A SEMANTIC SCHEMA FOR GEONAMES

Enzo Maltese
Feroz Farazi
Index

• Introduction
• GeoNames
• DERA Methodology
• Faceted Ontology
• Semantic Schema
• Result
• Conclusion

INSPIRE Conference – June 25th 2013

Feroz Farazi – University of Trento
Introduction

• With the proliferation of the Web, dataset publishing is also getting a momentum

• With Linked Open Data (LOD) initiative
  • Old datasets are published in machine readable format
  • Many new datasets are made available on the Web

• As part of LOD initiative essentially we observe that datasets are there with increased
  • Machine readability
  • Availability
  • Ease of finding through links
Introduction

• However, in the LOD initiative more emphasis is given on the quantity of the data than its quality

• In this work we proposed a semantic schema that can help improve the quality while enforcing it on the data

• Our approach is domain independent

• We performed a detailed case study using GeoNames
GeoNames

- GeoNames is a geo-spatial database consists of various locations of all countries
- It includes geographical data
  - Place names in various languages
  - Latitude
  - Longitude
  - Altitude
  - Class
- Latitude and longitude coordinates are stored according to the WGS84
GeoNames

- It currently contains around 8 million geographic names for around 7 million unique places

- The places are classified into 9 broader categories called feature classes

- They are further divided into 667 classes, most of them have a natural language description

- The data is available free of charge, can be used under a creative commons attribution license
DERA Methodology

• A methodology for developing ontology

• It allows building domain specific ontologies

• However, it can be used to construct ontology for any domain

• In DERA, a domain consists of three components namely entity class, relation and attribute.

\[
D = \langle E, R, A \rangle
\]

• E is a tuple \(<C, e>\), where C represents set of concepts and e represents set of entities
DERA Methodology

• Methodology:
  • Step 1: Identification of the relevant terms
  • Step 2: Analysis
  • Step 3: Synthesis
  • Step 4: Standardization

• Following the above steps leads to the creation of a set of facets
Faceted Ontology

- Facet is a hierarchy of homogeneous concepts describing an aspect of the domain (e.g., space), where each term in the hierarchy denotes an atomic concept.

- A **facet** is a distinctive property of a set of concepts that can help in differentiating one group from another.

- **Faceted ontology** is an ontology in which concepts are organized into facets.

- **S. R. Ranganathan** was the first to introduce the faceted approach.

- **GeoWordNet** is an example of a faceted ontology consists of facets such as *body of water*, *geological formation* and *administrative division*.
## Faceted Ontology

<table>
<thead>
<tr>
<th>Body of water</th>
<th>Landform</th>
</tr>
</thead>
<tbody>
<tr>
<td>Flowing body of water</td>
<td>Natural depression</td>
</tr>
<tr>
<td>Stream</td>
<td>Oceanic depression</td>
</tr>
<tr>
<td>Brook</td>
<td>Oceanic valley</td>
</tr>
<tr>
<td>River</td>
<td>Oceanic trough</td>
</tr>
<tr>
<td>Still body of water</td>
<td>Continental depression</td>
</tr>
<tr>
<td>Pond</td>
<td>Trough</td>
</tr>
<tr>
<td>Lake</td>
<td>Valley</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Populated place</th>
<th>Natural elevation</th>
</tr>
</thead>
<tbody>
<tr>
<td>City</td>
<td>Oceanic elevation</td>
</tr>
<tr>
<td>Town</td>
<td>Seamount</td>
</tr>
<tr>
<td>Village</td>
<td>Submarine hill</td>
</tr>
<tr>
<td></td>
<td>Continental elevation</td>
</tr>
<tr>
<td></td>
<td>Hill</td>
</tr>
<tr>
<td></td>
<td>Mountain</td>
</tr>
</tbody>
</table>
Semantic Schema

• The usage of a geospatial ontology does not solve all the problems

• In fact, GeoNames lacks sufficient constraints on the domain and range of the attributes

• Moreover, it lacks corresponding mechanisms to enforce them which can guarantee for an adequate quality of the data

• We define a semantic schema as a set of constrains on the domain and range of
  • the attributes (e.g. population)
  • the relations (e.g. capital)
Semantic Schema

- Such constraints should prevent the attribute population to have a negative value

- While it is fine for cities to have such attribute, this should be prevented for streams

- 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
Semantic Schema

- We assume it to apply also to city, town and village (more specific classes in the ontology)

- The purpose of the schema is expressly to define what is legal in terms of
  - attributes
  - relations
  - corresponding values

- Enforcing the schema corresponds to verifying the consistency of the dataset w.r.t. such constraints
## Semantic Schema

<table>
<thead>
<tr>
<th>Attribute Name</th>
<th>Definition</th>
<th>Domain (main class)</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population</td>
<td>the people who inhabit a territory or State</td>
<td>Populated Place</td>
<td>Long &gt; 0</td>
</tr>
<tr>
<td>Altitude</td>
<td>elevation above sea level</td>
<td>Location but Undersea</td>
<td>Float in [-423, 8848]</td>
</tr>
<tr>
<td>Area</td>
<td>the extent of a 2-dimensional surface enclosed within a boundary</td>
<td>Location</td>
<td>Float &gt; 0</td>
</tr>
<tr>
<td>Capital</td>
<td>A seat of government</td>
<td>Geo-political entity</td>
<td>Populated Place</td>
</tr>
</tbody>
</table>
Semantic Schema

• The range of altitude was set by referring to the altitude of
  • the Dead Sea (the lowest)
  • the Mount Everest (the highest)
Result

• In GeoNames the Dead Sea is represented with negative altitude set to $-405\text{ m}$.
  • Surprisingly, GeoNames contains other 45 locations with same altitude of the Dead Sea
  • two other locations are reported to be even lower than the Dead Sea (Nahal Amazyahu and `Arvat Sedom)

• The domain of population includes several unexpected classes such as airport, stream and garden
  • We removed population from corresponding entities in the ontology
Result

• 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 encoded as cities (e.g. Jerusalem) while all the others as capitals
  • This is not wrong, but at least this is not homogeneous
  • We represented them homogeneously
Conclusion

• 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 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
Thank you

Questions?