

# Semi-automatic semantification of institutional spatial datasets

**Vasilis Kopsachilis, Nikos Vachtsavanis and Michail Vaitis**

Department of Geography, University of the Aegean, Greece



UNIVERSITY OF THE AEGEAN

 <p><b>European Union</b> European Regional Development Fund</p>	 <p>HELLENIC REPUBLIC MINISTRY OF DEVELOPMENT AND INVESTMENTS SPECIAL SECRETARIAT FOR ERDF &amp; CF PROGRAMMES MANAGING AUTHORITY OF ΕΡΑνεΚ</p>	<p><b>ΕΡΑνεΚ 2014-2020</b> OPERATIONAL PROGRAMME COMPETITIVENESS ENTREPRENEURSHIP INNOVATION</p>	 <p><b>ΕΣΠΑ</b> 2014-2020 ανάπτυξη - εργασία - αλληλεγγύη Partnership Agreement 2014 - 2020</p>
<p>Co-financed by Greece and the European Union</p>			

30 May 2022 in Hersonissos, Greece

# Institutional Spatial Datasets

- National mapping agencies, regional or local authorities, universities, research institutes, ...
- Produce and manage high-quality and high-resolution spatial data
- Spatial datasets may be stored and disseminated in various formats (Shapefile, GeoJSON, RDBMS, WFS, etc..)
- Spatial datasets may present heterogeneity with regards to the thematic areas that they cover, their production methods and purposes, schema definitions, quality and documentation level

## The case of the University of the Aegean

- Departments and research labs produce spatial data independently for various purposes (research programs, student assignments, etc.)
- Spatial datasets located in various systems (RDBMS, SDIs, local computers)
- Lack of a common representation format and interface that could facilitate advanced data querying & integration

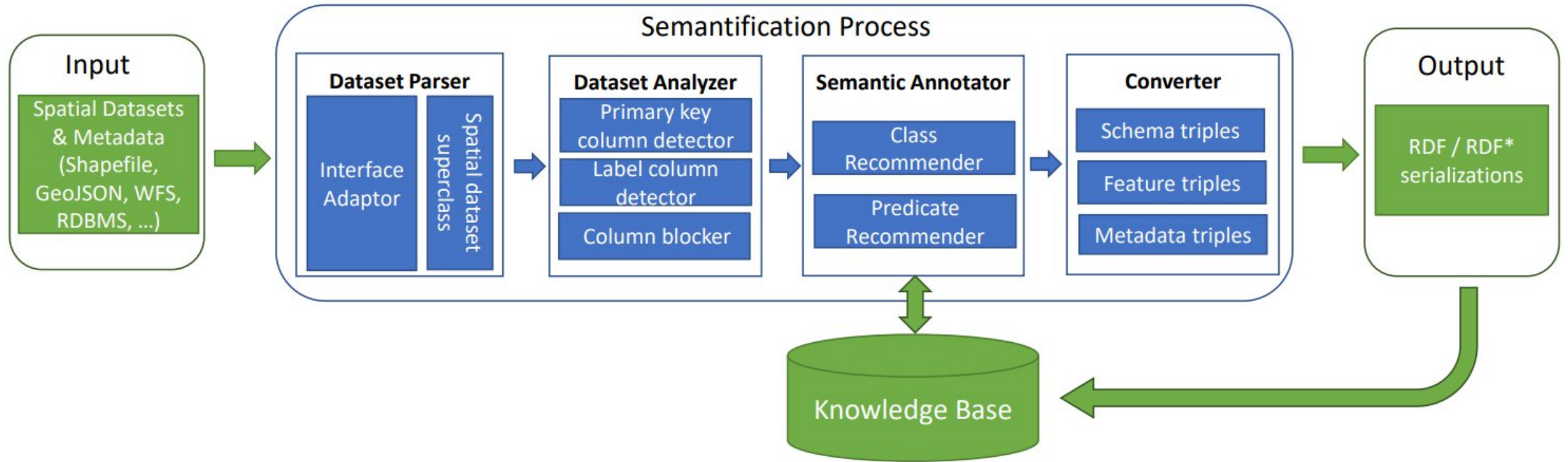
# Need for semantification

- Conversion of institutional spatial datasets to a common representation format, such as the RDF, and publish them as Linked Data
- Shared semantics that would provide capabilities for advanced querying, integration and reasoning among the institutional spatial data as well as with the entire Linked Open Data (LOD)
- The development of an “institutional” semantic Knowledge Base that would be part of the LOD
- Few local organizations participate actively in the linked data domain, possibly because of the lack of resources and expertise in the domain and the absence of easy-to-use semantification tools

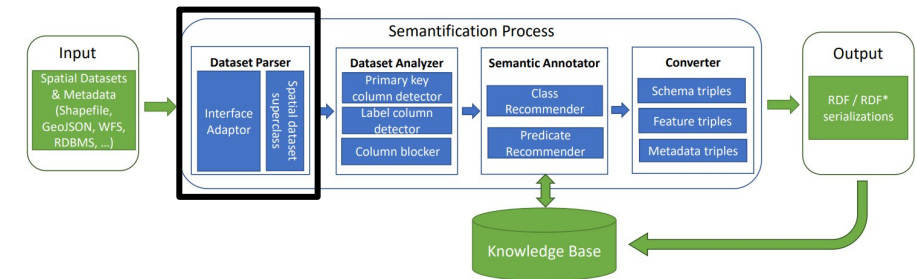
# Semantification Requirements

- The implementation of easy-to-use semantification tools that do not require deep expertise on linked data and knowledge of the schema of the datasets
- Support of semantic annotation recommendations based on existing semantic knowledge for guiding users and minimizing their involvement during semantification
- Handling of both geometric and thematic attributes
- Dynamic and incremental population of a knowledge base by the spatial datasets at hand
- Maintenance of provenance metadata about converted RDF
- A process that can be easily adapted by institutions that want to integrate their spatial data in LOD

# Design of the semantification process



# Dataset Parser



## Input

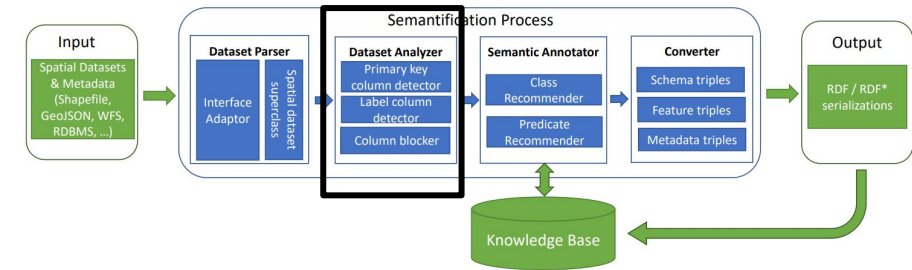
1. Spatial Datasets (Shapefiles, GeoJSON, RDBMS tables, WFS)
2. Metadata (e.g., INSPIRE-compliant XML files, CSW Records)

An Interface Adaptor parses dataset and extracts:

1. Schema-level information (the list of column names and their types)
2. The actual data (geographic features with their attribute values)
3. Metadata from spatial dataset file (dataset name, creation date, dataset format, publisher, description, geometry column, geometry type, original CRS and dataset spatial extent)
4. If available, it also parses metadata files (e.g., INSPIRE-compliant XML files)

The above information is modeled in a Spatial Dataset Superclass.

# Dataset Analyzer



## Primary Key column Detector

- Primary Key (PK) values will be used for assigning URIs to geographic features
- Candidate PK are integer and string columns that contain distinct, not null values
- String columns that contain large-length values are not candidate PK
- The module selects the most appropriate primary key column by giving priority to the candidate string columns

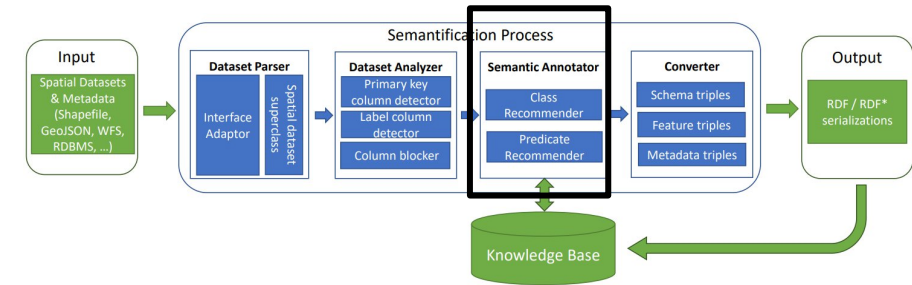
## Label columns Detector

- The label columns that will be annotated with the `rdfs:label` predicate
- Candidate label columns are string columns that contain small-length values
- Language detection for each candidate label column

## Column Blocker

- Columns that will not be converted to RDF
- Formed by the label columns (because they are already annotated), by columns that may refer to foreign keys and by user-defined/selected columns

# Semantic Annotator



Class and Predicate recommendations for semantic annotation based on the content of a Knowledge Base

## Class Recommendation

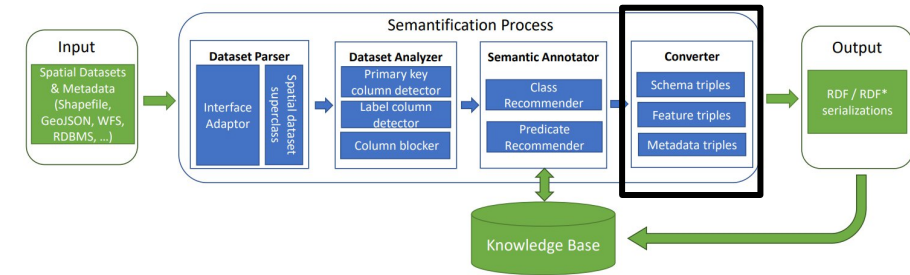
- The class that will be associated with the geographic features of the dataset
- The recommendations are based on the textual (Levenshtein) and semantic (WordNet WuPalmer) similarity between the dataset name and classes in the KB
- Users can select an alternative class that exists in the KB or create a new class
- In case of a new class, users are opted to provide a label (rdfs:label) and a description (rdfs:comment)

## Predicate Recommendation

- The predicates that will be associated with the columns of the dataset
- The recommendations are based on the textual (Levenshtein) and semantic (WordNet WuPalmer) similarity between the column name and predicates in the KB
- Users can select an alternative predicate that exists in the KB or create a new predicate
- In case of a new predicate, users are opted to provide a label (rdfs:label) and a description (rdfs:comment)



# Converter



## Schema Triples

- Definitions for new classes and predicates
- The resource URI is formed by a Base URI, the term “ontology” and the resource name, e.g.:

a) `http://semantics.aegean.gr/ontology/Ports`  
 Base URI                      Class Name

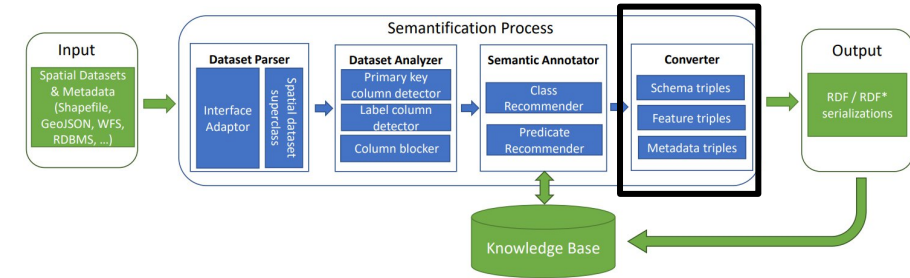
## Feature Triples

- Description of geographic features (instances)
- The instance URI is formed by the Base URI, the term “resource”, its class name and its PK value, e.g.:

b) `http://semantics.aegean.gr/resource/Ports_Pireus`  
 Base URI                      Class Name      PK Value

- An instance is declared (rdf:type) to be member of the annotated class and of the GeoSPARQL Feature class
- If label columns detected, the corresponding rdfs:labels are created
- The rest columns values are associated with their annotated predicates
- The geometric column is converted according to the GeoSPARQL vocabulary
  - If the dataset CRS is not WGS84, the geometry is reprojected and also converted according to the GeoSPARQL vocabulary
  - If the geometry is point, also the respective W3CBasic Geo triples are created

# Converter



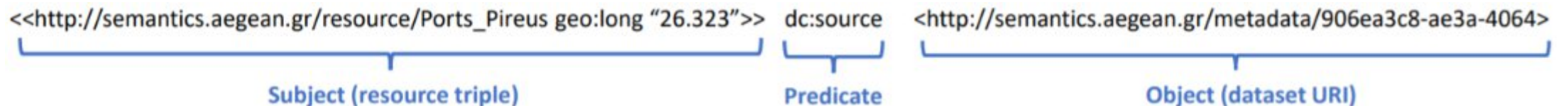
## Metadata Triples

- Contain the spatial dataset metadata (e.g., dataset name, creation date, publisher, CRS, spatial extent)
- A spatial dataset is declared to be member of a *SpatialDataset* class
- The Spatial Dataset URI is formed by the Base URI, the term “metadata” and a randomly generated ID



## Association of features triples with dataset metadata

- Triple-level metadata for capturing knowledge such as who created a piece of information and when.
- Each feature triple is associated with the dataset from which it originates, using the RDF\* model:
  - For each feature triple, a new RDF\* triple is created that in the subject position appears the triple itself, enclosed in ‘« »’, in the predicate position the ‘dc:source’ and in the object position the spatial dataset



# Example

“Ports” Shapefile

OIK_ID	CODE	NAME	PREFECTURE
1	252	Mytilene	LESVOS
2	456	Herakleion	CRETE
3	365	Pireus	ATTIKI



```

1 @prefix rdf:      <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
2 @prefix rdfs:    <http://www.w3.org/2000/01/rdf-schema#> .
3 @prefix geo:    <http://www.w3.org/2003/01/geo/wgs84_pos#> .
4 @prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
5 @prefix dc:     <http://purl.org/dc/terms/> .
6 @prefix uoa:    <http://semantics.aegean.gr/ontology/> .
7
8 ## Schema Triples
9 <http://semantics.aegean.gr/ontology/Ports>
10   a          rdfs:Class ;
11   rdfs:label "Ports" ;
12   rdfs:comment "This class describes ports" .
13
14 ## Metadata Triples
15 <http://semantics.aegean.gr/metadata/f904bdf9-526b-425e>
16   a          uoa:SpatialDataset ;
17   dc:created "Sat Feb 12 15:06:43 EET 2022" ;
18   dc:dateSubmitted "Sat Feb 12 15:15:59 EET 2022" ;
19   dc:format "Shapefile" ;
20   dc:publisher "Aegean University" ;
21   dc:title "ports" ;
22   uoa:CRS "Greek_Grid" ;
23   uoa:GeometryType "Point" .
24
25 ## Resource Triples
26 <http://semantics.aegean.gr/resources/Ports_Herakleion>
27   rdf:type uoa:Ports , geosparql:Feature ;
28   rdfs:label "Herakleion"@en ;
29   uoa:Code "456" ;
30   uoa:Id "2" ;
31   uoa:Prefecture "CRETE" ;
32   geosparql:hasGeometry _:b1 , _:b3 ;
33   geo:lat "39.36908053997243" ;
34   geo:long "26.15680836273347" .
35
36 _:b1 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/4326>POINT (26.15680836273347 39.36908053997243)" .
37
38 _:b3 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/2100>POINT (685647.9108665949 4359666.74695738)" .
39
40 << <http://semantics.aegean.gr/resources/Ports_Herakleion> rdfs:label "Herakleion"@en >>
41   dc:source <http://semantics.aegean.gr/metadata/f904bdf9-526b-425e> .
42
43 << <http://semantics.aegean.gr/resources/Ports_Herakleion> uoa:Prefecture "CRETE" >>
44   dc:source <http://semantics.aegean.gr/metadata/f904bdf9-526b-425e> .

```

# Implementation

- The semantification API is implemented in Java
- The GeoTools and JTS libraries are used for spatial dataset parsing and geometric transformations
- Apache Jena framework is used for RDF modelling and for sending SPARQL queries to the knowledge base
- The Knowledge Base is selected to be a Fuseki instance
- On top of the semantification API, a web application (RDFConverter) was developed that acts like a semantification wizard
- For testing purposes more than 100 Shapefiles from various sources were converted and loaded to an initially empty KB

# Demonstration – The RDF Converter App

<http://semantics.aegean.gr/RDFConverter>

RDF Converter   Help   Register

## RDF Converter

Convert your spatial datasets to RDF

Please enter your credentials to enter:

Email:

Password:

Login

# Demonstration – User Settings

psachilis ▾

### User Preferences

Knowledge Base URI: ⓘ	<input type="text" value="http://semantics.aegean.gr:3030/dat"/>	<input checked="" type="checkbox"/>
KB User:	<input type="text" value="admin"/>	
KB Password:	<input type="password"/>	
Base URI: ⓘ	<input type="text" value="http://semantics.aegean.gr"/>	
Base Prefix: ⓘ	<input type="text" value="uoa"/>	

# Demonstration – Dataset Upload

The screenshot shows the 'RDF Converter' web application interface. At the top, there is a green navigation bar with the text 'RDF Converter', 'Help', and a user profile 'vkopsachilis'. Below this, a large green header contains the title 'RDF Converter' and the subtitle 'Convert your spatial datasets to RDF'. A progress indicator below the header shows four steps: 1. Upload, 2. Analysis, 3. Annotation, and 4. Conversion. The 'Upload' step is currently active. The main content area is a grey box with the text 'Select a spatial dataset to begin the conversion' and a question mark icon. Below this text is a file selection button labeled 'Επιλογή αρχείων' and a message 'Δεν επιλέχθηκε κανένα αρχείο.' (No file was selected). A 'Next' button is located at the bottom right of the grey box. At the bottom of the page, there is a green footer with the text 'Designed by CartoGI Lab @ 2022' and 'Powered by w3.css'.

# Demonstration – Dataset Analysis

1  
Upload
2  
Analysis
3  
Annotation
4  
Conversion

### Dataset Analysis

This step presents the dataset preview and a set of recommended conversion options. You can fill missing dataset metadata and alter the conversion options at the respective panels.

#### Dataset Metadata

File Name:	ports	Creation Date:	Thu May 05 17:43:18 UTC 2022
Format:	Shapefile	CRS:	Greek_Grid
Publisher:	<input type="text"/>	Geometry:	Point
Description:	<input type="text"/>	Bounds:	ReferencedEnvelope[659079.4150601482 : 721535.8176560029, 4330523.150549255 : 4359666.74695738]
Source:	<input type="text"/>		

#### Features Preview

[Show on Map](#)

OIK_ID	NAME	PREFECTURE	CODE
1	Mytilene	LESVOS	252
2	Herakleion	CRETE	456
3	Pireus	ATTIKI	365

#### Conversion Options

Select a Primary Key: ?

Select label columns: ?

Select columns for conversion: ?

Back
Next



# Demonstration – Semantic Annotation

1 Upload      2 Analysis      3 **Annotation**      4 Conversion

## Semantic Annotation

This step presents the recommendations for the class and predicate dataset annotation. The recommendations are based on the Knowledge Base (KB) and Base URI you have set in the preferences page. You can alter the recommendations below.

### Class Annotation

Annotation options for the dataset: **Ports** ?

- Create a new class:
- Select an existing class:
- Select a recommended class:

### Predicate Annotation

Annotation options for the column: **PREFECTURE** ?

- Create a new predicate:  
Predicate Name:       Predicate Label:       Predicate Description:
- Select an existing predicate:
- Select a recommended predicate:

Annotation options for the column: **CODE** ?



- Create a new predicate:
- Select an existing predicate:
- Select a recommended predicate:



Demonstration – Convert and Upload Step

### Converted RDF

This is a preview of the final RDF triples. You can either download them in a Turtle file or upload them to your Knowledge Base. You can alter the output by providing different configuration and annotation options.

```
@prefix dc: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix uoa: <http://semantics.aegean.gr/ontology/> .

<< _b0 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/2100>POINT (659079.4150601482 4339719.937580696)" >>
  dc:source <http://semantics.aegean.gr/metadata/20a2221b-f6a1-442f-b8d9-4e674378e1e2> .

_b1 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/4326>POINT (26.15680836273347 39.36908053997243)" .
_b2 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/4326>POINT (26.563479172156914 39.098270543890166)" .

<< <http://semantics.aegean.gr/resources/port_Mytilene> uoa:hasGreekGovernmentGazette "252" >>
  dc:source <http://semantics.aegean.gr/metadata/20a2221b-f6a1-442f-b8d9-4e674378e1e2> .



uoas:hasPrefecture rdfs:type rdfs:Property;
  rdfs:comment "This predicate ..";
  rdfs:label "prefecture" .

<< <http://semantics.aegean.gr/resources/port_Herakleion> uoa:hasPrefecture "CRETE" >>
  dc:source <http://semantics.aegean.gr/metadata/20a2221b-f6a1-442f-b8d9-4e674378e1e2> .

<< _b1 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/4326>POINT (26.15680836273347 39.36908053997243)" >>
  dc:source <http://semantics.aegean.gr/metadata/20a2221b-f6a1-442f-b8d9-4e674378e1e2> .

<< <http://semantics.aegean.gr/resources/port_Pireus> geo:long "25.843836308383153"^^<http://www.w3.org/2001/XMLSchema#double> >>
  dc:source <http://semantics.aegean.gr/metadata/20a2221b-f6a1-442f-b8d9-4e674378e1e2> .

_b3 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSSG/0/2100>POINT (659079.4150601482 4339719.937580696)" >>
```

# Knowledge Base Content

## KB Explorer

Explore Aegean University Knowledge Base

## Classes

- <http://semantics.aegean.gr/ontology/Region>
- <http://semantics.aegean.gr/ontology/Prefecture>
- <http://semantics.aegean.gr/ontology/Municipality>
- <http://semantics.aegean.gr/ontology/Chamber>
- <http://semantics.aegean.gr/ontology/Hospital>
- <http://semantics.aegean.gr/ontology/Museum>
- <http://semantics.aegean.gr/ontology/University>
- <http://semantics.aegean.gr/ontology/Bank>
- <http://semantics.aegean.gr/ontology/canyon>
- <http://semantics.aegean.gr/ontology/airport>

View all

## Instances

- [http://semantics.aegean.gr/resources/Region\\_IonioI\\_NisoI](http://semantics.aegean.gr/resources/Region_IonioI_NisoI)
- [http://semantics.aegean.gr/resources/Region\\_Kriti](http://semantics.aegean.gr/resources/Region_Kriti)

## Predicates

- <http://semantics.aegean.gr/ontology/hasDescription>
- <http://semantics.aegean.gr/ontology/id>
- <http://semantics.aegean.gr/ontology/city>
- <http://semantics.aegean.gr/ontology/municipality>
- <http://semantics.aegean.gr/ontology/area>
- <http://semantics.aegean.gr/ontology/perimeter>
- <http://semantics.aegean.gr/ontology/name>
- [http://semantics.aegean.gr/ontology/angle\\_unit](http://semantics.aegean.gr/ontology/angle_unit)
- <http://semantics.aegean.gr/ontology/code>
- [http://semantics.aegean.gr/ontology/volcano\\_status](http://semantics.aegean.gr/ontology/volcano_status)

View all

## Datasets

- <http://semantics.aegean.gr/metadata/c920d9ef-787f-43be-90e8-59d0d7ab58cc>
- <http://semantics.aegean.gr/metadata/6fcd4b5-7f05-42bb-ab65->

<http://semantics.aegean.gr/ontology/Hospital>

Aegean Semantics

## Hospital

<http://semantics.aegean.gr/ontology/Hospital>

---

A hospital is a health care institution providing patient treatment with specialized health science and auxiliary healthcare staff and medical equipment.

Property	Value
rdfs:comment	<ul style="list-style-type: none"> <li>A hospital is a health care institution providing patient treatment with specialized health science and auxiliary healthcare staff and medical equipment.</li> </ul>
rdfs:label	<ul style="list-style-type: none"> <li>Hospital</li> </ul>
rdf:type	<ul style="list-style-type: none"> <li>&lt;<a href="http://www.w3.org/2000/01/rdf-schema#Class">http://www.w3.org/2000/01/rdf-schema#Class</a>&gt;</li> </ul>

[http://semantics.aegean.gr/resources/Prefecture\\_Lesvou](http://semantics.aegean.gr/resources/Prefecture_Lesvou)

Aegean Semantics

## Λέσβου

[http://semantics.aegean.gr/resources/Prefecture\\_Lesvou](http://semantics.aegean.gr/resources/Prefecture_Lesvou)

---

Property	Value
uoa:code	<ul style="list-style-type: none"> <li>54</li> </ul>
uoa:hasEconomicValue	<ul style="list-style-type: none"> <li>1415.103858</li> </ul>
geosparql:hasGeometry	<ul style="list-style-type: none"> <li>[4 anonymous resources]</li> </ul>
uoa:hasHumanPopulation	<ul style="list-style-type: none"> <li>103698.0</li> </ul>
uoa:hasUnemploymentRate2013	<ul style="list-style-type: none"> <li>21.0</li> </ul>
uoa:id	<ul style="list-style-type: none"> <li>10</li> <li>32</li> <li>7754</li> </ul>
rdfs:label	<ul style="list-style-type: none"> <li>Λέσβου (el)</li> </ul>
rdf:type	<ul style="list-style-type: none"> <li>&lt;<a href="http://semantics.aegean.gr/ontology/Prefecture">http://semantics.aegean.gr/ontology/Prefecture</a>&gt;</li> <li>&lt;<a href="http://www.opengis.net/ont/geosparql#Feature">http://www.opengis.net/ont/geosparql#Feature</a>&gt;</li> </ul>

# Initial Assessment

- The design of the semi-automatic semantification wizard is intuitive and allows the completion of the process easily and in short time even by non-experts on semantic web
- The annotation recommendations help users, without strong familiarity with the knowledge base content, to quickly determine the suitable classes and properties
- The process guarantees the instant population of the knowledge base with well-formed RDF that is ready to use
- We plan to conduct more detailed experiments in order to evaluate the overall performance of the semantification process and its ability to populate high-quality semantic content

# Further Improvement

- Design of more sophisticated rules for the dataset analysis step, e.g.:
  - improved primary and foreign key column detection
  - block duplicate columns
  - identify columns with specific content (telephones, emails, etc)
- Design of more sophisticated rules for the semantic annotation, e.g.:
  - instance-based methods for class and predicates annotation
  - integration of third-party semantic web search engine APIs (e.g., Linked Open Vocabularies, GeoLOD) for recommending resources from external KBs
- Design of a post processing process that will:
  - perform some cleaning (e.g., URI merging, substitution of literal with object properties)
  - establish sameAs links between local and external instances
  - perform ontology alignment in order to detect equivalency or hierarchy relations between local and external classes and properties
- Standardize the vocabulary for dataset metadata (e.g., RDF representation of INSPIRE metadata)
- Adoption of the alternative RDF\* annotation syntax
- Selection of alternative spatial vocabularies for representing geometries (other than W3C Basic Geo and GeoSPARQL)

# Thank you!

This research was funded by the Research e-Infrastructure [e- Aegean R&D Network], which is implemented within the framework of the “Regional Excellence” Action of the Operational Program “Competitiveness, Entrepreneurship and Innovation”. The action was co-funded by the European Regional Development Fund (ERDF) and the Greek State [Partnership and Cooperation Agreement 2014-2020].



30 May 2022 in Hersonissos, Greece