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

European Union
European Regional Development Fund

HELLENIC REPUBLIC
MINISTRY OF DEVELOPMENT AND INVESTMENTS
SPECIAL SECRETARIAT FOR ERDF & CF PROGRAMMES
MANAGING AUTHORITY OF EPAnEK

EPAnEK 2014–2020
OPERATIONAL PROGRAMME
COMPETITIVENESS
ENTREPRENEURSHIP
INNOVATION

ΕΣΠΑ 2014-2020
ανάπτυξη - εργασία - αλληλεγγύη
Partnership Agreement 2014 - 2020

Co-financed by Greece and the European Union

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

<u>The case of the University of the Aegean</u>

- 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

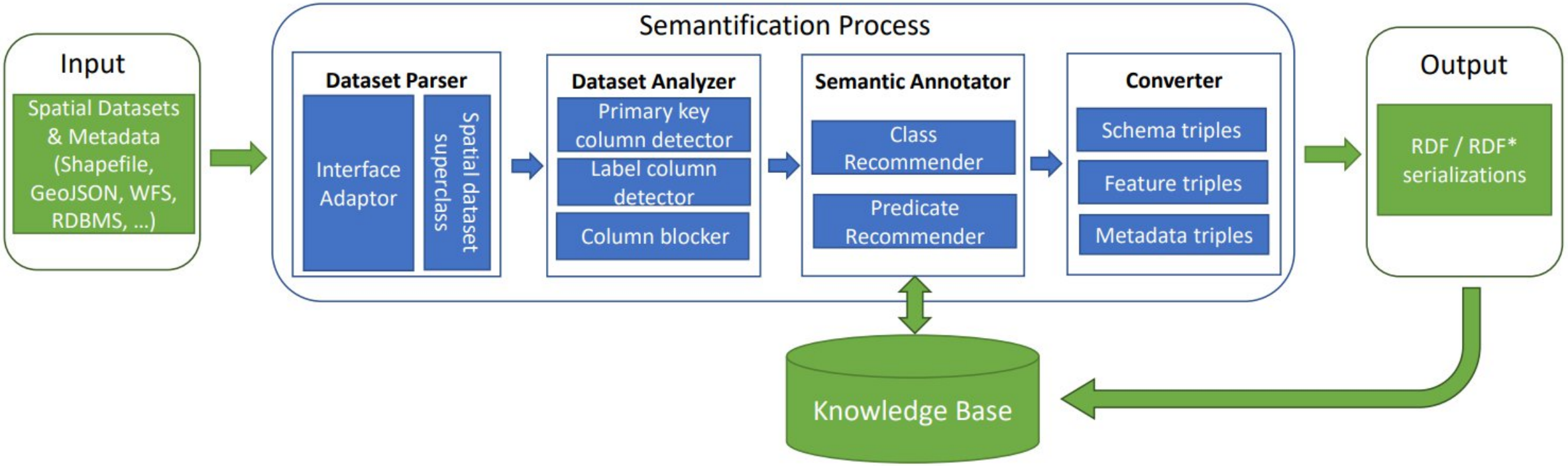
Need for semantification

- 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

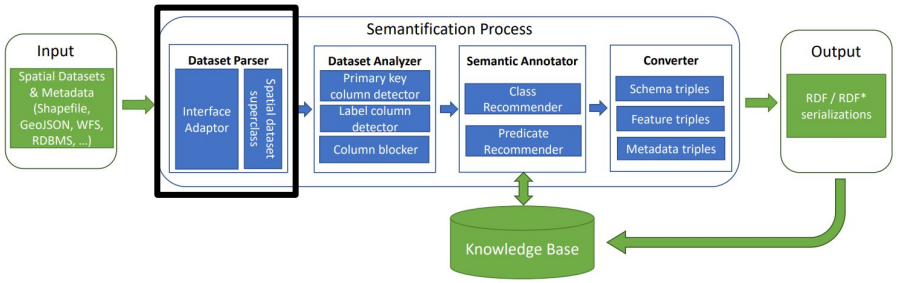
Semantification Requirements

- 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

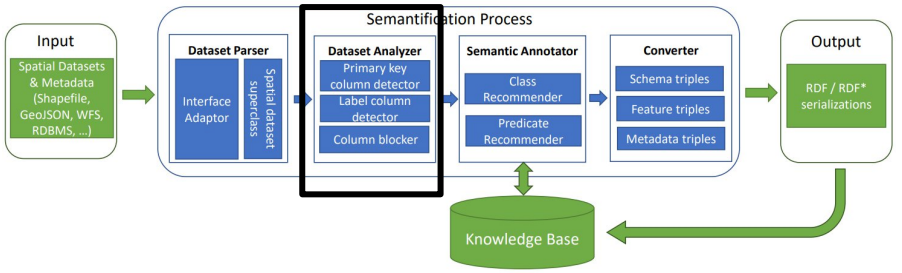
1. Spatial Datasets (Shapefiles, GeoJSON, RDBMS tables, WFS)
2. Metadata (e.g., INSPIRE-compliant XML files, CSW Records)

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

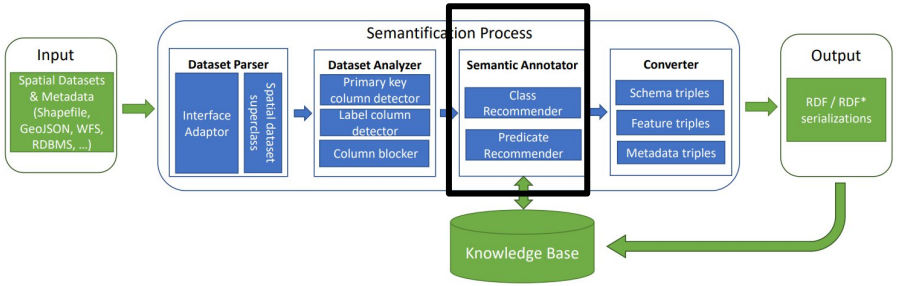
Label columns Detector

- 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

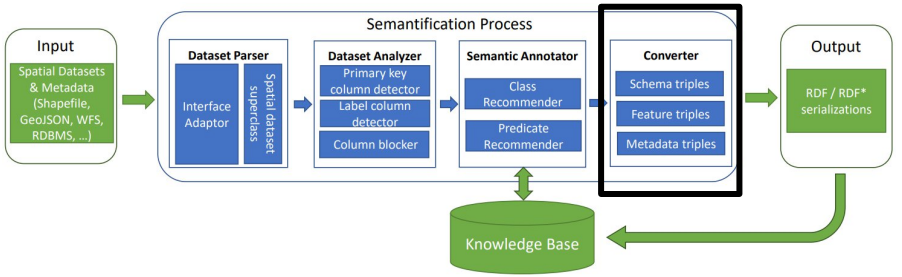
Class Recommendation

- 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

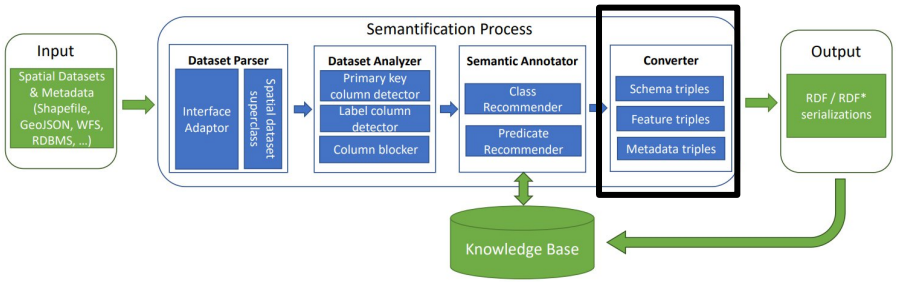
- Definitions for new classes and predicates
- The resource URI is formed by a Base URI, the term “ontology” and the resource name, e.g.:

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.:

- 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

c) http://semantics.aegean.gr/metadata/906ea3c8-ae3a-4064

Base URI — Dataset 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

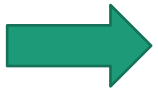
<<http://semantics.aegean.gr/resource/Ports_Pireus geo:long “26.323”>> dc:source <http://semantics.aegean.gr/metadata/906ea3c8-ae3a-4064>

Subject (resource triple) — Predicate — Object (dataset URI)

Example

“Ports” Shapefile

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



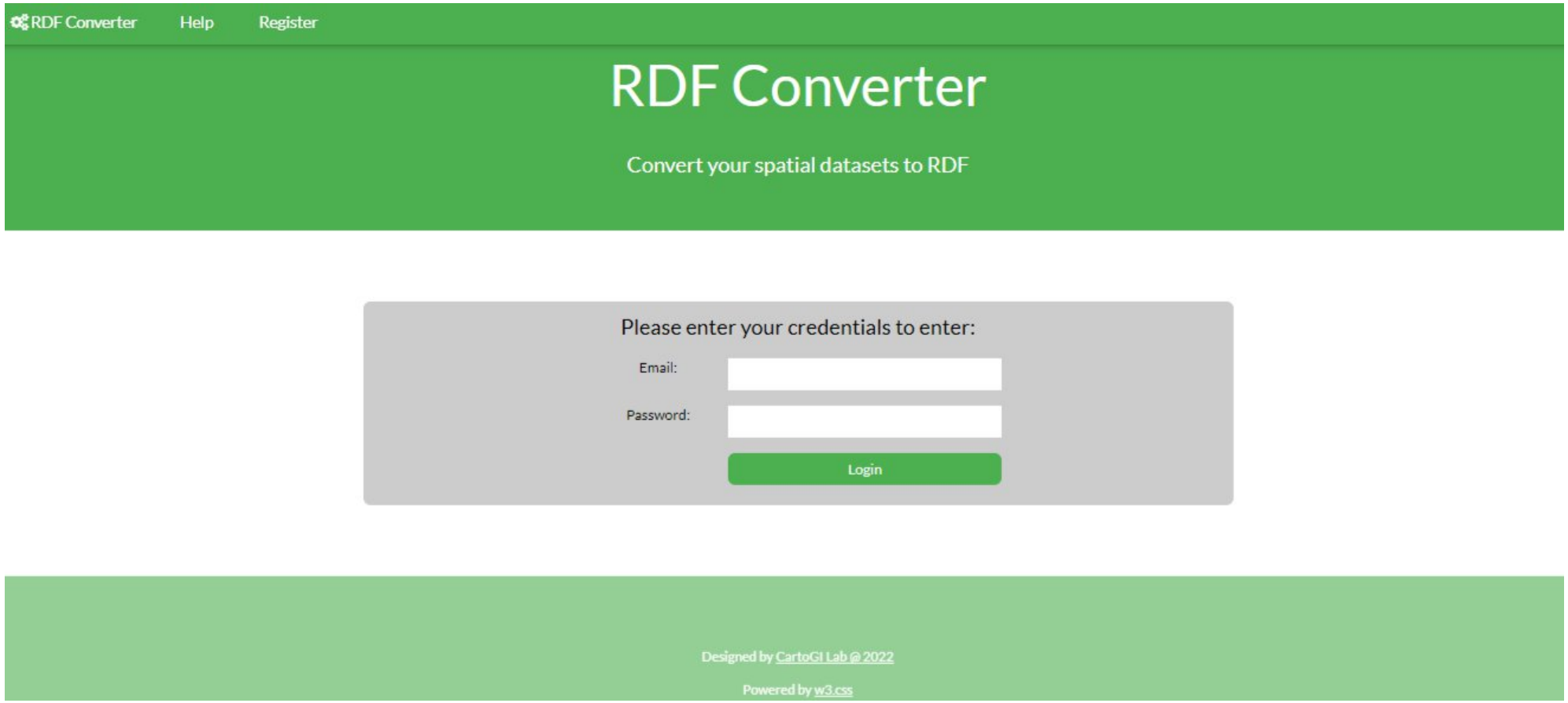
```
@prefix rdf:       <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:      <http://www.w3.org/2000/01/rdf-schema#> .
@prefix geo:       <http://www.w3.org/2003/01/geo/wgs84_pos#> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix dc:        <http://purl.org/dc/terms/>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

## Schema Triples
<http://semantics.aegean.gr/ontology/Ports>
        a                   rdfs:Class ;
        rdfs:label          "Ports" ;
        rdfs:comment        "This class describes ports" .

## Metadata Triples
<http://semantics.aegean.gr/metadata/f904bdf9-526b-425e>
        a                   uoa:SpatialDataset ;
        dc:created          "Sat Feb 12 15:06:43 EET 2022" ;
        dc:dateSubmitted    "Sat Feb 12 15:15:59 EET 2022" ;
        dc:format           "Shapefile" ;
        dc:publisher        "Aegean University" ;
        dc:title            "ports" ;
        uoa:CRS             "Greek_Grid" ;
        uoa:GeometryType    "Point" .

## Resource Triples
<http://semantics.aegean.gr/resources/Ports_Herakleion>
        rdf:type                uoa:Ports , geosparql:Feature ;
        rdfs:label              "Herakleion"@en ;
        uoa:Code                "456" ;
        uoa:Id                  "2" ;
        uoa:Prefecture          "CRETE" ;
        geosparql:hasGeometry   _:b1 , _:b3 ;
        geo:lat                 "39.36908053997243";
        geo:long                "26.15680836273347".

_:b1    geosparql:asWKT  "<http://www.opengis.net/def/crs/EPSG/0/4326>POINT (26.15680836273347 39.36908053997243)" .

_:b3    geosparql:asWKT  "<http://www.opengis.net/def/crs/EPSG/0/2100>POINT (685647.9108665949 4359666.74695738)" .

<< <http://semantics.aegean.gr/resources/Ports_Herakleion> rdfs:label "Herakleion"@en >>
        dc:source  <http://semantics.aegean.gr/metadata/f904bdf9-526b-425e> .

<< <http://semantics.aegean.gr/resources/Ports_Herakleion> uoa:Prefecture "CRETE" >>
        dc:source  <http://semantics.aegean.gr/metadata/f904bdf9-526b-425e> .
```

Implementation

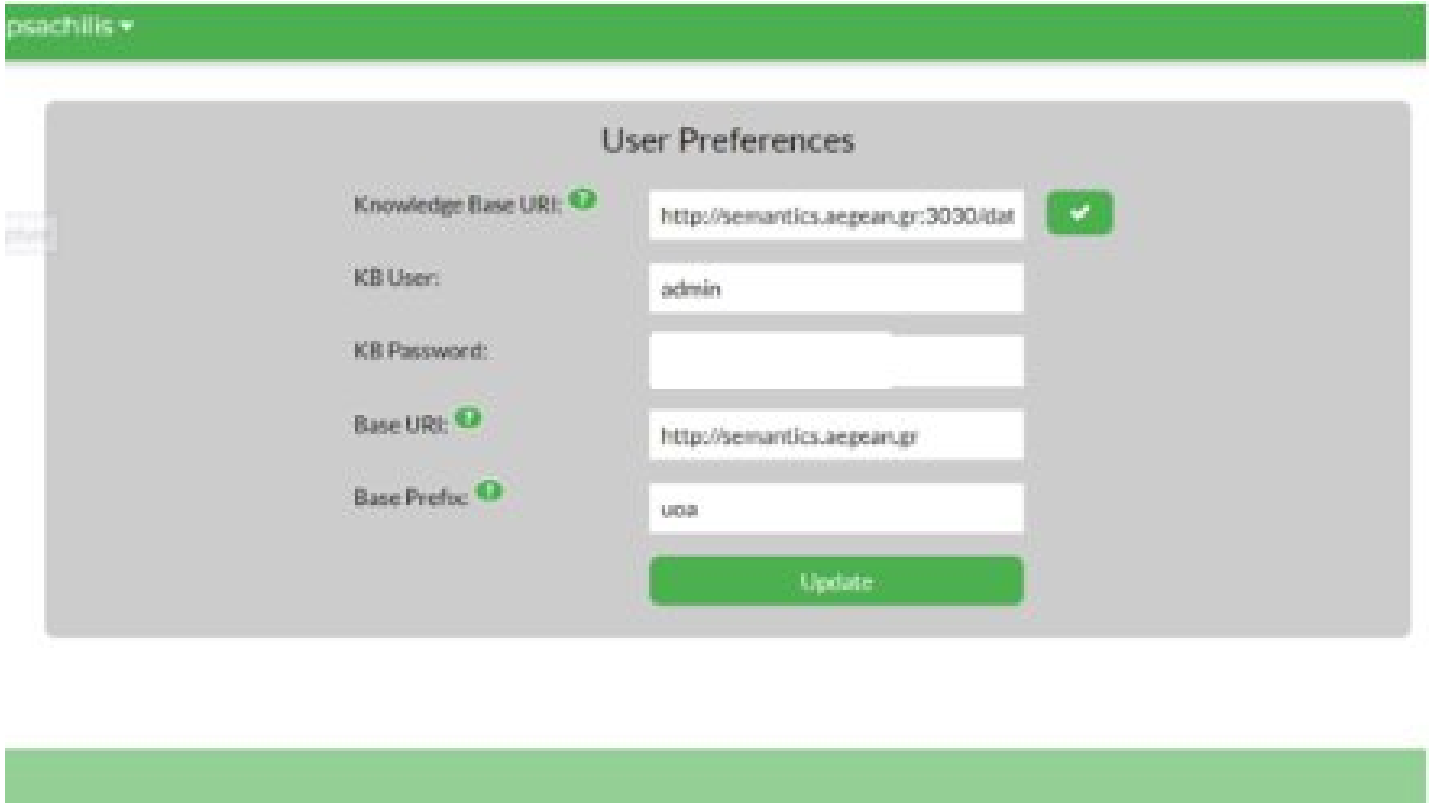
- The semantification API is implemented in Java
- The GeoTools and JTS libraries are used for spatial dataset parsing and geometric transformations
- Apache Jena framework is used for RDF modelling and for sending SPARQL queries to the knowledge base
- The Knowledge Base is selected to be a Fuseki instance
- 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 converter and loaded to an initially empty KB

Demonstration – The RDF Converter App

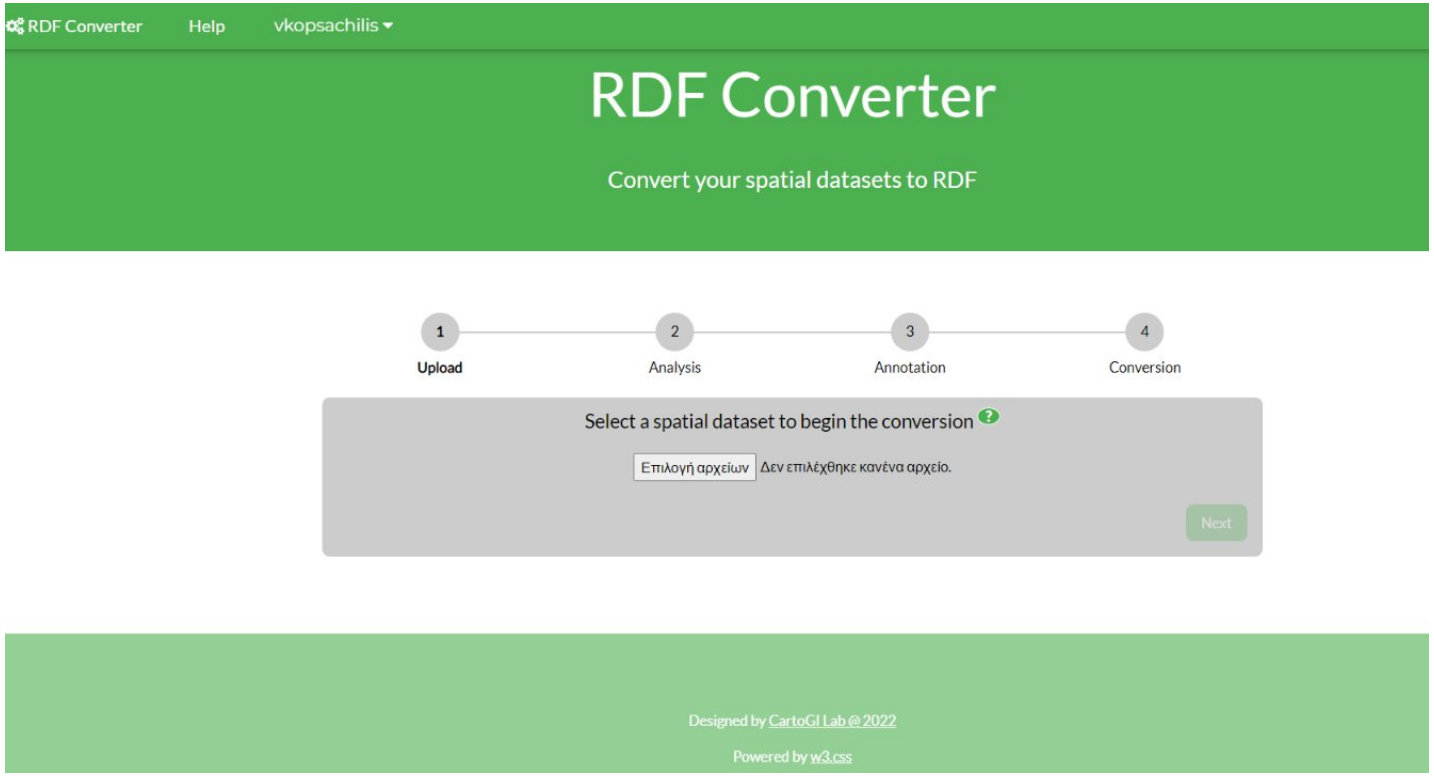
http://semantics.aegean.gr/RDFConverter



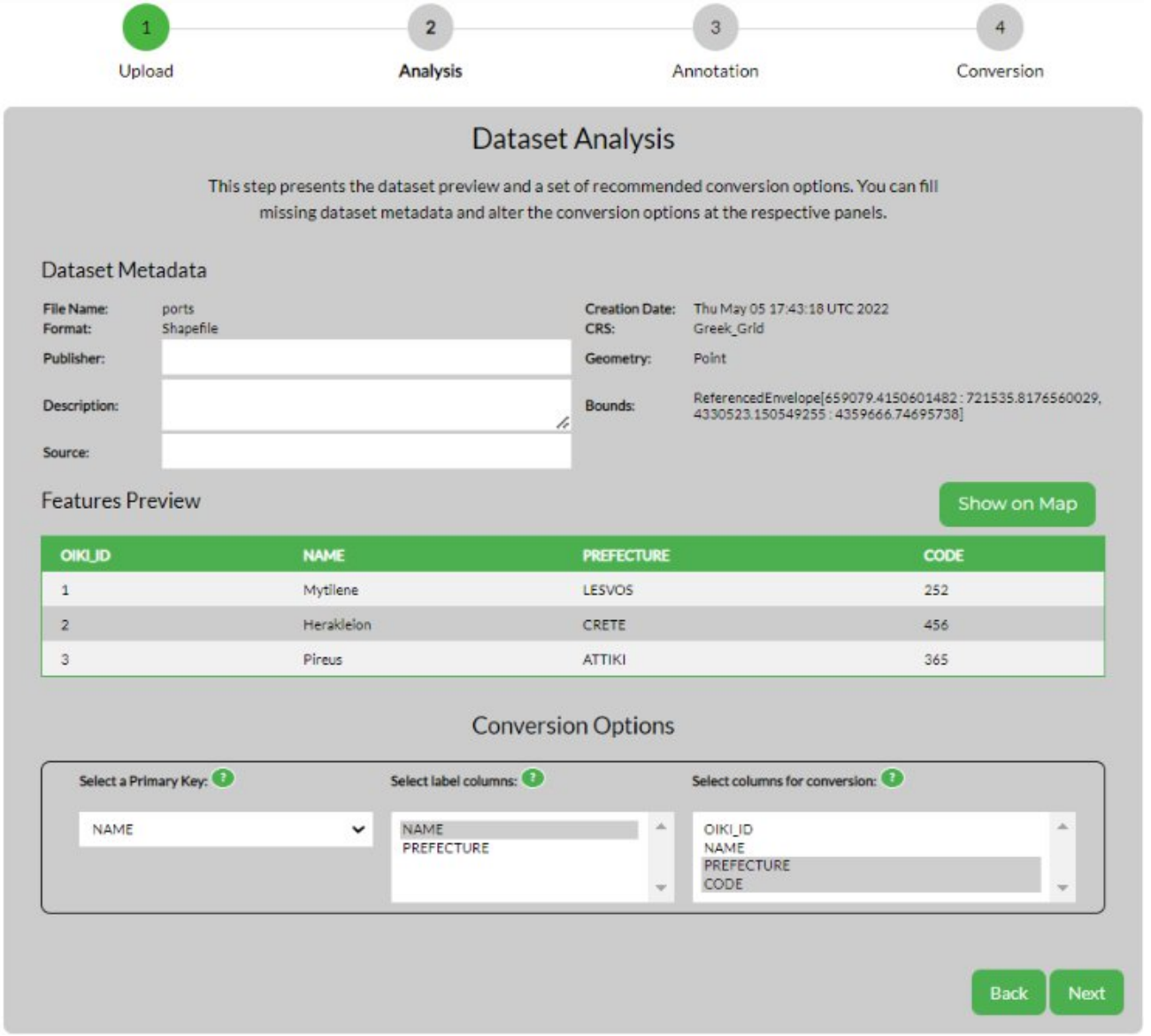
Demonstration – User Settings



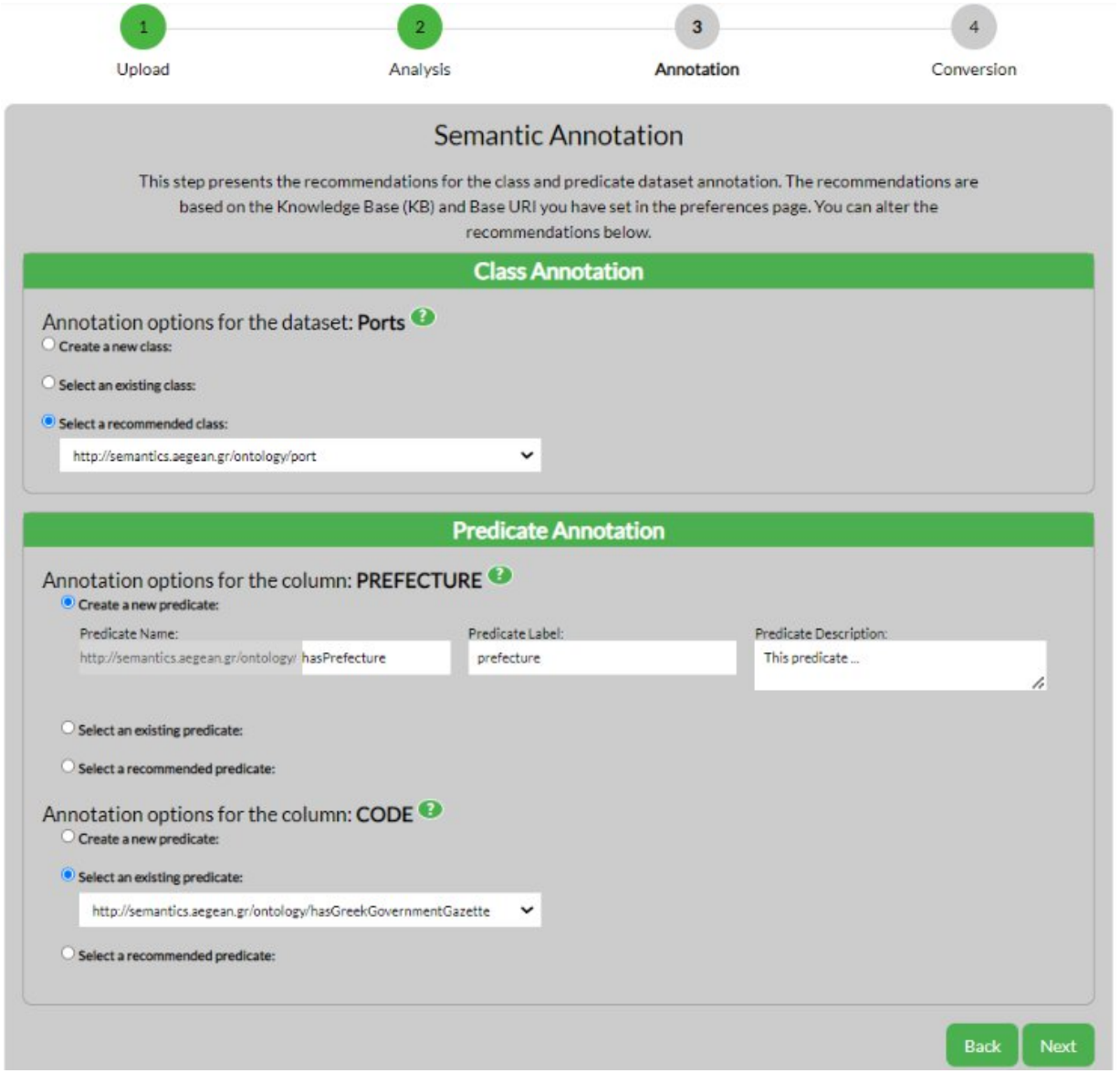
Demonstration – Dataset Upload



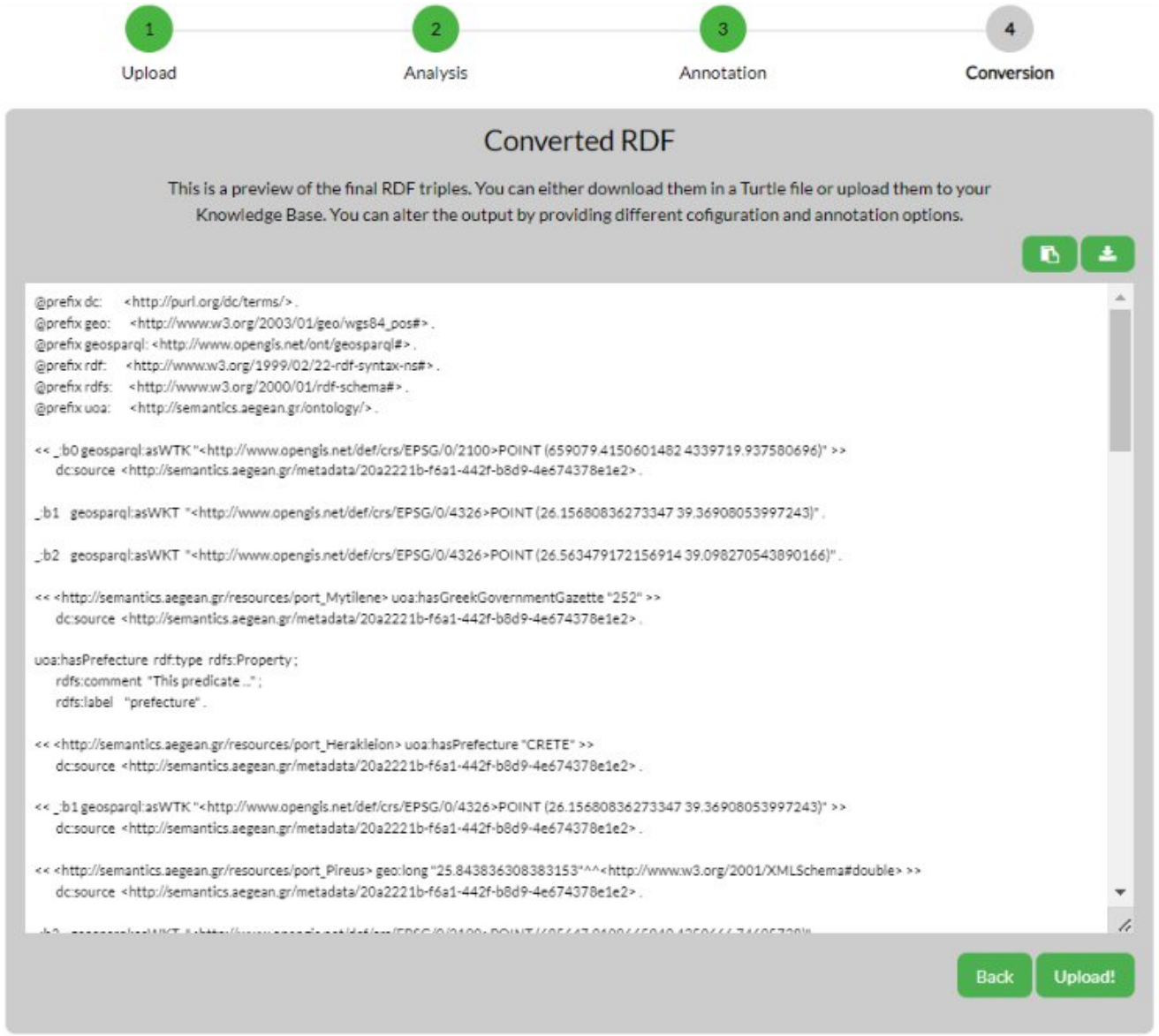
Demonstration – Dataset Analysis



Demonstration – Semantic Annotation



Demonstration – Convert and Upload Step



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

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/code
http://semantics.aegean.gr/ontology/volcano_status
View all

Instances

http://semantics.aegean.gr/resources/Region_IonioiNisoi
http://semantics.aegean.gr/resources/Region_Kriti

Datasets

http://semantics.aegean.gr/metadata/c920d9ef-787f-43be-90e8-59d0d7ab58cc
http://semantics.aegean.gr/metadata/6fcdb4b5-7f05-42bb-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	• A hospital is a health care institution providing patient treatment with specialized health science and auxiliary healthcare staff and medical equipment.
rdfs:label	• Hospital
rdf:type	• <http://www.w3.org/2000/01/rdf-schema#Class>

http://semantics.aegean.gr/resources/Prefecture_Lesvou

Aegean Semantics

Λέσβου

http://semantics.aegean.gr/resources/Prefecture_Lesvou

Property	Value
uoa:code	• 54
uoa:hasEconomicValue	• 1415.103858
geosparql:hasGeometry	• [4 anonymous resources]
uoa:hasHumanPopulation	• 103698.0
uoa:hasUnemploymentRate2013	• 21.0
uoa:id	• 10 • 32 • 7754
rdfs:label	• Λέσβου (el)
rdf:type	• <http://semantics.aegean.gr/ontology/Prefecture> • <http://www.opengis.net/ont/geosparql#Feature>

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