Semantic Modeling of Official Statistics - The case of the Greek Statistics

To allow for blinded review:   
do NOT indicate author information or affiliation

**Keywords:** Semantic Modeling, Official Statistics Ontology

# Introduction

The institutional and legal framework of the Hellenic Statistical System - ELSS in combination with the huge number and complexity of the statistics produced by its stakeholders, requires full support from computing systems through the automation of the information retrieval and knowledge management. To this end, we will examine in this paper the possibility provided by the semantic web software tools. In particular, in order to model the field of ELSS, we developed an OWL-ontology using the Protégé, and present the search results on the data of the corresponding knowledge base that we created in parallel with the ELSS ontology.

# Methods

Coordination of the services of the Ministry of Interior in order to support the program of certification of the produced statistics by Hellenic Statistical Authority - ELSTAT [1] highlighted specific management issues within the organization within the responsibilities of the Statistical Head [2]. Among the issues that triggered this work are the following.

* Statistics production services are characterized by heterogeneity in terms of administrative structure and on the processing methodology and data analysis
* a large number of regulatory provisions on the production of statistical products which often produce conflicting results
* non-interoperable software applications for processing and managing statistical products
* difficulties in understanding the terms and procedures for certifying statistics
* non-standardized procedures both during production and dissemination of statistical products
* possible overlaps in the production and diffusion of statistics.

Semantic Web technologies have very positive results to offer in addressing these issues and are mainly related to the modeling of knowledge and the intelligent management of statistics. In this paper, we focus on the modeling of knowledge in the field of official statistics of the Greek Statistical System through the development of a corresponding OWL ontology. We also highlight the possibility of intelligent search on the data stored in the knowledge base we developed along with ontology.

## Recording and presenting basic concepts of the ELSS

In accordance with the statistical law [2] and the provisions of the institutional and legal framework, as detailed on the official website of the ELSTAT, the ELSS bodies, which produce and disseminate statistical products, have the obligation to certify as official statistics by ELSTAT. An important parameter of the success of this effort is the understanding of the terms and procedures as well as the implementation of concrete steps / actions by the stakeholders. In detail, the definitions and necessary steps have been exhaustively depicted in the relevant circulars and instructions of ELSTAT [3].

As indicative terms and procedures we mention:

* The main components of the ELSS are the "National Authorities of ELSS" and the "ELSS bodies". In detail the composition of the two entities refers to the relevant decisions of the President of ELSTAT [3].
* Each statistical product produced by the ELSS institutions after being certified by ELSTAT is characterized as "Official Statistics", as provided for in the relevant provisions of the institutional and legal framework.
* Each statistical product is characterized by a unique code that follows it throughout its lifecycle, irrespective of the operator, the subject and the domain that belongs to it as well as the producer that produces it. These data are recorded in the corresponding tables of ELSTAT.
* Producer and production manager is considered to be the service belonging to Ministries or Independent Authorities or other bodies of the Greek Public Administration.
* Each entity in the system can be considered either as a producer or as a user of statistics on a case by case basis. In the meaning of the producer, it is understood that the body is independent of the organizational unit that has the substantive competence, whereas the user can be understood as either an entity or individuals / citizens.

## Developing an ontology for the modeling of knowledge in the field of Official Statistics of the ELS

Information retrieval with innovative - intelligent methods and the automated management of the knowledge produced by them have as a prerequisite the appropriate modeling of information and knowledge. An important factor in this direction is the annotation of data, using semantic terms and methods of the semantic web. Following the primary step in the study of the ELSS environment, which was the recording of the basic concepts, correlations and procedures of the field, we developed the ELSS ontology in order to create a knowledge base. The development of the ontology and the feeding of the base was done using Protégé 5.1.0, a modern open source tool [4].

The methodology we have followed includes the following stages:

* Initially, we developed all the necessary entities that corresponds one another to the ELSS concepts we developed above.
* Then we created the necessary snapshots to complete the set of entities and create the knowledge base.
* Then, the most important step of the process is the semantic annotation entities and instances. Properties and relations between entities play an essential role in rendering semantics to the concepts and entities of the database and is the key difference from a relational database. Their appropriate use also supports the efficient production of RDF triplets that essentially support the open and interconnected distribution of our information.
* We also defined, where appropriate, restrictions on the domain and range of classes as well as specific rules for extracting conclusions on the Protégé SWRL tool.

As mentioned above, the ELSS entities are in full correspondence with the classes of ontology as shown in figure1.

* We have used the *“Person”* class to attribute the roles of staff and citizens involved in the production and use of statistics respectively, while the *“Producer”* class corresponds to the bodies that produce and disseminate statistics.
* The *“Statistics”* class corresponds as the basic taxonomy of the statistics produced, based on the classification set by ELSTAT [3]. Individuals of the class and its subclasses are the statistics produced by the ELSS agencies.
* The *“OfficialStatistics”* class implements the “official statistics” entity
* The core components of the ELSS are implemented with the *“AgencyOfELSS”* and *“NationalAuthority”* classes.

Particularly important are the properties we have defined and which essentially materialize the semantic annotation of the entities figure1. From the properties we have developed, we consider as important, the *“produces”* property, used to implement the basic function of the ELSS in the production of statistics. Correspondingly, the inverse *“isProducedBy”* property is used to mark the relevant statistics produced by the operators.

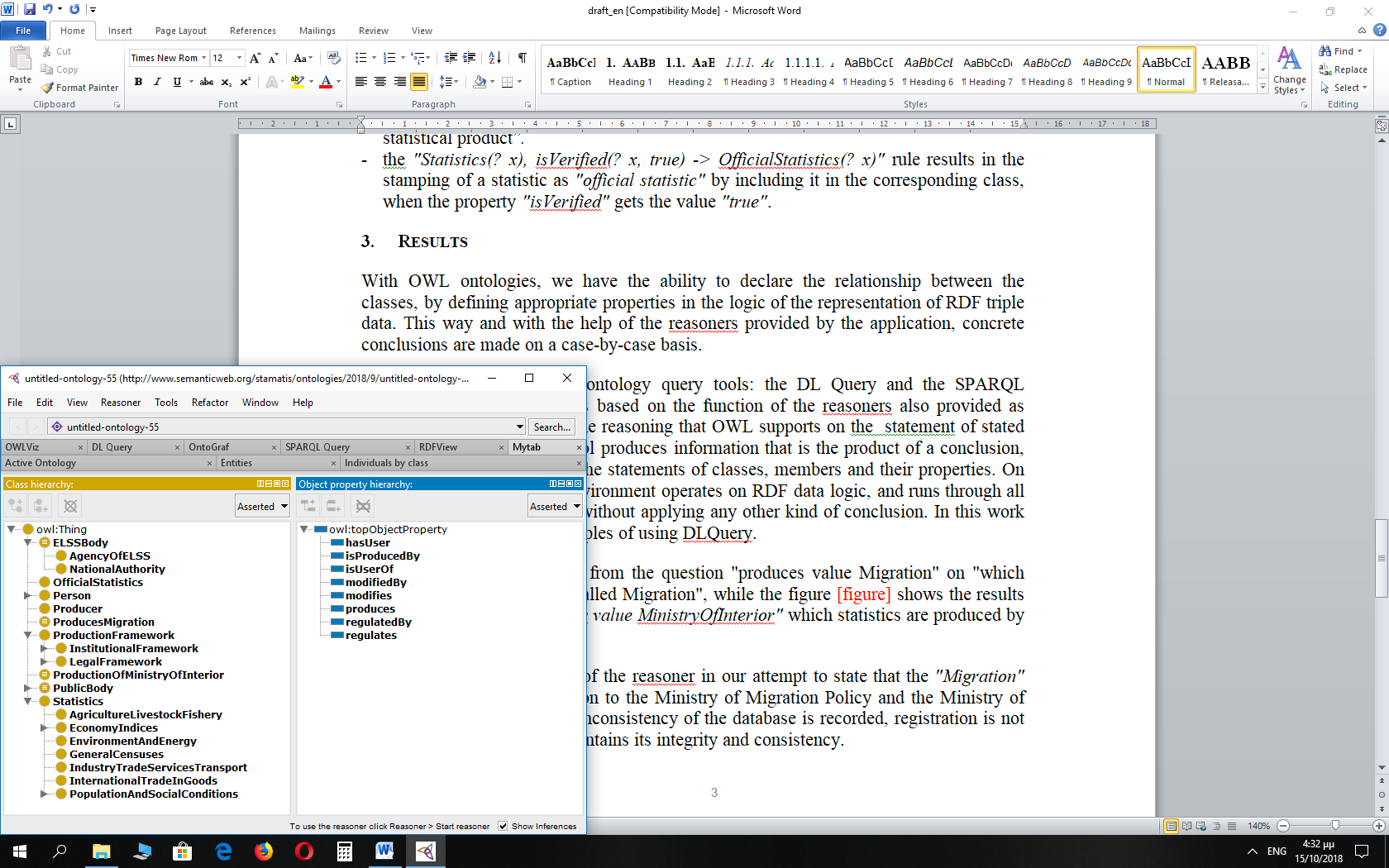


Figure 1. Class and Object hierarchy

Important for automated production of information and knowledge from our ontology is the definition of rules through the SWRL language and the corresponding Protégé tool as well as property statements through the domain and range definitions as appropriate. Another way of limiting is the statement of the equivalent class we have used for the class *"ELSSBody"*. For example:

* the *"produces(? x,? y)->Producer(? x)"* rule results in automatic updating of the Producer class when using the reasoner provided by the application by the respective individuals, implementing the logic that "Producer is someone who produces a statistical product”*.*
* the *"Statistics(? x), isVerified(? x, true) -> OfficialStatistics(? x)"* rule results in the stamping of a statistic as *"official statistic"* by including it in the corresponding class, when the property *"isVerified"* gets the value *"true"*.

# Results

With OWL ontologies, we have the ability to declare the relationship between the classes, by defining appropriate properties in the logic of the representation of RDF triples. This way and with the help of the reasoners provided by the application, concrete conclusions are made on a case-by-case basis.

Protégé provides two main ontology query tools: the DL Query and the SPARQL environment. The first tool is based on the function of the reasoners also provided as protégé plung-ins, and uses the reasoning that OWL supports on the statement of declared properties. The use of this tool produces information that is the product of a conclusion, without directly arising from the statements of classes, members and their properties. On the contrary, the SPARQL environment operates on RDF data logic, and runs through all of the declared RDF triplets without applying any other kind of conclusion. In this work we present some typical examples of using DLQuery.

Figure2a shows the results from the question "produces value Migration" on "which body produces the statistics called Migration", while the figure2b shows the results from the query "isProducedBy value MinistryOfInterior" which statistics are produced by the Ministry of the Interior ".

Figure2c shows the results of the reasoner in our attempt to state that the *"Migration"* statistic is produced in addition to the Ministry of Migration Policy and the Ministry of the Interior. In this case, the inconsistency of the database is recorded, registration is not accepted and the database maintains its integrity and consistency.

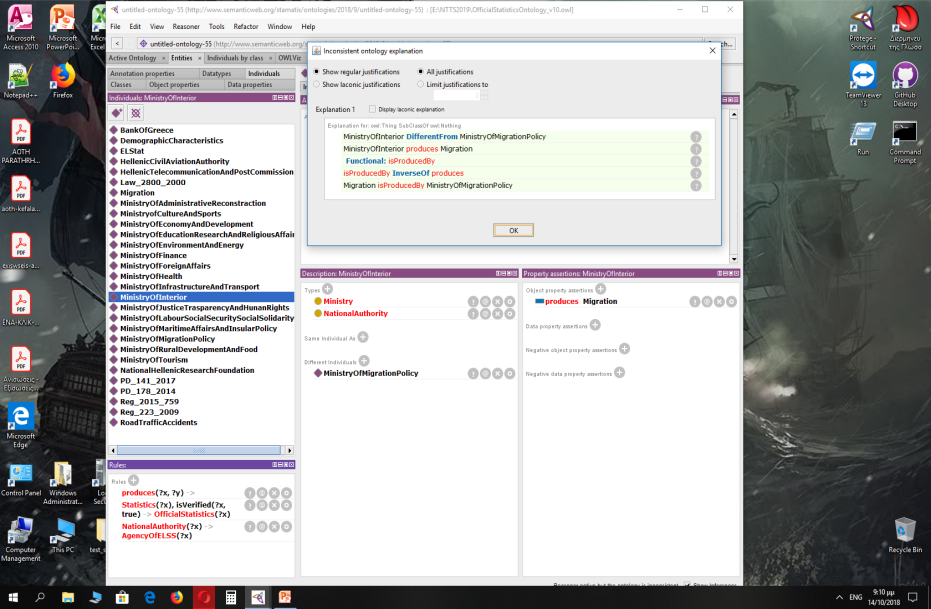
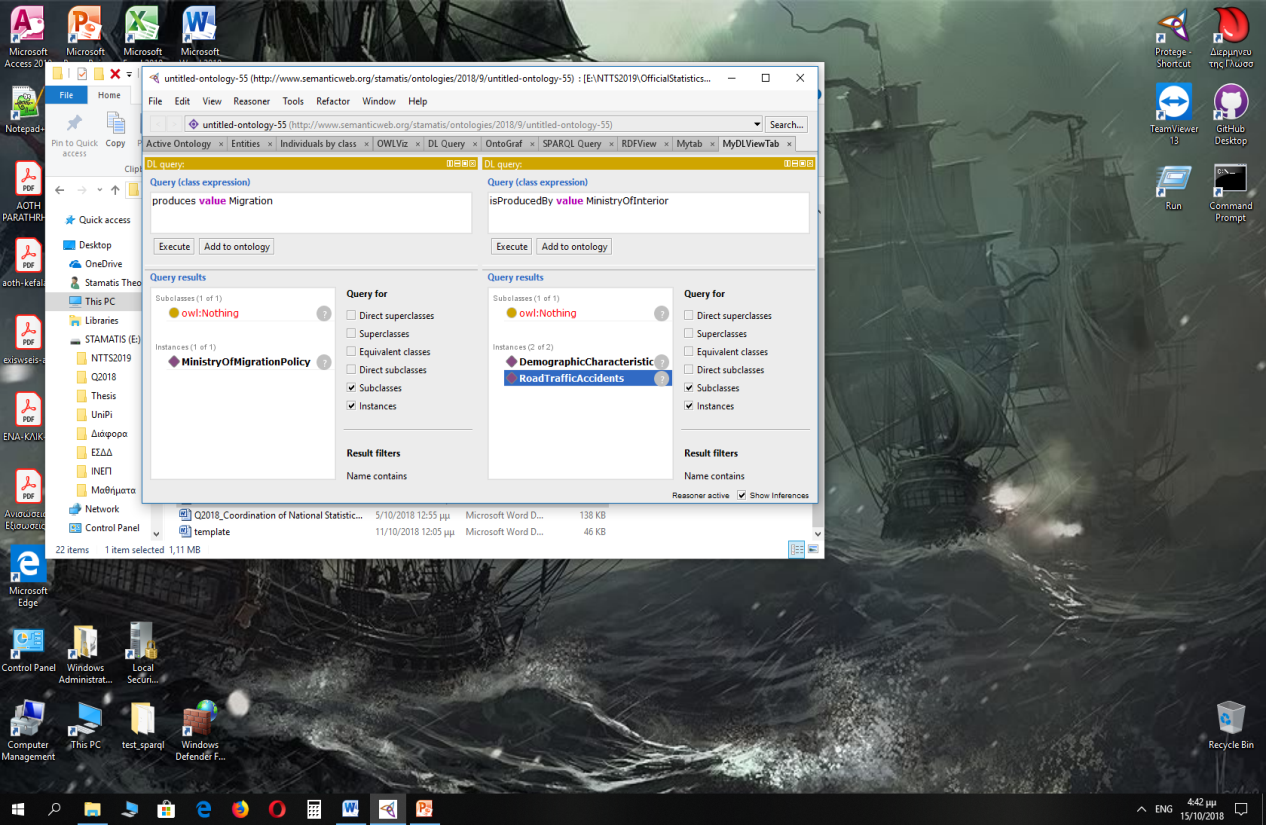


Figure 2a, 2b : Results from DLQury Figure 2c : Results from DLQuery

# Conclusions - Assessment & evaluation of ontology

The base we developed was tested using the reasoners tools provided by Protégé and the ontology is checked for its consistency and integrity. The ontology and the corresponding basis we developed are useful in many ways, including a) the possibility of semantic search of information based on relations between classes and between individuals, b) as a training material for those involved in the development and dissemination of statistical public bodies; c) the visualization of a complex system of concepts, entities and associations, d) future use of the database as an input file for the production of linked statistical data. This was made possible by modeling the relative knowledge of the object in the form of an OWL-ontology. In terms of ontology assessment, we have to mention the following advantages of the proposed modeling.

* Simplicity of class statements, instances, and relationships between them
* Ability to directly enrich the base without requiring a change of the existing structure from the beginning.
* Consistency with the actual ELSS data
* Possibility to query based on reasonable conclusions

References

1. Hellenic Republic, (2010), Statistical Law No 3832/2010, as in force
2. Christina Pierrakou, Stamatios Theocharis, Coordination of National Statistical System – Experience from the Hellenic Statistical System (ELSS), Q2018
3. Official site of ELSTAT, [www.statistics.gr](http://www.statistics.gr)
4. Official site of Protege, http://protege.stanford.edu/about.php#citing