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 SEMANTIC SCHEMA FOR GEONAMES

Vincenzo Maltese and Feroz Farazi

January 2013

Technical Report # DISI-13-004

A semantic schema for GeoNames

Vincenzo Maltese, Feroz Farazi
DISI, University of Trento, Trento, Italy

Abstract. As part of a broader strategy towards supporting semantic interoperability in geospatial applications, in this paper we present a semantic schema we designed for GeoNames and the qualitative improvements we obtained by enforcing it on the data.

Introduction. GeoNames (www.geonames.org) is a well-known geospatial dataset providing geographical data and metadata of around 7 million unique named places from all over the world collected from several sources. At top level, the places are categorized into 9 broader feature classes, further divided into 663 features which are arranged in a flat list with no relations between them. A special null class contains unclassified entities. Each class is associated one name and often a natural language description. Yet, a fixed terminology is an obstacle towards achieving semantic interoperability [6]. For example, if it is decided that the standard term to denote *a terminal where subways load and unload passengers* is *metro station*, it would fail in applications where the same concept is denoted with *subway station.* This weakness has been identified as one of the key issues for the future of the INSPIRE implementation [8, 9, 10, 11].

As part of the solution, geospatial ontologies - by providing alternative terms and semantic relations between them - represent a more flexible alternative [12, 13, 18]. They can be basically seen as *semantic standards*. Following this line, in our previous work [4, 5, 3] we came up with a methodology and a minimal set of guiding principles, based on the *faceted approach*, as originally used in library science [14], and developed a large-scale multilingual *geospatial faceted ontology* obtained from the refinement and extension of GeoNames, WordNet (wordnet.princeton.edu) and MultiWordNet (multiwordnet.fbk.eu). It accounts for the relevant classes, entities, their relations and attributes arranged into *facets*, each of them capturing a different aspect of the geospatial domain. For instance, it includes the facets *land formation*, *body of water* and *populated place* with corresponding more specific classes (exemplified in the picture aside). Following the faceted approach is known to guarantee the construction of very high quality ontologies in terms of robustness, extensibility, reusability, compactness and flexibility [15, 16]. This approach has been proven effective in geospatial applications. It is worth mentioning for instance the benefits obtained from the usage of such ontologies within the discovery service of the semantic geo-catalogue of the Autonomous Province of Trento in Italy [1, 17]. This work also put the basis for the release of its geographical data and metadata as linked open government data [2].

- **Landform**
  - Natural depression
    - Oceanic depression
      - Oceanic valley
      - Oceanic trough
    - Continental depression
      - Trough
      - Valley
  - Natural elevation
    - Oceanic elevation
      - Seamount
      - Submarine hill
    - Continental elevation
      - Hill
      - Mountain
- **Body of water**
  - Flowing body of water
    - Stream
      - Brook
      - River
  - Still body of water
    - Pond
    - Lake
- **Populated place**
  - City
  - Town
  - Village

Nevertheless, the usage of a geospatial ontology does not solve all the problems. In fact, GeoNames seems to lack of sufficient constraints on the domain and range of the attributes, and of corresponding mechanisms to enforce them which can guarantee for an adequate quality of the data. For instance, such constraints should prevent the attribute *population* to have a negative value and while it is fine for cities to have such attribute, this should be prevented for streams. This deficiency results in some unexpected mistakes. The solution we adopt is what we call a semantic schema.

The semantic schema. In this setting, we define a *semantic schema* as a set of constrains on the domain and range of the attributes (e.g. population) and the relations (e.g. capital) in the dataset. In particular, the schema is semantic-aware because the domain of attributes and relations, and the range of relations are always a class and its more specific classes taken from the geospatial ontology. For instance, if we specify that the domain of the attribute *population* is *populated place* (the main class), we assume it to apply also to *city*, *town* and *village* (more specific classes in the ontology). In the specific case of GeoNames, the range of attributes is instead a standard data type (e.g. integer, float or string). The purpose of the schema is expressly to define what is legal in terms of attributes, relations and corresponding values. Enforcing the schema corresponds to verifying the consistency of the dataset w.r.t. such constraints (see, e.g. [7]). Among others, the schema we defined includes the following constraints:

Attribute Name	**Definition**	**Domain (main class)**	**Range**
Population	the people who inhabit a territory or state	Populated Place	Long > 0
Altitude	elevation above sea level	Location but Undersea	Float in [-423, 8848]
Elevation	vertical distance above a reference point	Undersea	Float
Area	the extent of a 2-dimensional surface enclosed within a boundary	Location	Float > 0
Capital	A seat of government	Geo-political entity	Populated Place

Notice in particular how we distinguish between elevation and altitude and separate the first from the second when clear from the domain. On the contrary, in GeoNames only elevation is provided. In fact, while elevation refers to a generic distance from a reference point, altitude is a more specific notion as in this case the reference point is the sea level. The range of altitude was set by referring to the altitude of the Dead Sea (the lowest) and Mount Everest (the highest) as taken from Wikipedia. Enforcing the schema brought to some surprising results. For instance:

- Despite in GeoNames it is assumed that elevation has not to be provided for oceanic entities, we have found that 2,934 entities (e.g., *Mentawai Ridge*) of 33 different undersea classes (e.g., *oceanic ridge*, *oceanic valley*) have actually a value for it. We keep these values in the ontology by separating them from altitude.
- In GeoNames the Dead Sea is represented with negative altitude set to −405 m. Surprisingly, GeoNames contains other 45 locations with same altitude of the Dead Sea, and two other locations are reported to be even lower than the Dead Sea (Nahal Amazyahu and `Arvat Sedom). Manual checks were needed to verify their correctness.
- The domain of population includes several unexpected classes such as *airport*, *stream* and *garden*. We removed population from corresponding entities in the ontology.
- We found several entities with elevation set to -9999 that is used in GeoNames to encode an unknown value. We removed elevation from corresponding entities in the ontology.
- In the range of capital, 3 entities are registered as cities (e.g. Jerusalem) while all the others as capitals. This is not wrong, but at least this is not homogeneous. Actually, as no location is *essentially* a capital (the capital of a country may change in time; see also [19] about the distinction between rigid and not rigid properties), we set corresponding class to *populated place* for all of them.

- The area of *United States Minor Outlying Islands* is set to 0. We corrected it to 34200 m^2 as reported in Wikipedia.

Conclusions. In this paper we have stressed the need for an integrated approach to effectively support semantic interoperability between different geospatial applications. The proposed solution consists in the usage of a *geospatial faceted ontology* providing the terminology of the geospatial domain (which can be seen as a sort of more flexible *semantic standard*) and a semantic schema that, by establishing precise constraints on the domain and range of the attributes and the relations, guarantees a higher level of data quality.

Acknowledgments. The research leading to these results has received funding from the CUbRIK Collaborative Project, partially funded by the European Commission's 7th Framework ICT Programme for Research and Technological Development under the Grant agreement no. 287704.

References

1. P. Shvaiko, A. Ivanyukovich, L. Vaccari, V. Maltese, F. Farazi (2010). A semantic geo-catalogue implementation for a regional SDI. INSPIRE conference.
2. P. Shvaiko, F. Farazi, D. Ferrari, G. Ucelli, L. Vaccari, V. Maltese, V. Rizzi, A. Ivanyukovich (2012). Trentino government linked open geodata: first results. INSPIRE conference.
3. Giunchiglia, F., Maltese, V., Dutta, B. (2012). Domains and context: first steps towards managing diversity in knowledge. Journal of Web Semantics, 12-13, 53-63.
4. Giunchiglia, F., Dutta, B., Maltese, V., Farazi, F. (2012). A facet-based methodology for the construction of a large-scale geospatial ontology. Journal of Data Semantics, 1(1), 57-73.
5. Giunchiglia, F., Maltese, V., Farazi, F., and Dutta, B. (2010). GeoWordNet: a Resource for Geo-Spatial Applications. 7th Extended Semantic Web Conference (ESWC).
6. Kuhn, W. (2005). Geospatial semantics: Why, of What, and How? Journal of Data Semantics (JoDS), III, pp. 1–24
7. Gomez-Perez, A. (2001). Evaluation of ontologies. International Journal of Intelligent Systems, 16 (3), 36.
8. Crompvoets, J., Wachowicz, M., de Bree, F., Bregt, A. (2004). Impact assessment of the INSPIRE geo-portal. 10th EC GI&GIS workshop.
9. Smits, P., Friis-Christensen, A. (2007). Resource discovery in a European Spatial Data Infrastructure. Transactions on Knowledge and Data Engineering, 19(1), 85–95.
10. Lutz, M., Ostlander, N., Kechagioglou, X., Cao, H. (2009). Challenges for Metadata Creation and Discovery in a multilingual SDI - Facing INSPIRE. ISRSE conference.
11. Vaccari, L., Shvaiko, P., Marchese, M. (2009). A geo-service semantic integration in spatial data infrastructures. Journal of Spatial Data Infrastructures Research, 4, 24–51.
12. Egenhofer, M. J. (2002). Toward the Semantic GeoSpatial Web. In the 10th ACM International Symposium on Advances in Geographic Information Systems (ACM-GIS), 1-4.
13. Kolas, D., Dean, M., Hebeler, J. (2005). Geospatial Semantic Web: architecture of ontologies. First Int. Conference on GeoSpatial Semantics (GeoS), 183-194.
14. Ranganathan, S. R. (1967). Prolegomena to library classification. Asia Publishing House.
15. Spiteri, L. (1998). A Simplified Model for Facet Analysis. Journal of Information and Library Science, 23, 1-30.
16. Broughton, V. (2006). The need for a faceted classification as the basis of all methods of information retrieval. Aslib, 58 (1/2), 49-72.
17. Farazi, F., Maltese, V., Dutta, B., Ivanyukovich, A., Rizzi, V. (2012). A semantic geo-catalogue for a local administration. AI Review Journal, 1-20. (DOI) 10.1007/s10462-012-9353-z.
18. Tóth, K., Portele, C., Illert, A., Lutz, M., Nunes de Lima, V. (2012). A Conceptual Model for Developing Interoperability Specifications in Spatial Data Infrastructures. European Commission Joint Research Centre.
19. Guarino, N., Welty, C. (2002). Evaluating ontological decisions with OntoClean. Communications of the ACM, 45 (2), 61-65.