

Semi-automatic Semantification of Institutional Spatial Datasets

Vasilis Kopsachilis^a, Nikos Vachtsavanis^a and Michail Vaitis^a

^aUniversity of the Aegean, University Hill, Mytilene, 81100, Greece

Abstract

This paper describes a semi-automatic process for the semantic annotation of institutional non-RDF spatial datasets and their integration into an incrementally populated knowledge base. The knowledge base can serve as the organizational Spatial Knowledge Infrastructure which can be exploited for data querying and further integration purposes. The highlights of the process are the inclusion of analytical capabilities that produce annotation recommendations based on existing semantic knowledge for minimizing the user involvement during the semantification process, the dynamic and incremental building of the underlying schema based on the spatial datasets at hand, and the adoption of the RDF* model for maintaining triple-level metadata. This work is undertaken for the case of the University of the Aegean; however, the solution is generic and can be easily adapted by other organizations as well.

Keywords

Spatial datasets, Data transformation, Semantic annotation, Geographical knowledge bases


1. Introduction


Many organizations, such as national mapping agencies, regional or local authorities, universities and research institutes, produce and manage a substantial amount of high quality and resolution spatial data for covering their geographical areas of jurisdiction or sites where research work is carried out. These data are usually stored in a variety of formats, such as ESRI Shapefile, GeoJSON or Spatial Relational Database Management Systems (RDBMS), or disseminated through Spatial Data Infrastructures (SDI) as services, e.g., according to the Web Feature Service (WFS) format. The adoption of open and well-known formats and protocols may ensure the syntactic interoperability among datasets [1], however still institutional datasets may present high heterogeneity with regards to the thematic areas that they cover, their production methods and purposes, schema definitions, storage and dissemination formats, quality and documentation level. For example, within a university various data stores, databases or SDIs may developed and maintained independently by different Departments or Research Labs to cover different thematic domains and to serve different purposes and applications (e.g., an RDBMS with economic indicators for administrative units, an SDI about natural disaster protection, or scattered spatial datasets produced during research programs and student assignments). These factors hinder the

GeoLD 2022: 5th International Workshop on Geospatial Linked Data co-located with ESWC, May 30 2022, Hersonissos, Greece

✉ vkopsachilis@geo.aegean.gr (V. Kopsachilis); nicos.vachtsavanis@aegean.gr (N. Vachtsavanis); vaitis@aegean.gr (M. Vaitis)

ORCID 0000-0003-3824-3932 (V. Kopsachilis); 0000-0002-1269-6071 (M. Vaitis)

 © 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

possibility for achieving semantic interoperability between spatial data that could potentially unlock advanced capabilities for institutional data exploitation. A possible solution could be the conversion of the (dispersed across the various sources) spatial datasets to a common data representation format, such as the RDF, for the semantic representation and annotation of the data and their metadata, and their publication to the web according to the linked data principles. Such a process will provide additional capabilities for querying, integration and reasoning among the institutional spatial data as well as with the entire Linked Open Cloud (LOD). As [2] note regarding the availability of spatial data in LOD, most spatial data are published by global data providers (e.g., DBpedia and LinkedGeoData) and few local organizations participate actively to the linked data domain, possibly because the lack of resources and expertise in the domain and the absence of easy-to-use semantification tools.

Towards the goal of tabular data transformation to RDF, W3C outlines some approaches. On one hand, the Direct Mapping Recommendation [3] generates RDF by assigning table names to classes and column names to predicates. On the other hand, the R2RML Recommendation [4] is a mapping language that allows to define custom mappings between tabular data and RDF resources. Most existing semantic transformation tools apply the above recommendations or similar processes in a manual, semi-automatic or automatic manner. For example, GeoTriples [5] extends the R2RML language to allow the transformation of spatial datasets to RDF and includes sophisticated methods for mapping geometric attributes to the GeoSPARQL and stSPARQL vocabularies. TripleGeo [6] is an ETL tool that also focuses on the conversion of the geometric attributes of spatial datasets. Regarding thematic attributes, it allows the selection of columns that their values will be used as identifiers, names and types in order to generate the URI, the `rdfs:label` and the `rdf:type` triples respectively for the produced RDF entities. SemGIS [7] goes beyond simple transformation and describes a sophisticated approach for the automatic semantic integration of spatial datasets into the semantic web. Specifically, it tries to guess the class of the data to be integrated by performing content-based geometric and textual comparisons between the spatial dataset and existing semantic web resources. Also, it tries to guess the predicates to be mapped with each column by applying NLP and geocoding methods and performing a feature analysis that classifies columns according to their probable types (e.g., IDs, measurement units, addresses, phone numbers, email). The above works are pioneers in semanticfication in the geospatial domain, but in our opinion they present some shortcomings with regard to an easy-to-use process for the transformation of spatial datasets to RDF. For example, GeoTriples and TripleGeo support only the manual specification of the semantic annotation rules for the dataset thematic attributes, while, SemGIS does not provide many options for the geometric attributes transformation, such as CRS reprojection, and does not allow users to select alternative class and predicate annotations. Moreover, to the best of our knowledge, SemGIS is not provided as a publicly available application.

This paper describes the design and implementation of a semantification process that recommends annotations of spatial datasets based on existing semantic knowledge. The process acts like a semi-automated wizard, thus, it minimizes the human effort for semantification, while allowing users to intervene by defining alternative conversion options, such as the use of custom classes and predicates. The converted RDF graph, which consists of schema, feature and metadata triples, is fed in a knowledge base, which is the main knowledge source of the process. By this way, the process allows the interactive, dynamic and incremental building of

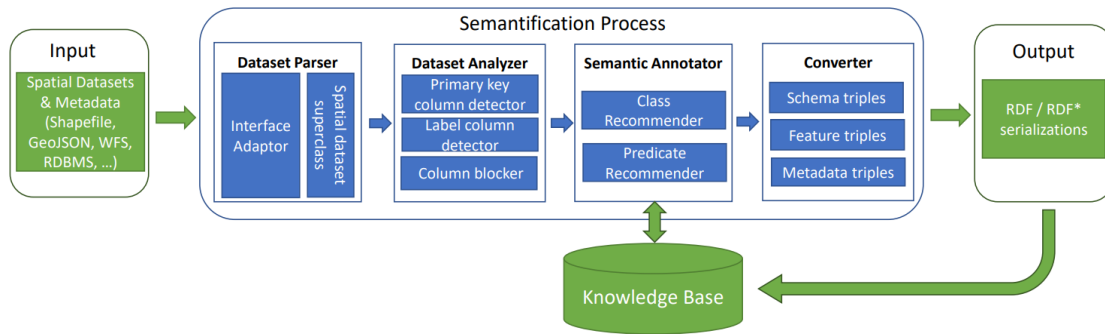


Figure 1: The semantification process

organizational Spatial Knowledge Infrastructures (SKI) [8]. The process is designed to handle a variety of spatial dataset formats and to manage both geometric and thematic attributes. A highlight of the process is the generation of triple-level metadata according to the RDF* model [9], that allows the enrichment of feature triples with useful provenance information. The process is applied on spatial datasets maintained by the University of the Aegean, Greece. However, the solution is generic and can be easily adapted by organizations that manage spatial data and want to build their own SKI. The paper concludes with a discussion on the initial results and pointers for future work.

2. Semantification Process

Fig. 1 depicts the overall process. The process is designed to accept as input various spatial dataset formats, such as Shapefile, GeoJSON, WFS and spatial RDBMS tables. Datasets may be heterogeneous, cover different thematic areas and are not required to follow a specific schema. In addition, the process is designed to handle XML files and CSW records in several dataset metadata standards, such as Dublin Core, INSPIRE or ISO19139, that may be associated with spatial datasets.

The semantification process begins with the parser module that contains an interface adaptor for each dataset format. The adaptor extracts schema-level information (i.e., a list of column names and their types) and the actual data (i.e., geographic features with their attribute values) of the dataset. Also, it extracts some basic metadata from the input file, such as dataset name, creation date, dataset format, publisher, description, geometry column, geometry type, original CRS and dataset spatial extent. Finally, if related metadata files exist, e.g. INSPIRE-compliant XML files, it tries to parse them. The module applies some string cleaning on dataset name and on column names in order to remove special characters and to convert non-Latin characters to Latin. All the above information and metadata (regardless the format of the input dataset) are modeled in a *spatial dataset superclass* that is used for the next steps of the process.

The analyzer module examines schema and data-level information of the dataset thematic attributes to decide a set of conversion options. First, it searches for candidate primary key columns so as their values to be used later for assigning URIs to geographic features. The list of

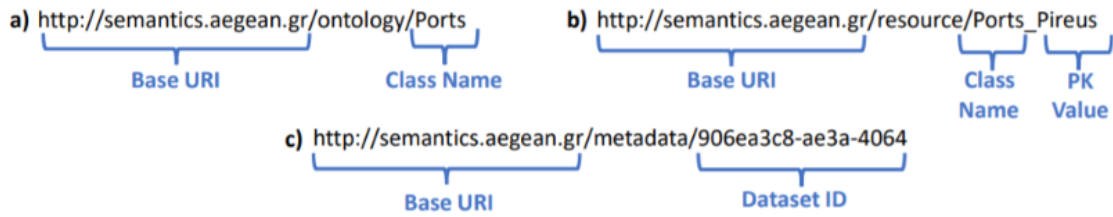


Figure 2: Examples of a) class URI, b) feature URI, and c) spatial dataset URI.

candidate primary keys is formed by integer and string columns that contain distinct not null values. String columns that contain large-length values are discarded from the list of candidate primary keys, since they usually provide additional verbal information about the geographic features. The analyzer then selects the most appropriate primary key column by giving priority to the candidate string columns. If no candidate columns are found, then no primary key column is selected. Second, the analyzer searches for columns with values that can be used later for assigning labels to features (using the `rdfs:label` predicate). The list of the candidate label columns is formed by string columns that contain small-length values. In addition, the analyzer tries to detect the language for each candidate label column. Third, the analyzer search for columns that will be blocked, i.e. not converted to RDF. The blocked columns list is formed by the label columns (because they are already annotated), by columns that may refer to foreign keys and by user-defined/selected columns. Possible foreign keys columns empirically set to be integer columns that contain not distinct values.

For the next step, the process annotates spatial data with suitable schema resources, extracted from an existing knowledge base. First, it tries to guess the type (class) of the spatial data by comparing the textual and semantic similarity between the spatial dataset name and existing classes in the knowledge base. For the textual comparison it applies the Levenshtein similarity and for the semantic the WordNet WuPalmer index. Classes that are similar above a certain threshold are ranked and the top-one is recommended to be used as the class for the spatial dataset. In case that no recommendations are generated, the user has the option to select another class from the knowledge base or to create a new class. In the latter case, the class will be added to the knowledge base and the user is opted to additionally provide a characteristic label and a description for the new class. The same procedure is applied to annotate the columns that will be converted (except for the label columns) with suitable predicates from the knowledge base. That is, the module searches for existing predicates with similar names to the column name and recommends them to the user. The user can keep the default recommendation, select an alternative predicate from the knowledge base or create a new predicate. In the latter case, the user is opted to provide a label and a description for the new predicate.

Finally, the spatial dataset is converted to RDF, containing triples about schema, features and metadata. Schema triples contain the definition of new classes and predicates that may have been created during the annotation step. Their URI is formed by the default base URI, the term “ontology” and the resource (class or predicate) name as depicted in Fig. 2a.

Feature triples contain the descriptions of spatial features. Specifically, each feature is assigned with a URI, formed by the default base URI, the term ‘resource’, the class name and the primary

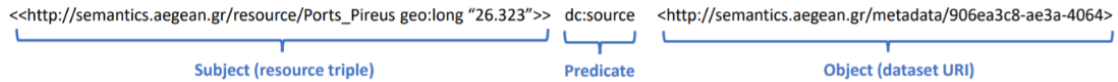


Figure 3: RDF* triple example

key value (Fig. 2b). If no primary key was found during the analysis step, a serial number is used. A feature (entity URI) is declared to be member (with the `rdf:type` predicate) of the annotated class and of the GeoSPARQL *Feature* class. Accordingly, the feature is associated with its labels using the `rdfs:label` predicate. For each of the rest columns, the feature is associated with the respective cell values using the annotated predicates. The feature's geometry is represented according to the GeoSPARQL vocabulary in the WKT serialization, which additionally embeds the CRS code of the geometry. If the geometry is not projected in WGS 84, the geometry is reprojected to WGS 84 and is also represented according to the GeoSPARQL vocabulary. Finally, if the geographic feature is point, then the geometry is also represented according to the W3C Basic Geo vocabulary (`long` and `lat` predicates).

Metadata triples contain the dataset-level metadata that were extracted during the parsing step. Each converted spatial dataset is assigned with a URI formed by the base URI, the term *metadata* and a randomly generated ID (Fig. 2c). Each dataset entity is declared to be member of a *SpatialDataset* class and is associated with its metadata (e.g., dataset name, publisher, creation date, spatial extent) using fixed predicates from the Dublin Core vocabulary.

The converter associate feature triples with the spatial dataset entity URI from which they originate. By this way, the knowledge base maintains triple-level metadata, allowing users to gain knowledge such as who created a piece of information and when. To this end, the RDF* model is used with the following syntax: 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 URI (Fig. 3).

The output of the process are files in RDF / RDF* serializations that can be directly imported in the selected knowledge base.

3. Implementation

The semantification API is implemented in Java. The GeoTools and JTS libraries are used for spatial dataset parsing and geometric transformations and the Apache Jena framework is used for RDF modelling and for sending SPARQL queries to the knowledge base, which is selected to be a Fuseki instance. On top of the semantification API, three applications, which share the same functionality but serve different system requirements, were developed: a) a desktop command-line, b) a desktop GUI and c) a web application. Screenshots of the web application are depicted in Figure 4. The applications provide user-friendly interfaces and target users with at least an elementary knowledge about semantic web concepts and about the content of the underlying knowledge base. The applications provide default recommendations for each step of the process, however they allow users to intervene by editing dataset metadata, selecting different primary key, label and blocked columns and annotating with alternative classes and predicates. Moreover, users have access to generic process customization options such as, setting

the URL of the default knowledge base, the base URI for annotation, the default output RDF serialization (e.g., RDF/XML, Turtle), and the CRS and spatial vocabulary (e.g., GeoSPARQL) that will be used for the geometric attributes transformation.

The semantification process was tested by converting more than a hundred spatial datasets to RDF and loading them to an initially empty knowledge base. The datasets are maintained by the University of the Aegean and stored in various institutional databases, SDIs and local computers. Since the current version of the applications support only shapefiles, the datasets were first converted to the shapefile format. Their thematic coverage span in several categories, including administrative units (prefectures, municipalities, etc), statistical indicators (population, GDP, etc), natural environment (rivers, volcanos, etc), natural phenomena (floods, earthquakes, etc), infrastructure (airports, hospitals, etc) and facilities (museums, schools, etc). Fig. 5 shows the semantification result for a 'ports' shapefile. Above is depicted the attribute table of the shapefile. Below is depicted the output, consisted of schema (lines 9-12), spatial dataset metadata (lines 15-23), feature (line 26-38) and RDF* triples (lines 40-44). For brevity, only the conversion for the Herakleion port and a subset of the RDF* triples are depicted.

4. Conclusion and Future Work

This paper described a semantification process for spatial datasets. Its main features are the provision of easy-to-use tools that semi-automatically build incrementally a semantic knowledge base and the adoption of the RDF* model for providing triple-level metadata. The design and implementation of the process are still in progress, thus, next we summarize our so far experience and we outline the roadmap for the further development. An initial assessment of the process indicates that 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. In particular, the annotation recommendations help users, without strong familiarity with the knowledge base content, to quickly determine the suitable classes and properties. Moreover, the process guarantees the instant population of the knowledge base with well-formed RDF that is ready to use. In the future, we plan to undertake more detailed experiments in order to evaluate the overall performance of the semantification process and its ability to populate high-quality semantic content.

As long as we are dealing with more spatial datasets, we will encounter with unseen disparities that will be valuable input for designing more sophisticated rules for the dataset analysis and the annotation steps of the process. For example, we may incorporate rules that better detect primary and foreign key columns, block duplicate columns or identify specific-content columns (e.g., telephones, emails) in order to annotate them with standard predicates from well-known ontologies. Also, as long as the knowledge base is populating with data, it will become 'smarter' regarding the semantic annotation recommendations. In this respect, we plan to improve recommendations by additionally employing instance-based methods that will recommend classes and predicates based on the textual, semantic and spatial similarity. In addition, the possibility of integrating APIs of third-party semantic web search engines (e.g., Linked Open Vocabularies [10], GeoLOD [2]) could be investigated in order to augment the annotation recommendations with classes and properties from external knowledge bases. Currently, the

semantification process primary focus is on the basic transformation aspect and puts aside other important tasks such as entity reconciliation and ontology alignment. This decision was taken in order to not burden the process with additional operations that could distract the user from the main process and diminish the performance of the process. However, these tasks will be part of a post-processing process that will perform some cleaning (e.g., merging URIs that refer to the same entity within knowledge base or finding literals that can be substituted by entity URIs), establish sameAs links between local and external instances, and perform ontology alignment in order to detect equivalency or hierarchy relations between local and external classes and properties. Regarding dataset metadata, these are currently represented using terms from the DC vocabulary. However, we plan to examine and standardize the vocabulary for representing spatial dataset metadata (e.g., by adapting the INSPIRE metadata template to RDF). Also, triple-level metadata are maintained according to the RDF* model, which in its default syntax requires the creation of additional triples and complicates the formation of SPARQL queries. A possible solution to these limitations would be the adoption of the alternative RDF* annotation syntax. Lastly, for the annotation of the geometric attributes, we use the GeoSPARQL and the W3C Basic Geo vocabularies since they are widely used and easy to use. However, in future versions we plan to provide the option for selection of additional spatial vocabularies.

Acknowledgments

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].

References

- [1] J. Nowak Da Costa, J. Nogueras-Iso, S. Peedell, Issues of multilinguality in creating a european sdi-the perspective for spatial data interoperability (2005).
- [2] V. Kopsachilis, M. Vaitis, Geolod: A spatial linked data catalog and recommender, *Big Data Cogn. Comput.* 5 (2021) 17. doi:10.3390/bdcc5020017.
- [3] W3C, A direct mapping of relational data to rdf, 2012. URL: <https://www.w3.org/TR/rdb-direct-mapping>.
- [4] W3C, R2rml: Rdb to rdf mapping language, 2012. URL: <https://www.w3.org/TR/2012/REC-r2rml-20120927/>.
- [5] K. Kyzirakos, D. Savva, I. Vlachopoulos, A. Vasileiou, N. Karalis, M. Koubarakis, S. Manegold, Geotriples: Transforming geospatial data into RDF graphs using R2RML and RML mappings, *J. Web Semant.* 52-53 (2018) 16–32. doi:10.1016/j.websem.2018.08.003.
- [6] K. Patroumpas, M. Alexakis, G. Giannopoulos, S. Athanasiou, Triplegeo: an ETL tool for transforming geospatial data into RDF triples, in: K. S. Candan, S. Amer-Yahia, N. Schweikardt, V. Christophides, V. Leroy (Eds.), *Proceedings of the Workshops of the*

- EDBT/ICDT 2014 Joint Conference (EDBT/ICDT 2014), Athens, Greece, March 28, 2014, volume 1133 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2014, pp. 275–278.
- [7] C. Prudhomme, T. Homburg, J. Ponciano, F. Boochs, C. Cruz, A. Roxin, Interpretation and automatic integration of geospatial data into the semantic web, *Computing* 102 (2020) 365–391.
- [8] I. Ivánová, J. Siao Him Fa, D. A. McMeekin, L. M. Arnold, R. Deakin, M. Wilson, From spatial data to spatial knowledge infrastructure: A proposed architecture, *Transactions in GIS* 24 (2020) 1526–1558.
- [9] O. Hartig, P.-A. Champin, Metadata for rdf statements: The rdf-star approach, 2021. URL: http://www.lotico.com/index.php/Metadata_for_RDF_Statements:_The_RDF-star_Approach.
- [10] P.-Y. Vandenbussche, G. Ateazing, M. Poveda-Villalón, B. Vatan, Linked open vocabularies (lov): A gateway to reusable semantic vocabularies on the web., *Semantic Web* 8 (2017) 437–452. URL: <http://dblp.uni-trier.de/db/journals/semweb/semweb8.html#VandenbusscheAP17>.

1
2
3
4

Upload
Analysis
Annotation
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:	Mon Feb 28 13:01:51 UTC 2022
Format:	Shapefile	CRS:	Greek_Grid
Publisher:	<input type="text"/>	Geometry:	Point
Description:	<input type="text"/>	Bounds:	References[Envelope[659079.4150601482; 721535.8176560029; 4330523.150549255; 4359666.74695738]]

Features Preview

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

Conversion Options

Select a Primary Key: 1

Select label columns: 1

Select columns for conversion: 1

NAME

NAME
PREFECTURE

OIKLID
NAME
PREFECTURE
CODE

Back
Next

1
2
3
4

Upload
Analysis
Annotation
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** 1

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

Predicate Annotation

Annotation options for the column: **PREFECTURE** 1

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

Predicate Name:
 Predicate Label:
 Predicate Description:

Back
Next

Figure 4: Semantification web application. Example of the dataset analysis (top) and semantic annotation (bottom) steps.

"Ports" Shapefile

OIKI_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/EPSG/0/4326>POINT (26.15680836273347 39.36908053997243)" .
37
38 _:b3 geosparql:asWKT "<http://www.opengis.net/def/crs/EPSG/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> .
```

Figure 5: Semantification result