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

EXTENDING A GEO-CATALOGUE WITH MATCHING CAPABILITIES

Feroz Farazi, Vincenzo Maltese, Biswanath Dutta and Alexander Ivanyukovich

May 2011

Technical Report # DISI-11-464

Also: At the IJCAI Workshop on Discovering Meaning On the Go in Large Heterogeneous Data (LHD 2011).

Extending a geo-catalogue with matching capabilities

Feroz Farazi
DISI
University of Trento
farazi@disi.unitn.it

Vincenzo Maltese
DISI
University of Trento
maltese@disi.unitn.it

Biswanath Dutta
DISI
University of Trento
bisu@disi.unitn.it

Alexander Ivanyukovich
Trient Consulting Group S.r.l.
Trento, Italy
a.ivanyukovich@trientgroup.it

Abstract

To achieve semantic interoperability, geo-spatial applications need to be equipped with tools able to understand user terminology that is typically different from the one enforced by standards. In this paper we summarize our experience in providing a semantic extension to the geo-catalogue of the Autonomous Province of Trento (PAT) in Italy. The semantic extension is based on the adoption of the S-Match semantic matching tool and on the use of a specifically designed faceted ontology codifying domain specific knowledge. We also briefly report our experience in the integration of the ontology with the geo-spatial ontology GeoWordNet.

1 Introduction

To be effective, geo-spatial applications need to provide powerful search capabilities to their users. On this respect, the INSPIRE[1] directive and regulations [EU Parliament, 2009; EU Commission, 2009] establish minimum criteria for the *discovery services* to support search within INSPIRE metadata elements. However, such services are often limited to only syntactically matching user queries to metadata describing geographical resources [Shvaiko *et al*., 2010]. In fact, current geographical standards tend to establish a fixed terminology to be used uniformly across applications thus failing in achieving semantic interoperability. 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 as *seaport.*

As part of the solution, domain specific geo-spatial ontologies need to be adopted. Unfortunately, existing geo-spatial ontologies are limited in coverage and quality [Giunchiglia *et al*., 2010b]. This motivated the creation of GeoWordNet[2] - a multi-lingual geo-spatial ontology providing knowledge about geographic classes, geo-spatial entities (locations), entities' metadata and part-of relations between them. It represents a significant improvement w.r.t. the state of the art, both in terms of quantity and quality of the knowledge provided. As such, it currently constitutes the best candidate to provide semantic support to geo-spatial applications.

One of the purposes of the Semantic Geo-Catalogue (SGC) project [Ivanyukovich *et al*., 2009] - promoted by the PAT - was to extend the geographical catalogue of the PAT with semantic search capabilities. 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.

In this paper we report our work in providing full support for semantic search to the geo-catalogue of the PAT. This was mainly achieved by integrating in the platform the S-Match[3] semantic matching tool [Giunchiglia *et al*., 2010a] and by adopting a specifically designed faceted ontology [Giunchiglia *et al*., 2009] 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. To increase the domain coverage of both resources, we integrated the faceted ontology with GeoWordNet. We conducted an evaluation of the proposed approach to show how simple queries can be semantically expanded using the tool.

The rest of this paper is organized as follows. Section 2 describes the overall system architecture by focusing on the semantic extension. 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, enlarging the background ontology and increasing the coverage of locations and corresponding metadata such as latitude and longitude coordinates. Section 7 provides an evaluation showing the effectiveness of the proposed approach. Section 8 provides a generalization of

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

[2] A significant part of GeoWordNet is in RDF and freely available at http://geowordnet.semanticmatching.org/

[3] Freely available at http://sourceforge.net/projects/s-match/

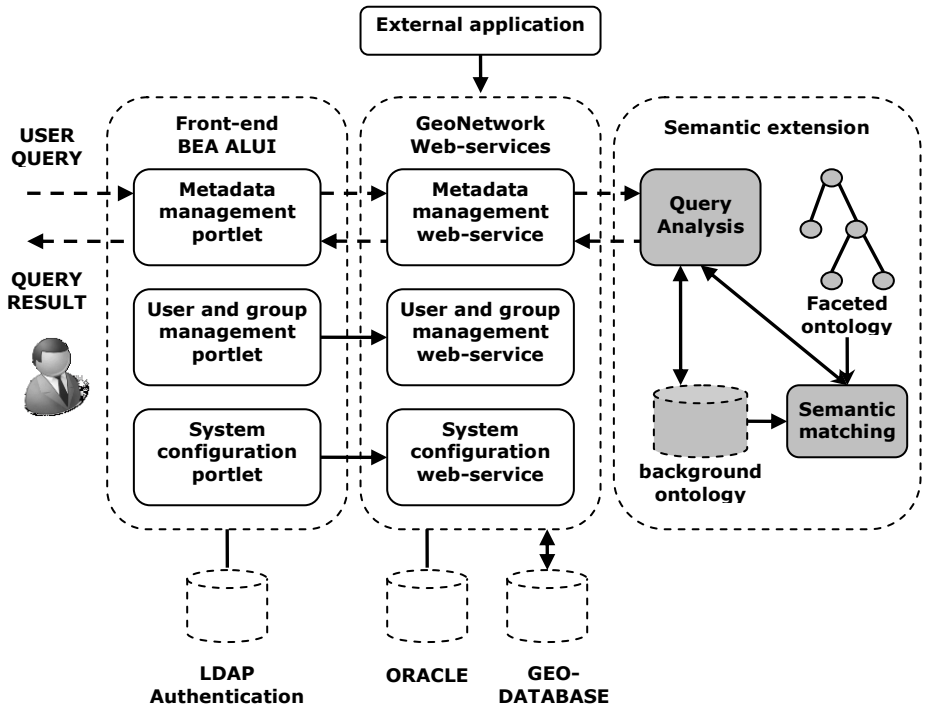

Fig. 1 – The architecture of the semantic geo-catalogue

the work done for the design of the faceted ontology of the PAT in the direction of a faceted ontology for the whole world. Section 9 concludes the paper.

2 The architecture of the geo-catalogue

The overall architecture is constituted by the front-end, business logic and back-end layers as from the standard three-tier paradigm [Shvaiko *et al.*, 2010]. The geo-catalogue is one of the services of the existing geo-cartographic portal[4] of the PAT. It has been implemented by adapting available open-source tool[5] conforming to the INSPIRE directive and 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[6] (the current engine of the geo-portal), external services are brought together using a portlet[7]-based scheme, where GeoNetwork is used as a back-end. Fig.1 provides an integrated view of the system architecture. At the front-end, the functionalities are realized as three portlets for:

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

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

The GeoNetwork catalogue search function was extended by providing ***semantic query processing*** support. To provide this support we used the S-Match open source *semantic matching operator*. Given two graph-like structures semantic matching operators identify 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 [Giunchiglia *et al.*, 2008]. For instance, it can identify that two nodes labeled *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 [Shvaiko *et al.*, 2010], 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. Knowledge about the geographical domain is codified into a faceted ontology. A faceted ontology is an ontology composed of several sub-trees, each codifying a different aspect of the given domain. In our case, it 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.

[4] http://www.territorio.provincia.tn.it/
[5] GeoNetwork Open Source, http://geonetwork-opensource.org
[6] http://download.oracle.com/docs/cd/E13174_01/alui/
[7] http://jcp.org/en/jsr/detail?id=168

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 with a single node), the query analysis component translates it into a formal language according to the knowledge in the background ontology[8]. 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 web-service component to query GeoNetwork and finally access the maps in the database.

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.

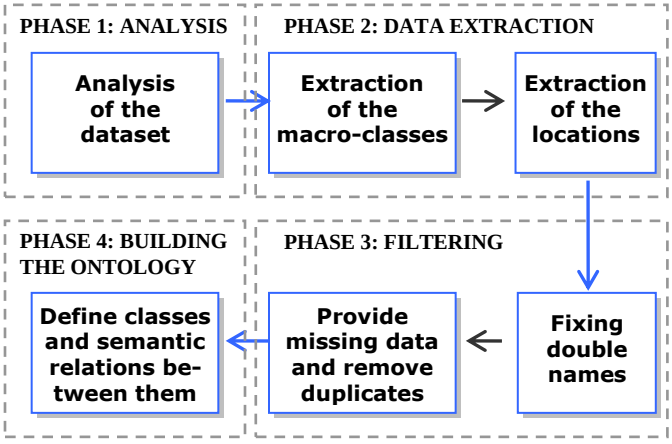

Fig. 2 – The phases for the dataset processing

The dataset of the PAT

The data are available in four files and are gathered from the PAT administration. The *features* file contains 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 18480 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.

Data extraction

We retrieved the PAT classes, that we call **macro-classes**, from the *features* file. Each class is associated an id (e.g., P110) and an Italian name (e.g., *Monti principali*). Names of the macro-classes need 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 manually created them.

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). The entity representing the Province of Trento is not explicitly defined in the dataset but it is clearly the root of the hierarchy, so we manually created it. A few locations from the files 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[9]. Since in GeoWordNet we store coordinates in WGS84 decimal format, for compatibility we converted them accordingly.

Filtering

A few location 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 is the official name of the location in the PAT. In the temporary database, we put the Italian and Ladin names as alternative names.

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. We created the missing locations and eliminated the duplicates. At the end of the importing we identified the objects reported in Table 1.

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

Table 1. Objects imported in the temporary database

4 Building the faceted ontology

As mentioned above, the macro-classes provided by the PAT are very generic. This is mainly due to the criteria used by PAT during categorization that were based not only on type but also on importance and population criteria. With the two-fold goal of refining them and determining the missing semantic relations between them, we analyzed the class names and created a multi-lingual faceted ontology.

Our final goal was to create an ontology that both reflects the specificity of the PAT and respects the canons of the analytico-synthetic approach [Ranganathan, 1967] for the

[8] S-Match uses WordNet by default but it is configurable

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

generation of a faceted ontology. A faceted (lightweight) ontology [Giunchiglia *et al*., 2009] is an ontology divided into sub-trees, called *facets*, each encoding a different dimension or aspect of the domain knowledge. As a result, it can be seen as a collection of lightweight ontologies [Giunchiglia and Zaihrayeu, 2009].

From macro-classes to atomic concepts

We started from the 45 macro-classes, which are not accompanied by any description. Therefore, by analysing the locations contained in the macro-classes, each class was manually disambiguated and refined (split, merged or renamed) and as a result new classes had to be created. 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 were found, but they were used to understand the classes corresponding to each macro-class. The identified classes were manually refined. Some of them required a deeper analysis.

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.

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* facet (in English and Italian).

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)

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 refinements. 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. 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 can successfully apply a heuristic then we classify the entity in the corresponding (more specific) class otherwise we select a more generic class, that is the root of a facet (same as the block name) in the worst case. For some macro-classes we reached a success rate of 98%. On average, nearly 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.

Concept integration

This step consisted in mapping atomic concepts from the faceted ontology to GeoWordNet concepts. We automated the disambiguation process with a little amount of manual intervention. Basically, we first manually identified the concept corresponding to the root of each facet - that we call the *facet concept* - and then we restricted the matching of the atomic concepts in the facet to the sub-tree rooted in the facet concept in GeoWordNet. For instance, we restricted the matching of *mountain* to only those concepts more specific than *natural elevation*. If a candidate was found the corresponding concept was selected, otherwise a more general concept, i.e. a suitable parent, was searched. If neither the concept nor the parent was identified, we went for manual intervention.

Entity matching and integration

Two partially overlapped entity repositories, the temporary database built from the PAT dataset (i.e. the populated fac-

eted ontology) and GeoWordNet, were integrated. The PAT dataset overall contains 20,162 locations. GeoWordNet contains nearly 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. We also automatically generated an Italian and English gloss for each entity. We used several rules, according to the language. In order to detect the duplicates we experimented different approaches. We found that in order to maximize accuracy two entities must match only if they have same name, coordinates, class, parent entities, children entities and alternative names. We allowed a tolerance in matching the coordinates of +/- 0.05, corresponding to +/- 5.5 Km. Note that while matching classes, we took into account the subsumption hierarchy of their concepts. For instance, 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.

7 Evaluation

In order to improve the results associated to a user query, S-Match is used to match terms in the query with the faceted ontology. The background knowledge used for the matching is WordNet, but the tool is developed in such a way to allow substituting it with any other ontology. The matching terms are used to enrich those in the query thus obtaining a semantic query expansion. It is expected that such richer queries, given in input to GeoNetwork, would return a higher number of results. To prove the effectiveness of the approach followed, in Table 2 we provide some examples of queries and the terms in their extension.

Query	Terms identified by S-Match
Watercourse	Rivulet, Stream, River
Falls	Cascade, Waterfall
Elevation	Natural elevation, Mountain, Highland, Glacier, Mountain range, Peak, Hill
Mount	Mountain pass, Mountain, Mountain range
Installation	Milestone, Hut, Farm, Highway, Railway, Road, Street, Transportation system, Provincial Road, Facility, Shelter
Water	Rivulet, Waterfall, Cascade, River, Body of water, Stream, Spring, Canal, Group of lakes, Lake
Transportation facility	Transportation system, Road, Street, Provincial Road, Milestone, Railway, Highway
Reef	

Table 2. Some query expansion results

The last example shows how typing the query *reef* would not produce any result. This depends on the fact that the faceted ontology strictly codifies the local specificities of the Province of Trento that does not present any marine environment (it is a mountain region far from the sea). In all the other cases it is evident how S-Match identifies semantically related terms (synonyms, less or more specific terms).

Table 3 shows real results in terms of documents found. The portal actually interacts with the users in Italian. The Italian query is translated in English, matched with the faceted ontology and, once processed, results are given back in Italian. Only terms matching with at least one document are returned. For instance, the query for *tracking* returns only *pista* (track) and *ponte* (bridge), since no documents are available in the repository for the term *tracking*.

Query	Expansion (with number of documents)
foresta	foresta (119), bosco (14)
fiume	fiume (18), alveo (16)
lago	lago (4), laghi (20)
strada	strada (14), strada provinciale (5)
connessione	connessione (3), ponte (6)
paese	località (15), provincia (348), città (4), comune (952), frazione (2), centri abitati (16)
tracking	pista (5), ponte (6)

Table 3. Some query expansion results

Note that by populating the ontology with locations and taking into account the part-of relations between them, also location names can be expanded. For instance, by providing the information that *Povo* is an administrative division in *Trento* it is possible to expand the term *Trento* with *Povo*. However, providing this support was out of the scope of the SGC project.

8 Extending the faceted ontology

The work done with the PAT for the construction of the faceted ontology can be generalized to cover the whole world. We recently worked on a methodology - mainly inspired by the faceted approach - and a minimal set of guiding principles aimed at modeling the spatial domain (and in general any domain) and at building the corresponding background knowledge taking into account the *classes*, the *entities*, their *attributes* and *relations* [Dutta *et al*., 2011]. We consider classes, relations, and attributes as the three fundamental components, or categories, of any domain. In this approach, the analysis of the domain allows the identification of the basic classes of real world objects. They are arranged, per *genus et differentia* (i.e. by looking at their commonalities and their differences), to construct the *facets*, each of them codifying a different aspect of the domain at hand. This allows being much more rigorous in the definition of the domain and its parts, in its maintenance and use [Giunchiglia *et al*., 2009].

We selected the classes from GeoWordNet and arranged them into 8 facets, each of them further divided into sub-facets: region, administrative division, populated place, facility, abandoned facility, land, landform and body of water.

The spatial relations we propose extend those in [Pullar and Egenhofer, 1988]. In addition to the standard direction, topological, ordinal, distance and fuzzy relations, we extend them by including relative level (e.g. above, below), longitudinal (e.g. in front, behind), side-wise (e.g. right, left), position in relation to border or frontier (e.g. adjacent, overlap) and other similar relations. We also consider functional relations. For example, in the context of lakes, primary inflow and primary outflow are two important relations.

An attribute is an abstraction belonging to or a characteristic of an object. This is a construct through which objects or individuals can be distinguished. Attributes are primarily *qualitative* and *quantitative* in nature. For example, we may mention depth (of a river), surface area (of a lake), length (of a highway) and altitude (of a hill). For each of these attributes, we may have both qualitative and quantitative values. We store the possible qualitative values in the background knowledge. This provides a controlled vocabulary for them. They are mostly *adjectives*. For example, for depth (of a river) the possible values are {wide, narrow}. Similarly, for altitude (of a hill) the possible values are {high, low}. We also make use of *descriptive* attributes. They are used to describe, usually with a short natural language sentence, a specific aspect of an entity. Typical examples are the history (of a monument) or the architectural style (of a building) or any user defined tag.

Our space domain overall includes 845 classes, 70 relations and 35 attributes. In comparing it with existing geo-spatial ontologies, like GeoNames and TGN, our space domain is much richer in all its aspects. Further details can be found in [Dutta *et al*., 2011].

9 Conclusions

We briefly reported our experience in providing a semantic extension to the geo-catalogue of the PAT. S-Match, once integrated with GeoNetwork, performs a semantic expansion of the query using a faceted ontology codifying the domain knowledge about geography of the PAT. This allows identifying information that would be more difficult to find using traditional information retrieval approaches.

To mutually increase their coverage, we have also integrated the faceted ontology with GeoWordNet. At this purpose we had to match their concepts and entities. The matching of the concepts was done by focusing on one facet at a time. The entity matching criteria needed to be tuned to maximize accuracy. We also briefly reported the methodology that we use to build domains and how we applied it to the space domain on top of GeoWordNet.

Acknowledgments

This work has been supported by the TasLab network project funded by the European Social Fund under the act n° 1637 (30.06.2008) of the PAT, by the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 231126 LivingKnowledge: Living-Knowledge - Facts, Opinions and Bias in Time and by "Live Memories - Active Digital Memories of Collective Life" funded by the PAT. We are thankful to Pavel Shvaiko, Aliaksandr Autayeu, Veronica Rizzi, Daniela Ferrari, Giuliana Ucelli, Monica Laudadio, Lydia Foess and Lorenzo Vaccari for their kind support.

References

[Dutta *et al*., 2011] B. Dutta, F. Giunchiglia, V. Maltese. A facet-based methodology for geo-spatial modelling. In *Proceedings of the GEOS, Vol. 6631*, 2011.

[EU Commission, 2009] European Commission. COMMISSION REGULATION (EC) No 976/2009 implementing Directive 2007/2/EC as regards the Network Services, 2009.

[EU Parliament, 2009] European Parliament. Directive 2007/2/EC establishing an Infrastructure for Spatial Information in the European Community (INSPIRE), 2009.

[Giunchiglia *et al*., 2008] F. Giunchiglia, F. McNeill, M. Yatskevich, J. Pane, P. Besana, and P. Shvaiko. Approximate structure-preserving semantic matching. In *Proceedings of ODBASE*, 2008.

[Giunchiglia and Zaihrayeu, 2009] F. Giunchiglia, I. Zaihrayeu. Lightweight Ontologies. *The Encyclopedia of Database Systems*, 2007

[Giunchiglia *et al*., 2009] 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.

[Giunchiglia *et al*., 2010a] F. Giunchiglia, A. Autayeu, J. Pane. S-Match: an open source framework for matching lightweight ontologies. *The Semantic Web journal*, 2010.

[Giunchiglia *et al*., 2010b] F. Giunchiglia, V. Maltese, F. Farazi, B. Dutta. GeoWordNet: a resource for geo-spatial applications. In *Proceedings of ESWC*, 2010.

[Ivanyukovich *et al*., 2009] A. Ivanyukovich, F. Giunchiglia, V. Rizzi, V. Maltese. SGC: Architettura del sistema. *TCG/INFOTN/2009/3/D0002R5 report*, 2009.

[Pullar and Egenhofer, 1988] D. Pullar, M. J. Egenhofer. Toward formal definitions of topological relations among spatial objects. In *Proceedings of the 3rd International Symposium on Spatial Data Handling, Sydney, Australia, pp. 165–176*, 1988.

[Ranganathan, 1967] S. R. Ranganathan. Prolegomena to library classification. *Asia Publishing House*, 1967.

[Shvaiko *et al*., 2010] P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, F. Farazi. A semantic geo-catalogue implementation for a regional SDI. In *Proceedings of the INPSIRE Conference*, 2010.