names and 8 name variants. For instance, for the location *Monzoni*, the Ladin name *Monciogn* and the name variant *Munciogn (poza)* are provided.

3.5 Provide missing data and remove duplicates

While importing the entities in the temporary database, we found that 8 municipalities and 39 wards were missing in the *ammcom* and *localita* files respectively, and 35 municipalities were duplicated in the *ammcom* file[9].

KIND OF OBJECT	NUMBER OF THE OBJECTS IMPORTED
macro-classes	44
locations	20,162
part-of relations	20,161
alternative names	7,929

Table 3. Objects imported in the temporary database

[8] https://www1.nga.mil/ProductsServices/GeodesyGeophysics/WorldGeodeticSystem/

[9] Note that the missing municipalities are due to the fact that they were merged with other municipalities on 1st January 2010, while the duplicates are related to administrative islands (regions which are not geometrically connected to the main area of each municipality).

We automatically created the missing locations and eliminated the duplicates. At the end of the importing we identified the objects reported in Table 3. Notice that here by part-of we mean a generic containment relation between locations. It can be administrative or topological containment.

4 Building the faceted ontology

As mentioned in the previous section, the macro-classes provided by the PAT are very generic and are meant to contain several different, but similar, kinds of locations. This is mainly due to the criteria used by PAT during categorization that where based not only on type but also on importance and population criteria. With the two-fold goal of refining them and determine the missing semantic relations between them, we analyzed the class names and created a multi-lingual faceted ontology. Our goal was to create an ontology that both reflected the specificity of the PAT and respected the canons of the analytico-synthetic approach [5] for the generation of a faceted ontology. A faceted (lightweight) ontology [2] is an ontology divided into subtrees, called facets, each encoding a different dimension or aspect of the domain knowledge. As a result, it can be seen as a collection of hierarchies. The ontology we built encodes the domain knowledge specific to the geographical scope of the PAT and it is therefore suitable for its application in the geo-portal of the PAT administration.

4.1 From macro-classes to atomic concepts

We started from the 45 macro-classes extracted from the *feature* file that we imported in the temporary database. Notice that they are not accompanied by any description. Therefore, analyzing the locations contained in each macro-class, each macro-class was manually disambiguated and refined - split, merged or renamed - and as a result new classes had to be created.

MACRO-CLASSES	CLASSES
P410 Capoluogo di Provincia	Province
P465 Malghe e rifugi	Shelter Farm Hut
P510 Antichita importanti P520 Antichita di importanza minore	Antiquity
P210 Corsi dacqua/laghi (1 ord.) P220 Corsi dacqua/laghi (2 ord.) P230 Corsi dacqua/Canali/Fosse/Cond. forz./Laghi (3 ord.) P240 Corsi dacqua/Canali/Fosse/Cond. forz./Laghi (>3 ord.-25.000) P241 Corsi dacqua/Canali/Fosse/Cond. forz./Laghi (>3 ord.)	Lake Group of lakes Stream River Rivulet Canal

Table 4. Examples of mapping from macro-categories to atomic concepts

This was done through a statistical analysis. Given a macro-class, corresponding locations were searched in GeoWordNet. We looked at all the locations in the part-of hierarchy rooted in the Province of Trento having same name and collected their classes. Only a little portion of the locations where found, but they were used to understand the classes corresponding to each macro-class. Some classes correspond to more than one macro-class. The identified classes were manually refined and some of them required a deeper analysis (with open discussions).

At the end of the process we generated 39 refined classes, including the class *province*, *municipality*, *ward* and *populated place* previously created. Each of these classes is what we call an atomic concept. Some examples are provided in Table 4. They represent examples of 1-to-1, 1-to-many, many-to-1 and many-to-many mappings respectively.

4.2 Arrange atomic concepts into hierarchies

By identifying semantic relations between atomic concepts and following the analytico-synthetic approach we finally created the faceted ontology of the PAT with five distinct facets: *antiquity*, *geological formation* (further divided into *natural elevation* and *natural depression*), *body of water*, *facility* and *administrative division*. As an example, below we provide the *body of water* and *geological formation* facets.

Body of water (Idrografia)
- Lake (Lago)
- Group of lakes (Gruppo di laghi)
- Stream (Corso d'acqua)
  - River (Fiume)
  - Rivulet (Torrente)
- Spring (Sorgente)
- Waterfall (Cascata)
  - Cascade (Cascatina)
- Canal (Canale)

Geological formation (Formazione geologica)
- **Natural elevation (Rilievo naturale)**
  - Highland (Altopiano)
  - Hill (Collina, Colle)
  - Mountain (Montagna, Monte)
  - Mountain range (Catena montuosa)
  - Peak (Cima)
  - Chain of peaks (Catena di picchi)
  - Glacier (Ghiacciaio, Vedretta)
- **Natural depression (Depressione naturale)**
  - Valley (Valle)
  - Mountain pass (Passo)

5 Populating the faceted ontology

Each location in the temporary database was associated a macro-class. The faceted ontology was instead built using the atomic concepts generated from their refinement. In order to populate the faceted ontology, we assigned each location in the temporary database to the corresponding atomic concept by applying some heuristics based on the entity names. They were mainly inspired by the statistical analysis discussed in the previous section. As first step, each macro-class was associated to a facet. Macro-classes associated to the same facet constitute what we call a block of classes. For instance, the macro-classes from P110 to P142 (11 classes) correspond to the *natural*

elevation block, including *inter-alia* mountains, peaks, passes and glaciers. Facet specific heuristics were applied to each block.

For instance, entities with name starting with *Monte* were considered as instances of the class *montagna* in Italian (*mountain* in English), while entities with name starting with *Passo* were mapped to the class *passo* in Italian (*pass* in English). The general criterion we used is that if we could successfully apply a heuristic we classified the entity in the corresponding class otherwise we choose a more generic class, which is the root of a facet (same as the block name) in the worst case. For some specific macro-classes we reached a success rate of 98%. On average, about 50% of the locations were put in a leaf class thanks to the heuristics.

Finally, we applied the heuristics beyond the boundary of the blocks for further refinement of the instantiation of the entities. The idea was to understand whether, by mistake, entities were classified in the wrong macro-class. For instance, in the *natural depression* block (the 5 macro-classes from P320 to P350), 6 entities have name starting with *Monte* and therefore they are supposed to be mountains instead. The right place for them is therefore the *natural elevation* facet. In total we found 48 potentially bad placed entities, which were checked manually. In 41.67% of the cases it revealed that the heuristics were valid, in only 8.33% of the cases the heuristics were invalid and the rest were unknown because of the lack of information available on the web about the entities. We moved those considered valid in the right classes.

6 Integration with GeoWordNet

With the previous step the locations in the temporary database were associated to an atomic concept in the faceted ontology. The next step consisted in integrating the faceted ontology and corresponding locations with GeoWordNet.

6.1 Concept integration

This step consisted in mapping atomic concepts from the faceted ontology to GeoWordNet concepts. While building GeoWordNet, we integrated GeoNames classes with WordNet by disambiguating their meaning manually [14]. This time we utilized the experience we gathered in the previous work to automate the disambiguation process with a little amount of manual intervention. An atomic concept from the faceted ontology might or might not be available in GeoWordNet. If available we identified the corresponding concept, otherwise we selected its most suitable parent. This can be done using the Italian or the English name of the class. In our case we used the Italian version of the name. The procedure is as follows:

1. ***Identification of the facet concepts.*** For each facet, the concept of its root node is manually mapped with GeoWordNet. We call it the *facet concept*.
2. ***Concept Identification.*** For each atomic concept C in the faceted ontology, check if the corresponding class name is available in GeoWordNet. If the name is available, retrieve all the candidate synsets/concepts for it. We restrict to noun senses only. For each candidate synset/concept check if it is more specific than

the facet concept. If yes, select it as the concept for C. If none of the concepts is more specific than the facet concept, parse the glosses of the candidate synsets. If the facet name is available in any of the glosses, select the corresponding candidate synset/concept as the concept of C.

3. ***Parent Identification.*** If the class name starts with either "group of" or "chain of", remove this string from the name and convert the remaining part to the singular form. Identify the synset/concept of the converted part. The parent of the identified concept is selected as the parent of the class. If the class name consists of two or more words, take the last word and retrieve its synset/concept. Assign this concept as the parent of the atomic concept corresponding to the class. If neither the concept nor the parent is identified yet, ask for manual intervention.

6.2 Entity matching

Two partially overlapped entity repositories, the temporary database built from the PAT dataset (corresponding to the populated faceted ontology) and GeoWordNet, were integrated. The PAT dataset overall contains 20,162 locations. GeoWordNet already contains around 7 million locations from all over the world, including some locations of the PAT. We imported all but the overlapping entities from the temporary database to GeoWordNet. In order to detect the duplicates we experimented with different approaches. The entity matching task was accomplished within/across the two datasets. We found that the following rules led to a satisfactory result; two entities match if:

Rule 1: name, class and coordinates are the same
Rule 2: name, class, coordinates and parent are the same
Rule 3: name, class, coordinates, parent, children and alternative names are the same

As it can be noticed, Rule 2 is an extension of Rule 1 and Rule 3 is an extension of Rule 2. Parent and children entities are identified using the part-of relation. For example, *Povo* is part-of *Trento*. We have found the following results:

1. ***Within GeoWordNet.*** Applying Rule 1 within GeoWordNet we found 15,665 matches. We found 12,112 matches using Rule 2. Applying Rule 3 we found 12,058 matches involving 22,641 entities. By deleting duplicates these entities can be reduced to 10,583 entities. In fact, if two (or more) entities match by Rule 3 we can safely reduce them by deleting one and keeping the other. Matching entities are clearly undistinguishable.
2. ***Within the temporary PAT database.*** There are 20,162 locations in the PAT dataset. Applying Rule 1 and Rule 2 we found 12 matches and 11 matches, respectively. The result did not change by applying Rule 3 as all of the matched entities are leaves and they have either the same or no alternative name. In total 22 entities matched and we could reduce them to 11.
3. ***Across the two datasets.*** By applying Rule 1 we found only 2 exact matches between the PAT dataset and GeoWordNet, which is far smaller than the number

we expected. Therefore, we checked the coordinates of the top level administrative division Trento of the PAT in these two databases manually. We found it with different, nevertheless, very close coordinates. In fact the coordinates were found as (46.0704, 11.1207) in GeoWordNet and (46.0615, 11.1108) in the PAT dataset, which motivated us to allow a tolerance while matching. The result is reported in Table 5. At the end the last one was applied, leading to 174 matches. It corresponds to Rule 3 with an offset of +/- 5.5 Km. We checked most of them manually and they are undistinguishable.

Same name	Same class	Same coordinates	Same parent	Same Children
1385	1160	2 (exact match)	0	0
		11 (using the offset +/-0.0001)	0	0
		341 (using the offset +/-0.001)	13	12
		712 (using the offset +/-0.01)	65	60
		891 (using the offset +/-0.05)	194	174

Table 5. Matching coordinates with tolerance

Note that while matching classes across datasets, we took into account the subsumption hierarchy of their concepts. For example, Trento as *municipality* in the PAT dataset is matched with Trento as *administrative division* in GeoWordNet because the former is more specific than the latter. Note also that the heuristic above aims only at minimizing the number of duplicated entities but it cannot prevent the possibility of still having some duplicates. However, further relaxing it would generate false positives. For instance, by dropping the condition of having same children we found 5% (1 over 20) of false matches.

6.3 Entity integration

With this step non overlapping locations and part-of relations between them were imported from the temporary database to GeoWordNet following the macro steps below:

1. For each location:
   a. Create a new entity in GeoWordNet
   b. Use the main name of the location to fill the name attribute both in English and Italian
   c. For each Italian alternative name add a value to the name attribute in Italian
   d. Create an instance-of entry between the entity and the corresponding class concept
2. Create part-of relations between the entities using the part-of hierarchy built as described in Section 3.3
3. Generate an Italian and English gloss for each entity created with previous steps

Note that natural language glosses were automatically generated. We used several rules, according to the language, for their generation. For instance, one in English is:

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

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

7 Conclusions

We briefly reported our experience with the geo-catalogue integration into the SDI of the PAT and in particular with its semantic extension. S-Match, initially designed as a standalone application, was integrated with GeoNetwork. S-Match performs a semantic expansion of the query using a faceted ontology codifying the necessary domain knowledge about geography of the PAT. This allows identifying information that would be more difficult to find using traditional information retrieval approaches. Future work includes extended support for Italian and the semantic expansion of the entities such as *Trento* into its (administrative and topological) parts.

In this work we have also dealt with data refinement, concept integration through parent or equivalent concept identification, ontology population using a heuristic-based approach and finally with entity integration through entity matching. In particular, with the data refinement, depending on the cases, most of the macro-classes needed to be split or merged so that their equivalent atomic concepts or parents could be found in the knowledge base used (GeoWordNet in our case). We accomplished the splitting/merging task manually supported by a statistical analysis, while the integration with the knowledge base was mostly automatic. Working on the PAT macro-classes helped in learning how to reduce manual work in dealing with potentially noisy sources. Entity integration was accomplished through entity matching, which was experimented within and across the entity repositories. The entity matching criteria that perform well within a single repository might need to expand or relax when the comparison takes place across the datasets. Note that entity type specific matchers might be necessary when dealing with different kinds of entities (e.g., persons, organizations, events).

Acknowledgements

This work has been partially supported by the TasLab network project funded by the European Social Fund under the act n° 1637 (30.06.2008) of the Autonomous Province of Trento, by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231126 LivingKnowledge: LivingKnowledge - Facts, Opinions and Bias in Time and by "Live- Memories - Active Digital Memories of Collective Life" funded by the Autonomous Province of Trento. We are thankful to our colleagues of the Informatica Trentina and in particular to Pavel Shvaiko for the fruitful discussions on the implementation of the geo-catalogue within the geo-portal of the Autonomous Province of Trento. We acknowledge Aliaksandr Autayeu for his support for the integration of S-Match. We are grateful to Veronica Rizzi for her technical support within the SGC project and to Biswanath Dutta for his suggestions for the creation of the faceted ontology. Finally, we want to thank Daniela

Ferrari, Giuliana Ucelli, Monica Laudadio, Lydia Foess and Lorenzo Vaccari of the PAT for their kind support.

References

1. P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, F. Farazi. A semantic geo-catalogue implementation for a regional SDI. In Proc. of the INPSIRE Conference, 2010.
2. 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.
3. F. Giunchiglia, I. Zaihrayeu. Lightweight Ontologies. The Encyclopedia of Database Systems, 2007
4. F. Giunchiglia, A. Autayeu, J. Pane. S-Match: an open source framework for matching lightweight ontologies. The Semantic Web journal, 2010.
5. S. R. Ranganathan. Prolegomena to library classification. Asia Publishing House (1967).
6. I. Cruz and W. Sunna. Structural alignment methods with applications to geospatial ontologies. Transactions in Geographic Information Science, 12(6):683–711, 2008.
7. J. Euzenat and P. Shvaiko. Ontology matching. Springer, 2007.
8. K. Janowicz, M. Wilkes, and M. Lutz. Similarity-based information retrieval and its role within spatial data infrastructures. In Proc. of GIScience, 2008.
9. P. Maué. An extensible semantic catalogue for geospatial web services. Journal of Spatial Data Infrastructures Research, 3:168–191, 2008.
10. K. Stock, M. Small, Y. Ou, and F. Reitsma. OGC catalogue services - OWL application profile of CSW. Technical report, Open Geospatial Consortium, 2009.
11. L. Vaccari, P. Shvaiko, and M. Marchese. A geo-service semantic integration in spatial data infrastructures. Journal of Spatial Data Infrastructures Research, 4:24–51, 2009.
12. P. Shvaiko, L. Vaccari, G. Trecarichi. Semantic Geo-Catalog: A Scenario and Requirements. In Proc. of the 4th workshop on Ontology Matching at ISWC, 2009.
13. F. Giunchiglia, F. McNeill, M. Yatskevich, J. Pane, P. Besana, and P. Shvaiko. Approximate structure-preserving semantic matching. In Proc. of ODBASE, 2008.
14. F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta. GeoWordNet: a resource for geo-spatial applications. In the Proc. of ESWC, 2010.
15. European Parliament, "Directive 2007/2/EC establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)", 2009.
16. European Commission, "COMMISSION REGULATION (EC) No 976/2009 implementing Directive 2007/2/EC as regards the Network Services," 2009.
17. M. Lutz, N. Ostlander, X. Kechagioglou, and H. Cao. Challenges for Metadata Creation and Discovery in a multilingual SDI - Facing INSPIRE. In Proc. of ISRSE, 2009.
18. J. Crompvoets, M. Wachowicz, F. de Bree, and A. Bregt. Impact assessment of the INSPIRE geo-portal. In Proc. of the 10th EC GI&GIS workshop, 2004.
19. P. Smits and A. Friis-Christensen. Resource discovery in a European Spatial Data Infrastructure. Transactions on Knowledge and Data Engineering, 19(1):85–95, 2007.
20. A. Ivanyukovich, F. Giunchiglia, V. Rizzi, V. Maltese. SGC: Architettura del sistema. Technical report, TCG/INFOTN/2009/3/D0002R5, 2009.
21. F. Giunchiglia, A. Villafiorita. T. Walsh, "Theories of Abstraction". AI Communications, IOS Press, Vol 10, n.3/4, pp. 167-176, 1997.
22. F. Giunchiglia, T. Walsh, "Abstract Theorem Proving". Proceedings of the 11th International Joint Conference on Artificial Intelligence (IJCAI'89), pp. 372-377, 1989.
23. Kuhn, W.: Geospatial semantics: Why, of What, and How? Journal of Data Semantics (JoDS) III, pp. 1–24 (2005)