Ontology based semantic model for health data interpretation

Christoniki Maga-Nteve[1], Nikos Tsolakis[1], Georgios Meditskos[1,2],
Anastasios Karakostas[1], Stefanos Vrochidis[1] and Ioannis Kompatsiaris[1]

[1]Information Technologies Institute, Centre of Research & Technology, Greece
{chmaga, tsolakin, akarakos, vrochidis, ikom}@iti.gr
[2]School of Informatics, Aristotle University of Thessaloniki, 54124, Greece {gmeditsk@csd.auth.gr}

Abstract. New opportunities for improved personalized healthcare have emerged due to the recent advances in the development of modern methods which reinforce personalized early risk prediction, prevention and intervention. Using semantic techniques for data integration has become pivotal as it can deliver different ways to represent data, automating the process of data integration, and providing the ability to query semantically. In this paper, we propose a new semantic data model in which health information derived from Parkinson's, Multiple Sclerosis, and Stroke (PMSS) patients is systematically analyzed to generate and improve knowledge that will be transferred to patient care in order to design and develop innovative health risk prediction and intervention tools. Furthermore, this project focuses on providing new opportunities for improved personalized healthcare and prevention that have been created by new designs and developments of innovative health risk prediction and intervention tools. A core ontology is currently being designed within the ALAMEDA project to deal with the semantic interoperability across heterogeneous datasets along with a semantic framework to concrete the generated heterogeneous data through a shared ontology. The ontology model development and the requirement elicitation will be done based on the components' capabilities and use case requirements. The heterogeneous and dynamic data will be subjected to annotation through the development of semantic models for data sharing and usage apart from being interpretable.

Keywords: Semantic interpretation, Ontology, Health-care data, Data interoperability

1 Introduction

Health-care ontologies are pivotal for knowledge representation and data integration as the health data have become very complex and there is an intense need to link disorders and applicable medication along with specific individual patient attributes so as to extract meaningful results. The use of ontologies facilitates easier processing of large datasets, while providing more effective solutions which support the way we manage health and wellbeing and the indispensable integration of knowledge and data [1]. In the healthcare domain, ontologies organize the knowledge as relations and instances to encode health records, lab results and diagnoses of patients while a specific data structure is required to generate the appropriate information and solution. They can also add context to the patient's data and provide a common framework for sharing and reuse of meaningful clinical outcomes.

Semantic models are able to describe the health-care concepts and the relationships between them by improving their effectiveness and efficiency. They have profited healthcare communities with methods based on multiple ontologies, assuring the data quality in a heterogeneous environment and organizing them so that it can be interpreted by computers without human intervention. Using semantic techniques for data integration has gained a lot of ground as they provide automated and multiple concepts to represent and process the data, and allow for the semantically query [2],[1]. With the goal of speeding up the modelling development process, a variety of possible knowledge resources can be reused. This approach has given different benefits to the developers, however, the existing methods and tools are not enough to guarantee a successful model. So, all these resources need to be evaluated in regard to their context-oriented usability and to adopt the requirements derived from the ontology needs.

In this paper, we present an Ontology based semantic model for health data interpretation, where the proposed model is able to harmonize information from multiple sources to provide context awareness. The use of the ontologies and the semantic web technologies will allow us to provide a conceptual model, supporting interoperability and flexibility. The ontologies will also address the semantic interoperability issues, management and integration of information models, as it will define the concepts and their relationships within the ALAMEDA's domain. The project's innovations will utilize new machine learning models, built upon lifestyle retrospective data as well as new streams of patient data that involve the monitoring of everyday activities, such as sleep behavior. The success of such applications will provide clinicians with the opportunity to modify interventions based on personalized data recordings. The main goals of this study are to implement the semantic models development for the annotation of the different modalities and to provide innovative solutions for context-aware data aggregation. For the purpose of this work, a health-care ontology is being designed to deal with health-care information and IoT devices and services.

The four steps based on Bravo, Reyes, Ortiz [3] that we follow to identify the modelling requirements are: a) the ontology requirements specification [4],where the main purpose is to define the requirements that the ontology should cover, b) the ontology design and c) construction, which are defined as an ordered series of phases that specify the procedures used in the engineering of an ontology or ontology system [3] and d) the ontology evaluation which is ensured in view of quality and correctness perspectives[5].The remainder of this paper is organized as follows. Section 2 summarizes the state-of-the-art on methodological reuse of existing technologies. Section 3 describes in detail how we draw out the requirements and the guidelines we follow based on the ALAMEDA needs. Sections 4 position our conceptual model and finally Section 5 concludes and presents some of our future work.

2 Related Work

The development of semantic web technologies provides a number of ontology-based approaches in different domains. Within the Semantic Web community, it is strongly encouraged to reuse existing ontology models. The main domains that can be covered are: Sensor Data, Context, Activity Recognition, Event and Healthcare ontologies.

Thus, we can build on existing resources and expand them, based on the specific needs of the ALAMEDA project, where is necessary. There are several state-of-the-art ontologies that can be utilized for modelling ALAMEDA's domains. In particular, some of the most commonly used ontologies, fused with healthcare interoperability standards, are the Fast Healthcare Interoperability Resources-HL7 (FHIR-HL7) [6] and the Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) [7]. The FHIR-HL7 describes the Resource Description Framework (RDF) representation of FHIR resources, while the SNOMED CT is a comprehensive medical terminology used for standardizing the storage, retrieval, and exchange of electronic health data and for the representation of medical concepts respectively. Additionally, the International Classification of Diseases and Related Health Problems (ICD-10) [8] ontology aims to create a knowledge base for use in the ICD coding system is it also frequently used. Other related ontologies are the PDON, a Parkinson's disease ontology for representation and modeling of the disease knowledge domain [9] and MSO, a multiple sclerosis ontology [10] integrated with Basic Formal Ontology [11]. While, much progress has been made in developing semantic models for healthcare interpretation, there is great potential for developing models about the three diseases that ALAMEDA addresses. The ALAMEDA ontology is concerned with creating a model that responds to the needs of patients with Stroke, MS and PD, providing semantic interoperability with respect to the personalized use cases of the project.

The Semantic Sensor Network ontology (SSN) [12], is used for the representation of sensors and sensor-like devices. The base of SSN is the Stimulus-Sensor-Observation pattern, a cornerstone for heavy-weight ontologies for the Semantic Sensor Web applications. The newest version of SSN, the Sensor, Observation, Sample, and Actuator (SOSA) allows the representation of sensors as a light-weight model [13]. The Semantic Smart Home [14] ontology captures knowledge relevant to activities of daily living, location, timing and people. The Event Model F is a formal model of events designed to facilitate interoperability in distributed event-based systems [15] and the Event Ontology [16] deals with the notion of reified events. The Dem@care Project contributes to the timely diagnosis, assessment, maintenance and promotion of self-independence of people with dementia [17]. The aforementioned ontologies can cover a subset of the domains involved in Healthcare systems and applications. Our proposed ontology seeks to respond to every aspect by reusing resources, comprised of modules for representing every need and can be easily adjustable and reusable.

3 Modelling Requirements

Ontologies can be defined as an explicit specification of a conceptualization [18], where a logical formulation of complex problems is provided. For a solid design of an ontology, it is essential to define a variety of stages. Some of them are the reuse of existing ontologies, the class definition and the relations between them. Initially, the domain and the scope of the ontology have to be determined based on their intended uses. The existing ontologies can be integrated into the ontology framework so that the new ontology will be developed from current dictionaries. The next step is to define the classes which the ontology will consist of and to allocate them in a hierarchical mode. Also relevant properties that characterize the relationships between the classes have to be introduced along with their individual instances.

This study will provide the annotation layer to endorse ALAMEDA with situational awareness, extracting and harnessing deep intelligence through the aggregation of heterogeneous information and knowledge. It will also allow the comprehensive representation and the harmonization between the IoT devises, while it will offer innovative solutions development for context-aware data aggregation, enabling the semantic federation of diverse infrastructures and services for capturing information.

In this context, we are developing a health-care and patient-oriented ontology to personalize the medical and social knowledge available. The ontology will provide knowledge structures that are maintainable whilst they will be used to support clinicians in multiple tasks. The objective is to enable sensors and data coming from multiple distributed information sources to be semantically accessible and discoverable, fostering the development of data processing applications that effectively utilize and combine multiple data sources and devices to deliver innovative services.

4 ALAMEDA's Conceptual Model

In ALAMEDA project, we will design a consensual and conceptual model able: a) to represent information that is made available via the questionnaires and the monitoring modules, b) to ensure the semantic interoperability of the information exchanged between the individual ALAMEDA systems components, and c) to achieve the semantic annotation of the generated data and to further extend it with domain knowledge pertinent to the ALAMEDA use cases. It is crucial for the ALAMEDA model to represent information related to the domains of the project such as physiological and cognitive assessment, clinical data, reported difficulties etc. Within the framework of the ontology design, we reused the Even Model F ontology and the Event Ontology, so as to construct the ALAMEDA Event module. The Event Ontology models the events, the environment and the changes that may happen and answers a series of critical questions about the actions, the places, the time and the person of an event. During the Sensors Ontology construction, we reused the SSN Ontology, which as it has been already mentioned is a fundamental ontology in the sensor representation domain.

4.1 Methodology Overview

There are several methodologies for ontology engineering to formalise and design an ontology. In this study, in order to develop the ALAMEDA's Model, we used the NeOn Methodology, which is based on a set of 9 scenarios and is well-documented and highly adaptable. While building the ALAMEDA Ontology, different phases have been determined. The first phase refers to the ontology requirements and the retrieval of the ontology requirements specification document. The role of domain experts is very important at this stage, since they define the use cases and propose optimal matching to ontology requirements as the model is ongoing. The second phase is related to the development of ontology at a primary level, where it will be defined which existing ontologies will be used, along with the information input. The third phase contains the implementation and enrichment of the ontology, using the OWL2 [19] language for knowledge representation, which provides Properties, Classes and Individuals.

4.2 Ontology Requirements Specification Document

The identification of the purpose of ALAMEDA Ontology, the scenarios that must be defined, the intended uses and end-users are the main elements of the specification of requirements. As proposed in the NeOn Methodology, the methodological guidelines were created, based on the state-of-the-art ontology development techniques and represented in the Ontology Requirements Specification Document (ORSD). The requirements are defined with respect to the Competency Questions (CQs), which are groups of questions that play a significant role in the ontology implementation, as they specify what knowledge has to be entailed in. A critical attribute of the CQs is that they define the functional requirements of the ontology. The ORSD is the output of the ontology requirements specification and provides information regarding the a) purpose which is the main general goal and function that the ontology should fulfill, b) scope which refers to the coverage and the degree of details that the ontology should have, c) implementation language which is the formal language that will be used for the ontology design, d) intended end-users and uses and e) the non-functional and functional requirements. In ALAMEDA project, there are clinical aims of the use cases that are critical components for the ontology framework. Parkinson disease is a common neurodegenerative disorder and its meaningful worsening of global status or of individual motor is a specific use case with inclusion criteria such as advanced Parkinson, age and more. Relapse risk prediction in Multiple Sclerosis in young to middle age person is a second important use case. The Stroke use case refers to patients who suffered from Stroke the last month and is monitored for neuro-rehabilitation. The proposed Ontology should be constructed with respect to these use cases so as to provide a shared vocabulary for the communication and exchange of information among the different system components, to represent, store and retrieve patients' profile data, sensors etc and to represent and query data made available by other analysis components of the ALAMEDA system.

4.3 ALAMEDA Ontology

The ORSD is a key factor for modelling the classes, properties and instances of the ontology. Moreover, we used further input and feedback made available during the stage of the ontology design by the clinical and technical partners involved in the relevant project. Another element for the current version of the ontology was the standards and best practices in the Semantic Web community which are available. The implementation of the ontology was done in Protégé [20], a tool for modelling ontologies that provides us with the ability to construct the appropriate modules.

ALAMEDA Ontology Modules
The ALAMEDA Ontology model consists of six modules and a main ontology, which acts as the parent of all the hierarchical relations. **The Model Ontology** represents all

the modules attached to ALAMEDA, where Home, Person, Lab, Event, Time, Sensors modules are some of the main classes. In Figure 1, a high-level figure of the model can be seen.

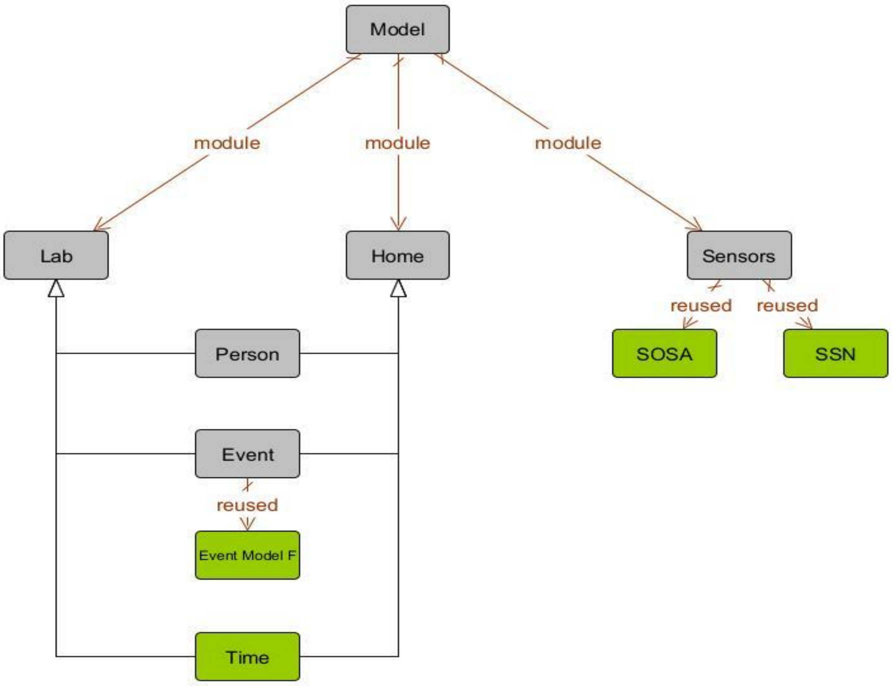

Fig. 1. A high-level figure of the ALAMEDA model

Home Ontology formalizes information relevant to behavioural interpretation and reported difficulties in the home environment. An example is the class *ReportedDifficulties* which is used to describe information about the problems (e.g. difficulty in exercising or bad mood etc) that a patient may face in the home environment. **Lab Ontology** formalizes the types of information relevant to the tests, assessments, patient's clinical & experimental records in the lab environment. The class Domain is used to provide information relevant to the specified domains, in order to describe the different types of clinical tests and their results. The class *MeasuredData* indicates the data that is essential to be shared during a task, while the class Task represents the possible types of tasks involved in the ALAMEDA. Finally, the class *ClinicalAssessment* is utilized so as to define the clinical characteristics that are collected during the clinical and medical phases taking place in the lab environment. **Person Ontology**: refers to patient's, clinician's and caregiver's socialdemographic data. It consists of 5 classes that display person, disease, gender, educational level and language. **Event Ontology** provides information relevant to the entities and the activities that take place in order to fulfill ALAMEDA purposes. The class *Entity* describes all the physical entities and consists of 2 subclasses *Person* and *Place*. The class *Activity* represents any activity the patient may be involved, while the class *MeasurementPattern* enacts the domain of the measurements that take place. Event Model F played a crucial role in the development of this module by using the participation pattern and making clear the roles and events in our model, while the Event Ontology describes the time, the agents etc of the ontology. **Sensors Ontology** describes information concerning the type and properties of the sensors used which are divided into two major groups: a) *FixedSensor* and b) *WearableSensor*. The class *FixedSensor* refers to sleep or location monitoring and the class *WearableSensor* refers to sensors like smartwatches or belt sensors that will be used. During the Sensors Ontology construction, we reused the SSN Ontology. The concepts introduced by SSN are very important in healthcare sensing environment. Main concepts that being reused are the class *Procedure,* which provides a way to specify observation and has an input and an output. Those input and output information is represented in the classes *Input* and *Output.* Other critical components of the SSN Ontology that being reused are the *Observation*, the *Platform*. **Time Ontology** presents the time, duration and information about the tasks of the ALAMEDA. It consists of classes like *DateTimeDescription*, *DayOfWeek* etc that provide specific information about date, time, day, duration and their values. Table 1 presents some of the most important object/data properties of our model, their definition, the module/class they exist in and one of their relations.

Table 1. ALAMEDA Modules Classes, Object/Data Properties and their use.

Module	Class/Subclass	Class	Name	Definition	Type
Lab	EDSS	MS	isFor	Indicates the disease that a test is used for	Object Property
Event	Activity	Person	hasAgent	Indicates which person is the performer of an activity	Object Property
Patient	ReportedDifficulties	Patient	forPatient	Indicates which patient is responsible for each self-assessment	Object Property
Lab	MocaTest		Type	Indicates the type of the test	DataProperty
Home	ADLSummary		Date	Indicates the date of the ADL activity	DataProperty

In Figure 2, an example of upper-level vocabulary for modelling a clinical test can be observed [21]. The class *CVLT-II* represents a measure of verbal learning and memory that is in Domain III – Mental and Cognitive Ability of the ALAMEDA Clinical Ethics. It is essentially a test for patients that suffer from MS. It relates the class *MS*, which is a subclass of the class *Disease*, via the object property isFor. In this way, we represent the tests that take place in the ALAMEDA Project and their relationship with a specific disease. The class *CVLT-II* has some data properties that provide information about the type of the test and the score that the patient has in each respective test. In this example, data property score is an integer, and provides information about the score and data property type:PM, which is an integer, provides information about the type of the test as dictated by the experts.

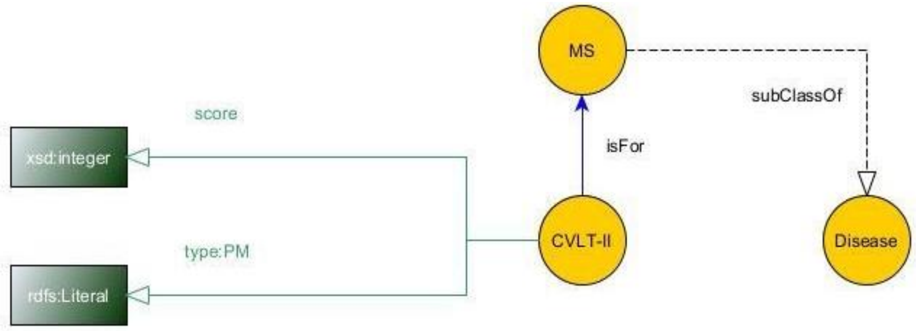

Fig. 2. An upper-level vocabulary for modelling a clinical test

4.4 Ontology Evaluation

In this section, we present the evaluation of the ontology, considering the quality, the structure and the consistency. The metrics of the current version of the ALAMEDA ontology, as provided by the ontology metrics view in Protégé can be seen in Table 2. The number of the classes, axioms, object properties, data properties etc, are considered base metrics and provide information regarding the quantity of the ontology components. There are 253 Classes, 67 Object properties and 42 Data properties.

Table 2. Base metrics.

Base Metrics	Value
Class count	253
Object property count	67
Data property count	42
SubClassOf axioms count	268
Disjoint classes axioms count	2
Inverse object properties axioms count	1

As the model is ongoing, its evaluation will be done by using some of the most common tools and methodologies. OntoClean is a methodology which validates the taxonomic relationships from the ontological adequacy standpoint [22]. It provides characterization of the basic elements of the ontology by using ontological notion. Furthermore, one of the most important tools for evaluating the consistency of the model is OOPS! (OntOlogy Pitfall Scanner), a tool that detects pitfalls and their consequences in the quality of the ontology and provides modifications and improvements of the pitfalls

[23]. This system provides pitfalls of different significance and categorize them in critical, important and minor pitfalls. Critical are the ones that is essential to be corrected, important are the ones that is not crucial, but are important to be corrected and minor pitfalls are the ones that are not crucial by any means, but their correction will provide quality to the ontology. The structure of the ontology is evaluated by using OntoMetrics, an online framework that provides information about the base and schema metrics of a semantic model [24]. Base metrics are the simple metrics, like the counting of classes, objects, etc, while schema metrics evaluate the design of the ontology. Some of the most common metrics that are used for evaluating the ontology using OntoMetrics are attribute richness, inheritance richness, relationship richness, axiom/class ration and class/relation ratio. Further steps, such as ontology population will enrich our model and will allow us to design and develop innovative health risk prediction and intervention tools. Classification tasks such as diagnosis prediction with taxonomical knowledge found in the ongoing ontology will be combined in order to support human-understandable explanations of the analysis.

5 Conclusion

One of the most challenging problems in healthcare systems is the interoperability between heterogeneous data, where a medium to share knowledge and exchange information both across people and services is essential. The Semantic Web services provide interoperability standards and vocabularies that can facilitate access to the necessary data in a secure and safe manner.

This paper describes a healthcare ontology-based model which interoperates between the systems and it is able to facilitate knowledge sharing in a complex environment. It is also able to manage and integrate patient specific data at home and lab environment with knowledge specific for this kind of patient. In addition, it presents the significance of this ontology, and provides users with the opportunity to gain insight and knowledge from their data and the criteria expected to be available through it.

Currently, the ontology development is ongoing, the requirement elicitation will be done based on components capabilities and use case requirements, while the heterogeneous and dynamic data will be subjected to annotation through the development of semantic models for data sharing and usage, besides being interpretable. An imminent challenge that needs to be addressed is to finalize the ORSD and the most appropriate semantic model for the acquired data that fits our case. In the future, we will include more restrictions and different types of properties on the interactions between intervention classes so as to implement a more efficient and accurate version of the ontology.

Acknowledgements. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No GA101017558.

References

1. Hammad R, Barhoush M, Abed-Alguni BH. A Semantic-Based Approach for Managing Healthcare Big Data: A Survey. J Healthc Eng. 2020.
2. H. Zhang, Q. Li, G. Yi et al., “An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival,” BMC Medical Informatics and Decision Making, vol. 18, no. 2, p. 41, 2018.
3. Bravo, Maricela, Hoyos Reyes, Luis Fernando, & Reyes Ortiz, José A. Methodology for ontology design and construction (2019).
4. Suárez-Figueroa M.C., Gómez-Pérez A., Villazón-Terrazas B. How to Write & Use the Ontology Requirements Specification Document. Meersman R., Dillon T., Herrero P. On the Move to Meaningful Internet Systems, Lecture Notes in Computer Science, vol 5871. Springer, (2009).
5. Hlomani, H. and D. Stacey. “Approaches, methods, metrics, measures, and subjectivity in ontology evaluation : A survey.” (2014)
6. FHIR Homepage, http://hl7.org/fhir/
7. SNOMED-CT, https://bioportal.bioontology.org/ontologies/SNOMEDCT
8. ICD10, https://bioportal.bioontology.org/ontologies/ICD10
9. E. Younesi, A. Malhotra, M. Gündel, P. Scordis, A. Kodamullil, M. Page, B. Müller, S. Springstubbe, U. Wüllner, D. Scheller, M. Hofmann-Apitius.PDON: Parkinson's disease ontology for representation and modeling of the Parkinson's disease knowledge domain (2015).
10. https://bioportal.bioontology.org/ontologies/MSO
11. Basic Formal Ontology, https://basic-formal-ontology.org/
12. Compton M. , Barnaghi P., Bermudez L., García Castro R., Corcho O., Cox S., Graybeal J., Hauswirth M., Henson C., Herzog A., Huang V.,Janowicz K., Kelsey D., Phuoc D., Lefort L., Leggieri M., Neuhaus H., Nikolov A., Page K., Taylor K. (2012). The SSN Ontology of the W3C Semantic Sensor Network Incubator Group. Web Semantics: Science, Services and Agents on the World Wide Web.
13. K. Janowicz, M. Compton, “The Stimulus-Sensor-Observation Ontology Design-Pattern and its Integration into the Semantic Sensor Network Ontology”, International Workshop on Semantic Sensor Networks, vol. 668, CEUR-WS, 2010.
14. L. Chen, D. C. Nugent, H. Wang, “A Knowledge-Driven Approach to Activity Recognition in Smart Homes”, IEEE Trans. Knowl. Data Eng. 24(6): 961-974, 2012.
15. A. Scherp, T. Franz, C. Saathoff, and S. Staab, “A core ontology on events for representing occurrences in the real world”, In Multimedia Tools and Applications, 58(2):293–331, 2012.
16. http://motools.sourceforge.net/event/event.html.
17. http://demcare.eu/
18. Grûber, T.: A translation approach to portable ontology specification, Knowledge Acquisition 5(2) (1993) 199- 220
19. http://www.w3.org/TR/owl2-overview/
20. https://protege.stanford.edu/
21. Falco, R., Gangemi, A., Peroni, S., Shotton, D., & Vitali, F.,2014. Modelling OWL ontologies with Graffoo. European Semantic Web Conference (320-325).
22. N.Guarino, C. Welty. “An Overview of OntoClean”. (2004)

23. http://oops.linkeddata.es/
24. https://ontometrics.informatik.uni-rostock.de/ontologymetrics/